Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Machine Learning

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: `jean@seas.upenn.edu`

March 1, 2025

Contents

Chapter 1

Introduction

Chapter 2

Groups, Rings, and Fields

In the following four chapters, the basic algebraic structures (groups, rings, fields, vector spaces) are reviewed, with a major emphasis on vector spaces. Basic notions of linear algebra such as vector spaces, subspaces, linear combinations, linear independence, bases, quotient spaces, linear maps, matrices, change of bases, direct sums, linear forms, dual spaces, hyperplanes, transpose of a linear maps, are reviewed.

2.1 Groups, Subgroups, Cosets

The set $\mathbb{R}$ of real numbers has two operations $+ \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (addition) and $* \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (multiplication) satisfying properties that make $\mathbb{R}$ into an abelian group under $+$, and $\mathbb{R} - \{0\} = \mathbb{R}^*$ into an abelian group under $*$. Recall the definition of a group.

Definition 2.1. A *group* is a set G equipped with a binary operation $\cdot \colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in G is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c.$ (associativity);

(G2) $a \cdot e = e \cdot a = a.$ (identity);

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e.$ (inverse).

A group G is *abelian* (or *commutative*) if

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$

A set M together with an operation $\cdot \colon M \times M \to M$ and an element e satisfying only Conditions (G1) and (G2) is called a *monoid*. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.

Example 2.1.

1. The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is an abelian group under addition, with identity element 0. However, $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is not a group under multiplication.

2. The set $\mathbb{Q}$ of rational numbers (fractions p/q with $p, q \in \mathbb{Z}$ and $q \neq 0$) is an abelian group under addition, with identity element 0. The set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ is also an abelian group under multiplication, with identity element 1.

3. Given any nonempty set S, the set of bijections $f\colon S \to S$, also called *permutations of* S, is a group under function composition (i.e., the multiplication of f and g is the composition $g \circ f$), with identity element the identity function id_S. This group is not abelian as soon as S has more than two elements. The permutation group of the set $S = \{1, \ldots, n\}$ is often denoted $\mathfrak{S}_n$ and called the *symmetric group* on n elements.

4. For any positive integer $p \in \mathbb{N}$, define a relation on $\mathbb{Z}$, denoted $m \equiv n \pmod{p}$, as follows:
$$m \equiv n \pmod{p} \quad \text{iff} \quad m - n = kp \quad \text{for some } k \in \mathbb{Z}.$$
The reader will easily check that this is an equivalence relation, and, moreover, it is compatible with respect to addition and multiplication, which means that if $m_1 \equiv n_1 \pmod{p}$ and $m_2 \equiv n_2 \pmod{p}$, then $m_1 + m_2 \equiv n_1 + n_2 \pmod{p}$ and $m_1 m_2 \equiv n_1 n_2 \pmod{p}$. Consequently, we can define an addition operation and a multiplication operation of the set of equivalence classes $\pmod{p}$:
$$[m] + [n] = [m + n]$$
and
$$[m] \cdot [n] = [mn].$$
The reader will easily check that addition of residue classes $\pmod{p}$ induces an abelian group structure with $[0]$ as zero. This group is denoted $\mathbb{Z}/p\mathbb{Z}$.

5. The set $\mathbb{R}[X]$ of all polynomials in one variable X with real coefficients,
$$P(X) = a_0 X^n + a_1 X^{n-1} + \cdots + a_{n-1} X + a_n,$$
(with $a_i \in \mathbb{R}$), is an abelian group under addition of polynomials. The identity element is the zero polynomial.

6. The set of $n \times n$ invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *general linear group* and is usually denoted by $\mathbf{GL}(n, \mathbb{R})$ (or $\mathbf{GL}(n, \mathbb{C})$).

7. The set of $n \times n$ invertible matrices A with real (or complex) coefficients such that $\det(A) = 1$ is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *special linear group* and is usually denoted by $\mathbf{SL}(n, \mathbb{R})$ (or $\mathbf{SL}(n, \mathbb{C})$).

8. The set of $n \times n$ matrices Q with real coefficients such that

$$QQ^\top = Q^\top Q = I_n$$

is a group under matrix multiplication, with identity element the identity matrix I_n; we have $Q^{-1} = Q^\top$. This group is called the *orthogonal group* and is usually denoted by $\mathbf{O}(n)$.

9. The set of $n \times n$ invertible matrices Q with real coefficients such that

$$QQ^\top = Q^\top Q = I_n \quad \text{and} \quad \det(Q) = 1$$

is a group under matrix multiplication, with identity element the identity matrix I_n; as in (6), we have $Q^{-1} = Q^\top$. This group is called the *special orthogonal group* or *rotation group* and is usually denoted by $\mathbf{SO}(n)$.

The groups in (5)–(8) are nonabelian for $n \geq 2$, except for $\mathbf{SO}(2)$ which is abelian (but $\mathbf{O}(2)$ is not abelian).

It is customary to denote the operation of an abelian group G by $+$, in which case the inverse a^{-1} of an element $a \in G$ is denoted by $-a$.

The identity element of a group is *unique*. In fact, we can prove a more general fact:

Proposition 2.1. *For any binary operation $\cdot\colon M \times M \to M$, if $e' \in M$ is a left identity and if $e'' \in M$ is a right identity, which means that*

$$e' \cdot a = a \quad \text{for all} \quad a \in M \tag{G2l}$$

and

$$a \cdot e'' = a \quad \text{for all} \quad a \in M, \tag{G2r}$$

then $e' = e''$.

Proof. If we let $a = e''$ in equation (G2l), we get

$$e' \cdot e'' = e'',$$

and if we let $a = e'$ in equation (G2r), we get

$$e' \cdot e'' = e',$$

and thus

$$e' = e' \cdot e'' = e'',$$

as claimed. ∎

Proposition 2.1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a *unique inverse*. This is a consequence of a slightly more general fact:

Proposition 2.2. *In a monoid M with identity element e, if some element $a \in M$ has some left inverse $a' \in M$ and some right inverse $a'' \in M$, which means that*

$$a' \cdot a = e \tag{G3l}$$

and

$$a \cdot a'' = e, \tag{G3r}$$

then $a' = a''$.

Proof. Using (G3l) and the fact that e is an identity element, we have

$$(a' \cdot a) \cdot a'' = e \cdot a'' = a''.$$

Similarly, Using (G3r) and the fact that e is an identity element, we have

$$a' \cdot (a \cdot a'') = a' \cdot e = a'.$$

However, since M is monoid, the operation $\cdot$ is associative, so

$$a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',$$

as claimed. □

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Another important property about inverse elements in monoids is stated below.

Proposition 2.3. *In a monoid M with identity element e, if a and b are invertible elements of M, where a^{-1} is the inverse of a and b^{-1} is the inverse of b, then ab is invertible and its inverse is given by $(ab)^{-1} = b^{-1}a^{-1}$.*

Proof. Using associativity and the fact that e is the identity element we have

$$\begin{aligned}
(ab)(b^{-1}a^{-1}) &= a(b(b^{-1}a^{-1})) && \text{associativity} \\
&= a((bb^{-1})a^{-1}) && \text{associativity} \\
&= a(ea^{-1}) && b^{-1} \text{ is the inverse of } b \\
&= aa^{-1} && e \text{ is the identity element} \\
&= e. && a^{-1} \text{ is the inverse of } a.
\end{aligned}$$

We also have

$$\begin{aligned}
(b^{-1}a^{-1})(ab) &= b^{-1}(a^{-1}(ab)) && \text{associativity}\\
&= b^{-1}((a^{-1}a)b) && \text{associativity}\\
&= b^{-1}(eb) && a^{-1} \text{ is the inverse of } a\\
&= b^{-1}b && e \text{ is the identity element}\\
&= e. && b^{-1} \text{ is the inverse of } b.
\end{aligned}$$

Therefore $b^{-1}a^{-1}$ is the inverse of ab. $\square$

Observe that the inverse of ba is $a^{-1}b^{-1}$. Proposition 2.3 implies that the set of invertible elements of a monoid M is a group, also with identity element e.

Definition 2.2. If a group G has a finite number n of elements, we say that G is a group of *order* n. If G is infinite, we say that G has *infinite order*. The order of a group is usually denoted by $|G|$ (if G is finite).

Given a group G, for any two subsets $R, S \subseteq G$, we let

$$RS = \{r \cdot s \mid r \in R,\ s \in S\}.$$

In particular, for any $g \in G$, if $R = \{g\}$, we write

$$gS = \{g \cdot s \mid s \in S\},$$

and similarly, if $S = \{g\}$, we write

$$Rg = \{r \cdot g \mid r \in R\}.$$

From now on, we will drop the multiplication sign and write g_1g_2 for $g_1 \cdot g_2$.

Definition 2.3. Let G be a group. For any $g \in G$, define L_g, the *left translation by* g, by $L_g(a) = ga$, for all $a \in G$, and R_g, the *right translation by* g, by $R_g(a) = ag$, for all $a \in G$.

The following simple fact is often used.

Proposition 2.4. *Given a group G, the translations L_g and R_g are bijections.*

Proof. We show this for L_g, the proof for R_g being similar.

If $L_g(a) = L_g(b)$, then $ga = gb$, and multiplying on the left by g^{-1}, we get $a = b$, so L_g injective. For any $b \in G$, we have $L_g(g^{-1}b) = gg^{-1}b = b$, so L_g is surjective. Therefore, L_g is bijective. $\square$

Definition 2.4. Given a group G, a subset H of G is a *subgroup of* G iff

(1) The identity element e of G also belongs to H ($e \in H$);

(2) For all $h_1, h_2 \in H$, we have $h_1 h_2 \in H$;

(3) For all $h \in H$, we have $h^{-1} \in H$.

The proof of the following proposition is left as an exercise.

Proposition 2.5. *Given a group G, a subset $H \subseteq G$ is a subgroup of G iff H is nonempty and whenever $h_1, h_2 \in H$, then $h_1 h_2^{-1} \in H$.*

If the group G is finite, then the following criterion can be used.

Proposition 2.6. *Given a finite group G, a subset $H \subseteq G$ is a subgroup of G iff*

(1) $e \in H$;

(2) H is closed under multiplication.

Proof. We just have to prove that Condition (3) of Definition 2.4 holds. For any $a \in H$, since the left translation L_a is bijective, its restriction to H is injective, and since H is finite, it is also bijective. Since $e \in H$, there is a unique $b \in H$ such that $L_a(b) = ab = e$. However, if a^{-1} is the inverse of a in G, we also have $L_a(a^{-1}) = aa^{-1} = e$, and by injectivity of L_a, we have $a^{-1} = b \in H$. □

Example 2.2.

1. For any integer $n \in \mathbb{Z}$, the set
$$n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$$
is a subgroup of the group $\mathbb{Z}$.

2. The set of matrices
$$\mathbf{GL}^+(n, \mathbb{R}) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid \det(A) > 0\}$$
is a subgroup of the group $\mathbf{GL}(n, \mathbb{R})$.

3. The group $\mathbf{SL}(n, \mathbb{R})$ is a subgroup of the group $\mathbf{GL}(n, \mathbb{R})$.

4. The group $\mathbf{O}(n)$ is a subgroup of the group $\mathbf{GL}(n, \mathbb{R})$.

5. The group $\mathbf{SO}(n)$ is a subgroup of the group $\mathbf{O}(n)$, and a subgroup of the group $\mathbf{SL}(n, \mathbb{R})$.

6. It is not hard to show that every 2×2 rotation matrix $R \in \mathbf{SO}(2)$ can be written as

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \text{with} \quad 0 \leq \theta < 2\pi.$$

   Then $\mathbf{SO}(2)$ can be considered as a subgroup of $\mathbf{SO}(3)$ by viewing the matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

   as the matrix

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

7. The set of 2×2 upper-triangular matrices of the form

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \quad a, b, c \in \mathbb{R},\ a, c \neq 0$$

   is a subgroup of the group $\mathbf{GL}(2, \mathbb{R})$.

8. The set V consisting of the four matrices

$$\begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix}$$

   is a subgroup of the group $\mathbf{GL}(2, \mathbb{R})$ called the *Klein four-group*.

Definition 2.5. If H is a subgroup of G and $g \in G$ is any element, the sets of the form gH are called *left cosets of H in G* and the sets of the form Hg are called *right cosets of H in G*. The left cosets (resp. right cosets) of H induce an equivalence relation $\sim$ defined as follows: For all $g_1, g_2 \in G$,

$$g_1 \sim g_2 \quad \text{iff} \quad g_1 H = g_2 H$$

(resp. $g_1 \sim g_2$ iff $Hg_1 = Hg_2$). Obviously, $\sim$ is an equivalence relation.

Now, we claim the following fact:

Proposition 2.7. *Given a group G and any subgroup H of G, we have $g_1 H = g_2 H$ iff $g_2^{-1} g_1 H = H$ iff $g_2^{-1} g_1 \in H$, for all $g_1, g_2 \in G$.*

Proof. If we apply the bijection $L_{g_2^{-1}}$ to both $g_1 H$ and $g_2 H$ we get $L_{g_2^{-1}}(g_1 H) = g_2^{-1} g_1 H$ and $L_{g_2^{-1}}(g_2 H) = H$, so $g_1 H = g_2 H$ iff $g_2^{-1} g_1 H = H$. If $g_2^{-1} g_1 H = H$, since $1 \in H$, we get $g_2^{-1} g_1 \in H$. Conversely, if $g_2^{-1} g_1 \in H$, since H is a group, the left translation $L_{g_2^{-1} g_1}$ is a bijection of H, so $g_2^{-1} g_1 H = H$. Thus, $g_2^{-1} g_1 H = H$ iff $g_2^{-1} g_1 \in H$. $\square$

It follows that the equivalence class of an element $g \in G$ is the coset gH (resp. Hg). Since L_g is a bijection between H and gH, the cosets gH all have the same cardinality. The map $L_{g^{-1}} \circ R_g$ is a bijection between the left coset gH and the right coset Hg, so they also have the same cardinality. Since the distinct cosets gH form a partition of G, we obtain the following fact:

Proposition 2.8. *(Lagrange) For any finite group G and any subgroup H of G, the order h of H divides the order n of G.*

Definition 2.6. Given a finite group G and a subgroup H of G, if $n = |G|$ and $h = |H|$, then the ratio n/h is denoted by $(G : H)$ and is called the *index of H in G*.

The index $(G : H)$ is the number of left (and right) cosets of H in G. Proposition 2.8 can be stated as

$$|G| = (G : H)|H|.$$

The set of left cosets of H in G (which, in general, is **not** a group) is denoted G/H. The “points” of G/H are obtained by “collapsing” all the elements in a coset into a single element.

Example 2.3.

1. Let n be any positive integer, and consider the subgroup $n\mathbb{Z}$ of $\mathbb{Z}$ (under addition). The coset of any integer $m \in \mathbb{Z}$ is

$$m + n\mathbb{Z} = \{m + nk \mid k \in \mathbb{Z}\}.$$

In particular, the coset of 0 is the set $n\mathbb{Z}$. By dividing m by n, we have $m = nq + r$ for some unique r such that $0 \leq r \leq n - 1$. But then we see that r is the smallest nonnegative element of the coset $m + n\mathbb{Z}$. This implies that there is a bijection betwen the cosets of the subgroup $n\mathbb{Z}$ of $\mathbb{Z}$ and the set of residues $\{0, 1, \ldots, n - 1\}$ modulo n, or equivalently a bijection with $\mathbb{Z}/n\mathbb{Z}$.

2. The cosets of $\mathbf{SL}(n, \mathbb{R})$ in $\mathbf{GL}(n, \mathbb{R})$ are the sets of matrices

$$A\,\mathbf{SL}(n, \mathbb{R}) = \{AB \mid B \in \mathbf{SL}(n, \mathbb{R})\},\ A \in \mathbf{GL}(n, \mathbb{R}).$$

Since A is invertible, $\det(A) \neq 0$, and we can write $A = (\det(A))^{1/n}((\det(A))^{-1/n}A)$ if $\det(A) > 0$ and $A = (-\det(A))^{1/n}((-\det(A))^{-1/n}A)$ if $\det(A) < 0$. But we have $(\det(A))^{-1/n}A \in \mathbf{SL}(n, \mathbb{R})$ if $\det(A) > 0$ and $-(-\det(A))^{-1/n}A \in \mathbf{SL}(n, \mathbb{R})$ if $\det(A) < 0$, so the coset $A\,\mathbf{SL}(n, \mathbb{R})$ contains the matrix

$$(\det(A))^{1/n}I_n \quad \text{if} \quad \det(A) > 0, \quad -(-\det(A))^{1/n}I_n \quad \text{if} \quad \det(A) < 0.$$

It follows that there is a bijection between the cosets of $\mathbf{SL}(n, \mathbb{R})$ in $\mathbf{GL}(n, \mathbb{R})$ and $\mathbb{R}$.

3. The cosets of $\mathbf{SO}(n)$ in $\mathbf{GL}^+(n, \mathbb{R})$ are the sets of matrices
$$A\,\mathbf{SO}(n) = \{AQ \mid Q \in \mathbf{SO}(n)\},\ A \in \mathbf{GL}^+(n, \mathbb{R}).$$
It can be shown (using the polar form for matrices) that there is a bijection between the cosets of $\mathbf{SO}(n)$ in $\mathbf{GL}^+(n, \mathbb{R})$ and the set of $n \times n$ symmetric, positive, definite matrices; these are the symmetric matrices whose eigenvalues are strictly positive.

4. The cosets of $\mathbf{SO}(2)$ in $\mathbf{SO}(3)$ are the sets of matrices
$$Q\,\mathbf{SO}(2) = \{QR \mid R \in \mathbf{SO}(2)\},\ Q \in \mathbf{SO}(3).$$
The group $\mathbf{SO}(3)$ moves the points on the sphere S^2 in $\mathbb{R}^3$, namely for any $x \in S^2$,
$$x \mapsto Qx \quad \text{for any rotation } Q \in \mathbf{SO}(3).$$
Here,
$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$
Let $N = (0, 0, 1)$ be the north pole on the sphere S^2. Then it is not hard to show that $\mathbf{SO}(2)$ is precisely the subgroup of $\mathbf{SO}(3)$ that leaves N fixed. As a consequence, all rotations QR in the coset $Q\,\mathbf{SO}(2)$ map N to the same point $QN \in S^2$, and it can be shown that there is a bijection between the cosets of $\mathbf{SO}(2)$ in $\mathbf{SO}(3)$ and the points on S^2. The surjectivity of this map has to do with the fact that the action of $\mathbf{SO}(3)$ on S^2 is transitive, which means that for any point $x \in S^2$, there is some rotation $Q \in \mathbf{SO}(3)$ such that $QN = x$.

It is tempting to define a multiplication operation on left cosets (or right cosets) by setting
$$(g_1H)(g_2H) = (g_1g_2)H,$$
but this operation is not well defined in general, unless the subgroup H possesses a special property. In Example 2.3, it is possible to define multiplication of cosets in (1), but it is not possible in (2) and (3).

The property of the subgroup H that allows defining a multiplication operation on left cosets is typical of the kernels of group homomorphisms, so we are led to the following definition.

Definition 2.7. Given any two groups G and G', a function $\varphi\colon G \to G'$ is a *homomorphism* iff
$$\varphi(g_1g_2) = \varphi(g_1)\varphi(g_2), \quad \text{for all } g_1, g_2 \in G.$$

Taking $g_1 = g_2 = e$ (in G), we see that
$$\varphi(e) = e',$$
and taking $g_1 = g$ and $g_2 = g^{-1}$, we see that
$$\varphi(g^{-1}) = (\varphi(g))^{-1}.$$

Example 2.4.

1. The map $\varphi\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ given by $\varphi(m) = m \bmod n$ for all $m \in \mathbb{Z}$ is a homomorphism.
2. The map $\det\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbb{R}$ is a homomorphism because $\det(AB) = \det(A)\det(B)$ for any two matrices A, B. Similarly, the map $\det\colon \mathbf{O}(n) \to \mathbb{R}$ is a homomorphism.

If $\varphi\colon G \to G'$ and $\psi\colon G' \to G''$ are group homomorphisms, then $\psi \circ \varphi\colon G \to G''$ is also a homomorphism. If $\varphi\colon G \to G'$ is a homomorphism of groups, and if $H \subseteq G$, $H' \subseteq G'$ are two subgroups, then it is easily checked that

$$\mathrm{Im}\ \varphi = \varphi(H) = \{\varphi(g) \mid g \in H\}$$

is a subgroup of G' and

$$\varphi^{-1}(H') = \{g \in G \mid \varphi(g) \in H'\}$$

is a subgroup of G. In particular, when $H' = \{e'\}$, we obtain the *kernel*, $\mathrm{Ker}\ \varphi$, of φ.

Definition 2.8. If $\varphi\colon G \to G'$ is a homomorphism of groups, and if $H \subseteq G$ is a subgroup of G, then the subgroup of G',

$$\mathrm{Im}\ \varphi = \varphi(H) = \{\varphi(g) \mid g \in H\},$$

is called the *image of H by φ*, and the subgroup of G,

$$\mathrm{Ker}\ \varphi = \{g \in G \mid \varphi(g) = e'\},$$

is called the *kernel* of φ.

Example 2.5.

1. The kernel of the homomorphism $\varphi\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is $n\mathbb{Z}$.
2. The kernel of the homomorphism $\det\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbb{R}$ is $\mathbf{SL}(n, \mathbb{R})$. Similarly, the kernel of the homomorphism $\det\colon \mathbf{O}(n) \to \mathbb{R}$ is $\mathbf{SO}(n)$.

The following characterization of the injectivity of a group homomorphism is used all the time.

Proposition 2.9. *If $\varphi\colon G \to G'$ is a homomorphism of groups, then $\varphi\colon G \to G'$ is injective iff* $\mathrm{Ker}\ \varphi = \{e\}$. *(We also write* $\mathrm{Ker}\ \varphi = (0)$.*)*

Proof. Assume φ is injective. Since $\varphi(e) = e'$, if $\varphi(g) = e'$, then $\varphi(g) = \varphi(e)$, and by injectivity of φ we must have $g = e$, so $\mathrm{Ker}\ \varphi = \{e\}$.

Conversely, assume that $\mathrm{Ker}\ \varphi = \{e\}$. If $\varphi(g_1) = \varphi(g_2)$, then by multiplication on the left by $(\varphi(g_1))^{-1}$ we get

$$e' = (\varphi(g_1))^{-1}\varphi(g_1) = (\varphi(g_1))^{-1}\varphi(g_2),$$

and since φ is a homomorphism $(\varphi(g_1))^{-1} = \varphi(g_1^{-1})$, so

$$e' = (\varphi(g_1))^{-1}\varphi(g_2) = \varphi(g_1^{-1})\varphi(g_2) = \varphi(g_1^{-1}g_2).$$

This shows that $g_1^{-1}g_2 \in \text{Ker}\ \varphi$, but since $\text{Ker}\ \varphi = \{e\}$ we have $g_1^{-1}g_2 = e$, and thus $g_2 = g_1$, proving that φ is injective. ☐

Definition 2.9. We say that a group homomorphism $\varphi\colon G \to G'$ is an *isomorphism* if there is a homomorphism $\psi\colon G' \to G$, so that

$$\psi \circ \varphi = \text{id}_G \quad \text{and} \quad \varphi \circ \psi = \text{id}_{G'}. \tag{$\dagger$}$$

If φ is an isomorphism we say that the groups G and G' are *isomorphic*. When $G' = G$, a group isomorphism is called an *automorphism*.

The reasoning used in the proof of Proposition 2.2 shows that if a a group homomorphism $\varphi\colon G \to G'$ is an isomorphism, then the homomorphism $\psi\colon G' \to G$ satisfying Condition ($\dagger$) is unique. This homomorphism is denoted φ^{-1}.

The left translations L_g and the right translations R_g are automorphisms of G.

Suppose $\varphi\colon G \to G'$ is a bijective homomorphism, and let φ^{-1} be the inverse of φ (as a function). Then for all $a, b \in G$, we have

$$\varphi(\varphi^{-1}(a)\varphi^{-1}(b)) = \varphi(\varphi^{-1}(a))\varphi(\varphi^{-1}(b)) = ab,$$

and so

$$\varphi^{-1}(ab) = \varphi^{-1}(a)\varphi^{-1}(b),$$

which proves that φ^{-1} is a homomorphism. Therefore, we proved the following fact.

Proposition 2.10. *A bijective group homomorphism $\varphi\colon G \to G'$ is an isomorphism.*

Observe that the property

$$gH = Hg, \quad \text{for all } g \in G. \tag{$*$}$$

is equivalent by multiplication on the right by g^{-1} to

$$gHg^{-1} = H, \quad \text{for all } g \in G,$$

and the above is equivalent to

$$gHg^{-1} \subseteq H, \quad \text{for all } g \in G. \tag{$**$}$$

This is because $gHg^{-1} \subseteq H$ implies $H \subseteq g^{-1}Hg$, and this for all $g \in G$.

Proposition 2.11. *Let $\varphi\colon G \to G'$ be a group homomorphism. Then $H = \text{Ker}\ \varphi$ satisfies Property ($**$), and thus Property ($*$).*

Proof. We have

$$\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g^{-1}) = \varphi(g)e'\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e',$$

for all $h \in H = \mathrm{Ker}\ \varphi$ and all $g \in G$. Thus, by definition of $H = \mathrm{Ker}\ \varphi$, we have $gHg^{-1} \subseteq H$. ☐

Definition 2.10. For any group G, a subgroup N of G is a *normal subgroup* of G iff

$$gNg^{-1} = N, \quad \text{for all } g \in G.$$

This is denoted by $N \lhd G$.

Proposition 2.11 shows that the kernel Ker φ of a homomorphism $\varphi\colon G \to G'$ is a normal subgroup of G.

Observe that if G is abelian, then *every* subgroup of G is normal.

Consider Example 2.2. Let $R \in \mathbf{SO}(2)$ and $A \in \mathbf{SL}(2, \mathbb{R})$ be the matrices

$$R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then

$$A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and we have

$$ARA^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \end{pmatrix},$$

and clearly $ARA^{-1} \notin \mathbf{SO}(2)$. Therefore $\mathbf{SO}(2)$ is not a normal subgroup of $\mathbf{SL}(2, \mathbb{R})$. The same counter-example shows that $\mathbf{O}(2)$ is not a normal subgroup of $\mathbf{GL}(2, \mathbb{R})$.

Let $R \in \mathbf{SO}(2)$ and $Q \in \mathbf{SO}(3)$ be the matrices

$$R = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then

$$Q^{-1} = Q^{\top} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

and we have

$$\begin{aligned} QRQ^{-1} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \end{aligned}$$

Observe that $QRQ^{-1} \notin \mathbf{SO}(2)$, so $\mathbf{SO}(2)$ is not a normal subgroup of $\mathbf{SO}(3)$.

Let T and $A \in \mathbf{GL}(2, \mathbb{R})$ be the following matrices

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We have

$$A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = A,$$

and

$$ATA^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

The matrix T is upper triangular, but ATA^{-1} is not, so the group of 2×2 upper triangular matrices is not a normal subgroup of $\mathbf{GL}(2, \mathbb{R})$.

Let $Q \in V$ and $A \in \mathbb{GL}(2, \mathbb{R})$ be the following matrices

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

We have

$$A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and

$$AQA^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 0 & -1 \end{pmatrix}.$$

Clearly $AQA^{-1} \notin V$, which shows that the Klein four group is not a normal subgroup of $\mathbf{GL}(2, \mathbb{R})$.

The reader should check that the subgroups $n\mathbb{Z}$, $\mathbf{GL}^+(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{R})$, and $\mathbf{SO}(n, \mathbb{R})$ as a subgroup of $\mathbf{O}(n, \mathbb{R})$, are normal subgroups.

If N is a normal subgroup of G, the equivalence relation $\sim$ induced by left cosets (see Definition 2.5) is the same as the equivalence induced by right cosets. Furthermore, this equivalence relation is a *congruence*, which means that: For all $g_1, g_2, g'_1, g'_2 \in G$,

(1) If $g_1N = g_1'N$ and $g_2N = g_2'N$, then $g_1g_2N = g_1'g_2'N$, and

(2) If $g_1N = g_2N$, then $g_1^{-1}N = g_2^{-1}N$.

As a consequence, we can define a group structure on the set $G/\sim$ of equivalence classes modulo $\sim$, by setting

$$(g_1N)(g_2N) = (g_1g_2)N.$$

Definition 2.11. Let G be a group and N be a normal subgroup of G. The group obtained by defining the multiplication of (left) cosets by

$$(g_1N)(g_2N) = (g_1g_2)N, \quad g_1, g_2 \in G$$

is denoted G/N, and called the *quotient of G by N*. The equivalence class gN of an element $g \in G$ is also denoted $\overline{g}$ (or $[g]$). The map $\pi\colon G \to G/N$ given by

$$\pi(g) = \overline{g} = gN$$

is a group homomorphism called the *canonical projection*.

Since the kernel of a homomorphism is a normal subgroup, we obtain the following very useful result.

Proposition 2.12. *Given a homomorphism of groups $\varphi\colon G \to G'$, the groups $G/\mathrm{Ker}\ \varphi$ and $\mathrm{Im}\ \varphi = \varphi(G)$ are isomorphic.*

Proof. Since φ is surjective onto its image, we may assume that φ is surjective, so that $G' = \mathrm{Im}\ \varphi$. We define a map $\overline{\varphi}\colon G/\mathrm{Ker}\ \varphi \to G'$ as follows:

$$\overline{\varphi}(\overline{g}) = \varphi(g), \quad g \in G.$$

We need to check that the definition of this map does not depend on the representative chosen in the coset $\overline{g} = g\,\mathrm{Ker}\ \varphi$, and that it is a homomorphism. If g' is another element in the coset $g\,\mathrm{Ker}\ \varphi$, which means that $g' = gh$ for some $h \in \mathrm{Ker}\,\varphi$, then

$$\varphi(g') = \varphi(gh) = \varphi(g)\varphi(h) = \varphi(g)e' = \varphi(g),$$

since $\varphi(h) = e'$ as $h \in \mathrm{Ker}\ \varphi$. This shows that

$$\overline{\varphi}(\overline{g'}) = \varphi(g') = \varphi(g) = \overline{\varphi}(\overline{g}),$$

so the map $\overline{\varphi}$ is well defined. It is a homomorphism because

$$\begin{aligned}
\overline{\varphi}(\overline{g}\overline{g'}) &= \overline{\varphi}(\overline{gg'}) \\
&= \varphi(gg') \\
&= \varphi(g)\varphi(g') \\
&= \overline{\varphi}(\overline{g})\overline{\varphi}(\overline{g'}).
\end{aligned}$$

The map $\overline{\varphi}$ is injective because $\overline{\varphi}(\overline{g}) = e'$ iff $\varphi(g) = e'$ iff $g \in \mathrm{Ker}\ \varphi$, iff $\overline{g} = \overline{e}$. The map $\overline{\varphi}$ is surjective because φ is surjective. Therefore $\overline{\varphi}$ is a bijective homomorphism, and thus an isomorphism, as claimed. $\square$

Proposition 2.12 is called the *first isomorphism theorem*.

A useful way to construct groups is the *direct product* construction.

Definition 2.12. Given two groups G an H, we let $G \times H$ be the Cartesian product of the sets G and H with the multiplication operation $\cdot$ given by

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2).$$

It is immediately verified that $G \times H$ is a group called the *direct product* of G and H.

Similarly, given any n groups $G_1, \ldots, G_n$, we can define the direct product $G_1 \times \cdots \times G_n$ is a similar way.

If G is an abelian group and $H_1, \ldots, H_n$ are subgroups of G, the situation is simpler. Consider the map

$$a \colon H_1 \times \cdots \times H_n \to G$$

given by

$$a(h_1, \ldots, h_n) = h_1 + \cdots + h_n,$$

using $+$ for the operation of the group G. It is easy to verify that a is a group homomorphism, so its image is a subgroup of G denoted by $H_1 + \cdots + H_n$, and called the *sum* of the groups H_i. The following proposition will be needed.

Proposition 2.13. *Given an abelian group G, if H_1 and H_2 are any subgroups of G such that $H_1 \cap H_2 = \{0\}$, then the map a is an isomorphism*

$$a \colon H_1 \times H_2 \to H_1 + H_2.$$

Proof. The map is surjective by definition, so we just have to check that it is injective. For this, we show that $\mathrm{Ker}\, a = \{(0, 0)\}$. We have $a(a_1, a_2) = 0$ iff $a_1 + a_2 = 0$ iff $a_1 = -a_2$. Since $a_1 \in H_1$ and $a_2 \in H_2$, we see that $a_1, a_2 \in H_1 \cap H_2 = \{0\}$, so $a_1 = a_2 = 0$, which proves that $\mathrm{Ker}\, a = \{(0, 0)\}$. $\square$

Under the conditions of Proposition 2.13, namely $H_1 \cap H_2 = \{0\}$, the group $H_1 + H_2$ is called the *direct sum* of H_1 and H_2; it is denoted by $H_1 \oplus H_2$, and we have an isomorphism $H_1 \times H_2 \cong H_1 \oplus H_2$.

2.2 Cyclic Groups

Given a group G with unit element 1, for any element $g \in G$ and for any natural number $n \in \mathbb{N}$, we define g^n as follows:

$$\begin{aligned} g^0 &= 1 \\ g^{n+1} &= g \cdot g^n. \end{aligned}$$

For any integer $n \in \mathbb{Z}$, we define g^n by

$$g^n = \begin{cases} g^n & \text{if } n \geq 0 \\ (g^{-1})^{(-n)} & \text{if } n < 0. \end{cases}$$

The following properties are easily verified:

$$\begin{aligned} g^i \cdot g^j &= g^{i+j} \\ (g^i)^{-1} &= g^{-i} \\ g^i \cdot g^j &= g^j \cdot g^i, \end{aligned}$$

for all $i, j \in \mathbb{Z}$.

Define the subset $\langle g \rangle$ of G by

$$\langle g \rangle = \{g^n \mid n \in \mathbb{Z}\}.$$

The following proposition is left as an exercise.

Proposition 2.14. *Given a group G, for any element $g \in G$, the set $\langle g \rangle$ is the smallest abelian subgroup of G containing g.*

Definition 2.13. A group G is *cyclic* iff there is some element $g \in G$ such that $G = \langle g \rangle$. An element $g \in G$ with this property is called a *generator* of G.

The Klein four group V of Example 2.2 is abelian, but not cyclic. This is because V has four elements, but all the elements different from the identity have order 2.

Cyclic groups are quotients of $\mathbb{Z}$. For this, we use a basic property of $\mathbb{Z}$. Recall that for any $n \in \mathbb{Z}$, we let $n\mathbb{Z}$ denote the set of multiples of n,

$$n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}.$$

Proposition 2.15. *Every subgroup H of $\mathbb{Z}$ is of the form $H = n\mathbb{Z}$ for some $n \in \mathbb{N}$.*

Proof. If H is the trivial group $\{0\}$, then let $n = 0$. If H is nontrivial, for any nonzero element $m \in H$, we also have $-m \in H$ and either m or $-m$ is positive, so let n be the smallest positive integer in H. By Proposition 2.14, $n\mathbb{Z}$ is the smallest subgroup of H containing n. For any $m \in H$ with $m \neq 0$, we can write

$$m = nq + r, \quad \text{with} \quad 0 \leq r < n.$$

Now, since $n\mathbb{Z} \subseteq H$, we have $nq \in H$, and since $m \in H$, we get $r = m - nq \in H$. However, $0 \leq r < n$, contradicting the minimality of n, so $r = 0$, and $H = n\mathbb{Z}$. $\square$

Given any cyclic group G, for any generator g of G, we can define a mapping $\varphi\colon \mathbb{Z} \to G$ by $\varphi(m) = g^m$. Since g generates G, this mapping is surjective. The mapping φ is clearly a group homomorphism, so let $H = \operatorname{Ker}\varphi$ be its kernel. By a previous observation, $H = n\mathbb{Z}$ for some $n \in \mathbb{Z}$, so by the first homomorphism theorem, we obtain an isomorphism

$$\overline{\varphi}\colon \mathbb{Z}/n\mathbb{Z} \longrightarrow G$$

from the quotient group $\mathbb{Z}/n\mathbb{Z}$ onto G. Obviously, if G has finite order, then $|G| = n$. In summary, we have the following result.

Proposition 2.16. *Every cyclic group G is either isomorphic to $\mathbb{Z}$, or to $\mathbb{Z}/n\mathbb{Z}$, for some natural number $n > 0$. In the first case, we say that G is an infinite cyclic group, and in the second case, we say that G is a cyclic group of order n.*

The quotient group $\mathbb{Z}/n\mathbb{Z}$ consists of the cosets $m + n\mathbb{Z} = \{m + nk \mid k \in \mathbb{Z}\}$, with $m \in \mathbb{Z}$, that is, of the equivalence classes of $\mathbb{Z}$ under the equivalence relation $\equiv$ defined such that

$$x \equiv y \quad \text{iff} \quad x - y \in nZ \quad \text{iff} \quad x \equiv y \pmod{n}.$$

We also denote the equivalence class $x + n\mathbb{Z}$ of x by $\overline{x}$, or if we want to be more precise by $[x]_n$. The group operation is given by

$$\overline{x} + \overline{y} = \overline{x + y}.$$

For every $x \in \mathbb{Z}$, there is a unique representative, $x \bmod n$ (the nonnegative remainder of the division of x by n) in the class $\overline{x}$ of x, such that $0 \le x \bmod n \le n - 1$. For this reason, we often identity $\mathbb{Z}/n\mathbb{Z}$ with the set $\{0, \ldots, n-1\}$. To be more rigorous, we can give $\{0, \ldots, n-1\}$ a group structure by defining $+_n$ such that

$$x +_n y = (x + y) \bmod n.$$

Then, it is easy to see that $\{0, \ldots, n-1\}$ with the operation $+_n$ is a group with identity element 0 isomorphic to $\mathbb{Z}/n\mathbb{Z}$.

We can also define a multiplication operation $\cdot$ on $\mathbb{Z}/n\mathbb{Z}$ as follows:

$$\overline{a} \cdot \overline{b} = \overline{ab} = \overline{ab \bmod n}.$$

Then, it is easy to check that $\cdot$ is abelian, associative, that 1 is an identity element for $\cdot$, and that $\cdot$ is distributive on the left and on the right with respect to addition. This makes $\mathbb{Z}/n\mathbb{Z}$ into a *commutative ring*. We usually suppress the dot and write $\overline{a}\,\overline{b}$ instead of $\overline{a} \cdot \overline{b}$.

Proposition 2.17. *Given any integer $n \ge 1$, for any $a \in \mathbb{Z}$, the residue class $\overline{a} \in \mathbb{Z}/n\mathbb{Z}$ is invertible with respect to multiplication iff* $\gcd(a, n) = 1$.

Proof. If $\overline{a}$ has inverse $\overline{b}$ in $\mathbb{Z}/n\mathbb{Z}$, then $\overline{a}\,\overline{b} = 1$, which means that

$$ab \equiv 1 \pmod{n},$$

that is $ab = 1 + nk$ for some $k \in \mathbb{Z}$, which is the Bezout identity

$$ab - nk = 1$$

and implies that $\gcd(a, n) = 1$. Conversely, if $\gcd(a, n) = 1$, then by Bezout's identity there exist $u, v \in \mathbb{Z}$ such that

$$au + nv = 1,$$

so $au = 1 - nv$, that is,

$$au \equiv 1 \pmod{n},$$

which means that $\overline{a}\,\overline{u} = 1$, so $\overline{a}$ is invertible in $\mathbb{Z}/n\mathbb{Z}$. □

Definition 2.14. The group (under multiplication) of invertible elements of the ring $\mathbb{Z}/n\mathbb{Z}$ is denoted by $(\mathbb{Z}/n\mathbb{Z})^*$. Note that this group is abelian and only defined if $n \geq 2$.

The *Euler* φ*-function* plays an important role in the theory of the groups $(\mathbb{Z}/n\mathbb{Z})^*$.

Definition 2.15. Given any positive integer $n \geq 1$, the *Euler* φ*-function* (or Euler *totient function*) is defined such that $\varphi(n)$ is the number of integers a, with $1 \leq a \leq n$, which are relatively prime to n; that is, with $\gcd(a, n) = 1$.[1]

Then, by Proposition 2.17, we see that the group $(\mathbb{Z}/n\mathbb{Z})^*$ has order $\varphi(n)$.

For $n = 2$, $(\mathbb{Z}/2\mathbb{Z})^* = \{1\}$, the trivial group. For $n = 3$, $(\mathbb{Z}/3\mathbb{Z})^* = \{1, 2\}$, and for $n = 4$, we have $(\mathbb{Z}/4\mathbb{Z})^* = \{1, 3\}$. Both groups are isomorphic to the group $\{-1, 1\}$. Since $\gcd(a, n) = 1$ for every $a \in \{1, \ldots, n-1\}$ iff n is prime, by Proposition 2.17 we see that $(\mathbb{Z}/n\mathbb{Z})^* = \mathbb{Z}/n\mathbb{Z} - \{0\}$ iff n is prime.

2.3 Rings and Fields

The groups $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}/n\mathbb{Z}$, and $\mathrm{M}_n(\mathbb{R})$ are more than abelian groups, they are also commutative rings. Furthermore, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields. We now introduce rings and fields.

Definition 2.16. A *ring* is a set A equipped with two operations $+\colon A \times A \to A$ (called *addition*) and $*\colon A \times A \to A$ (called *multiplication*) having the following properties:

(R1) A is an abelian group w.r.t. $+$;

(R2) $*$ is associative and has an identity element $1 \in A$;

[1] We allow $a = n$ to accomodate the special case $n = 1$.

(R3) $*$ is distributive w.r.t. $+$.

The identity element for addition is denoted 0, and the additive inverse of $a \in A$ is denoted by $-a$. More explicitly, the axioms of a ring are the following equations which hold for all $a, b, c \in A$:

$$\begin{aligned}
a + (b + c) &= (a + b) + c && \text{(associativity of } +) && (2.1)\\
a + b &= b + a && \text{(commutativity of } +) && (2.2)\\
a + 0 &= 0 + a = a && \text{(zero)} && (2.3)\\
a + (-a) &= (-a) + a = 0 && \text{(additive inverse)} && (2.4)\\
a * (b * c) &= (a * b) * c && \text{(associativity of } *) && (2.5)\\
a * 1 &= 1 * a = a && \text{(identity for } *) && (2.6)\\
(a + b) * c &= (a * c) + (b * c) && \text{(distributivity)} && (2.7)\\
a * (b + c) &= (a * b) + (a * c) && \text{(distributivity)} && (2.8)
\end{aligned}$$

The ring A is *commutative* if

$$a * b = b * a \quad \text{for all } a, b \in A.$$

From (2.7) and (2.8), we easily obtain

$$\begin{aligned}
a * 0 &= 0 * a = 0 && (2.9)\\
a * (-b) &= (-a) * b = -(a * b). && (2.10)
\end{aligned}$$

Note that (2.9) implies that if $1 = 0$, then $a = 0$ for all $a \in A$, and thus, $A = \{0\}$. The ring $A = \{0\}$ is called the *trivial ring*. A ring for which $1 \neq 0$ is called *nontrivial*. The multiplication $a * b$ of two elements $a, b \in A$ is often denoted by ab.

Example 2.6.

1. The additive groups $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, are commutative rings.

2. For any positive integer $n \in \mathbb{N}$, the group $\mathbb{Z}/n\mathbb{Z}$ is a group under addition. We can also define a multiplication operation by

$$\overline{a} \cdot \overline{b} = \overline{ab} = \overline{ab \bmod n},$$

   for all $a, b \in \mathbb{Z}$. The reader will easily check that the ring axioms are satisfied, with $\overline{0}$ as zero and $\overline{1}$ as multiplicative unit. The resulting ring is denoted by $\mathbb{Z}/n\mathbb{Z}$.[2]

3. The group $\mathbb{R}[X]$ of polynomials in one variable with real coefficients is a ring under multiplication of polynomials. It is a commutative ring.

[2]The notation $\mathbb{Z}_n$ is sometimes used instead of $\mathbb{Z}/n\mathbb{Z}$ but it clashes with the notation for the *n-adic integers* so we prefer not to use it.

4. Let d be any positive integer. If d is not divisible by any integer of the form m^2, with $m \in \mathbb{N}$ and $m \geq 2$, then we say that d is *square-free*. For example, $d = 1, 2, 3, 5, 6, 7, 10$ are square-free, but $4, 8, 9, 12$ are not square-free. If d is any square-free integer and if $d \geq 2$, then the set of real numbers

$$\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} \in \mathbb{R} \mid a, b \in \mathbb{Z}\}$$

is a commutative a ring. If $z = a + b\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$, we write $\overline{z} = a - b\sqrt{d}$. Note that $z\overline{z} = a^2 - db^2$.

5. Similarly, if $d \geq 1$ is a positive square-free integer, then the set of complex numbers

$$\mathbb{Z}[\sqrt{-d}] = \{a + ib\sqrt{d} \in \mathbb{C} \mid a, b \in \mathbb{Z}\}$$

is a commutative ring. If $z = a + ib\sqrt{d} \in \mathbb{Z}[\sqrt{-d}]$, we write $\overline{z} = a - ib\sqrt{d}$. Note that $z\overline{z} = a^2 + db^2$. The case where $d = 1$ is a famous example that was investigated by Gauss, and $\mathbb{Z}[\sqrt{-1}]$, also denoted $\mathbb{Z}[i]$, is called the ring of *Gaussian integers*.

6. The group of $n \times n$ matrices $\mathrm{M}_n(\mathbb{R})$ is a ring under matrix multiplication. However, it is not a commutative ring.

7. The group $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ is a ring under the operation $f \cdot g$ defined such that

$$(f \cdot g)(x) = f(x)g(x)$$

for all $x \in (a, b)$.

Definition 2.17. Given a ring A, for any element $a \in A$, if there is some element $b \in A$ such that $b \neq 0$ and $ab = 0$, then we say that a is a *zero divisor*. A ring A is an *integral domain* (or an *entire ring*) if $0 \neq 1$, A is commutative, and $ab = 0$ implies that $a = 0$ or $b = 0$, for all $a, b \in A$. In other words, an integral domain is a nontrivial commutative ring with no zero divisors besides 0.

Example 2.7.

1. The rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, are integral domains.

2. The ring $\mathbb{R}[X]$ of polynomials in one variable with real coefficients is an integral domain.

3. For any positive integer, $n \in \mathbb{N}$, we have the ring $\mathbb{Z}/n\mathbb{Z}$. Observe that if n is composite, then this ring has zero-divisors. For example, if $n = 4$, then we have

$$2 \cdot 2 \equiv 0 \pmod{4}.$$

The reader should prove that $\mathbb{Z}/n\mathbb{Z}$ is an integral domain iff n is prime (use Proposition 2.17).

4. If d is a square-free positive integer and if $d \geq 2$, the ring $\mathbb{Z}[\sqrt{d}]$ is an integral domain. Similarly, if $d \geq 1$ is a square-free positive integer, the ring $\mathbb{Z}[\sqrt{-d}]$ is an integral domain. Finding the invertible elements of these rings is a very interesting problem.

5. The ring of $n \times n$ matrices $\mathrm{M}_n(\mathbb{R})$ has zero divisors.

A homomorphism between rings is a mapping preserving addition and multiplication (and 0 and 1).

Definition 2.18. Given two rings A and B, a *homomorphism between A and B* is a function $h\colon A \to B$ satisfying the following conditions for all $x, y \in A$:

$$\begin{aligned} h(x+y) &= h(x) + h(y) \\ h(xy) &= h(x)h(y) \\ h(0) &= 0 \\ h(1) &= 1. \end{aligned}$$

Actually, because B is a group under addition, $h(0) = 0$ follows from

$$h(x+y) = h(x) + h(y).$$

Example 2.8.

1. If A is a ring, for any integer $n \in \mathbb{Z}$, for any $a \in A$, we define $n \cdot a$ by

$$n \cdot a = \underbrace{a + \cdots + a}_{n}$$

if $n \geq 0$ (with $0 \cdot a = 0$) and

$$n \cdot a = -(-n) \cdot a$$

if $n < 0$. Then, the map $h\colon \mathbb{Z} \to A$ given by

$$h(n) = n \cdot 1_A$$

is a ring homomorphism (where 1_A is the multiplicative identity of A).

2. Given any real $\lambda \in \mathbb{R}$, the evaluation map $\eta_\lambda\colon \mathbb{R}[X] \to \mathbb{R}$ defined by

$$\eta_\lambda(f(X)) = f(\lambda)$$

for every polynomial $f(X) \in \mathbb{R}[X]$ is a ring homomorphism.

Definition 2.19. A ring homomorphism $h\colon A \to B$ is an *isomorphism* iff there is a ring homomorphism $g\colon B \to A$ such that $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. An isomorphism from a ring to itself is called an *automorphism*.

As in the case of a group isomorphism, the homomorphism g is unique and denoted by h^{-1}, and it is easy to show that a bijective ring homomorphism $h\colon A \to B$ is an isomorphism.

Definition 2.20. Given a ring A, a subset A' of A is a *subring* of A if A' is a subgroup of A (under addition), is closed under multiplication, and contains 1.

For example, we have the following sequence in which every ring on the left of an inlcusion sign is a subring of the ring on the right of the inclusion sign:

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}.$$

The ring $\mathbb{Z}$ is a subring of both $\mathbb{Z}[\sqrt{d}]$ and $\mathbb{Z}[\sqrt{-d}]$, the ring $\mathbb{Z}[\sqrt{d}]$ is a subring of $\mathbb{R}$ and the ring $\mathbb{Z}[\sqrt{-d}]$ is a subring of $\mathbb{C}$.

If $h\colon A \to B$ is a homomorphism of rings, then it is easy to show for any subring A', the image $h(A')$ is a subring of B, and for any subring B' of B, the inverse image $h^{-1}(B')$ is a subring of A.

As for groups, the *kernel* of a ring homomorphism $h\colon A \to B$ is defined by

$$\mathrm{Ker}\ h = \{a \in A \mid h(a) = 0\}.$$

Just as in the case of groups, we have the following criterion for the injectivity of a ring homomorphism. The proof is identical to the proof for groups.

Proposition 2.18. *If $h\colon A \to B$ is a homomorphism of rings, then $h\colon A \to B$ is injective iff* $\mathrm{Ker}\ h = \{0\}$. *(We also write* $\mathrm{Ker}\ h = (0)$.*)*

The kernel of a ring homomorphism is an abelian subgroup of the additive group A, but in general it is not a subring of A, because it may not contain the multiplicative identity element 1. However, it satisfies the following closure property under multiplication:

$$ab \in \mathrm{Ker}\ h \quad \text{and} \quad ba \in \mathrm{Ker}\ h \quad \text{for all } a \in \mathrm{Ker}\ h \text{ and all } b \in A.$$

This is because if $h(a) = 0$, then for all $b \in A$ we have

$$h(ab) = h(a)h(b) = 0h(b) = 0 \quad \text{and} \quad h(ba) = h(b)h(a) = h(b)0 = 0.$$

Definition 2.21. Given a ring A, an additive subgroup $\mathfrak{I}$ of A satisfying the property below

$$ab \in \mathfrak{I} \quad \text{and} \quad ba \in \mathfrak{I} \quad \text{for all } a \in \mathfrak{I} \text{ and all } b \in A \tag{$*_{\text{ideal}}$}$$

is called a *two-sided ideal*. If A is a commutative ring, we simply say an *ideal*.

It turns out that for any ring A and any two-sided ideal $\mathfrak{I}$, the set $A/\mathfrak{I}$ of additive cosets $a + \mathfrak{I}$ (with $a \in A$) is a ring called a *quotient ring*. Then we have the following analog of Proposition 2.12, also called the *first isomorphism theorem*.

Proposition 2.19. *Given a homomorphism of rings $h\colon A \to B$, the rings $A/\mathrm{Ker}\ h$ and $\mathrm{Im}\ h = h(A)$ are isomorphic.*

A field is a commutative ring K for which $K - \{0\}$ is a group under multiplication.

Definition 2.22. A set K is a *field* if it is a ring and the following properties hold:

(F1) $0 \neq 1$;

(F2) For every $a \in K$, if $a \neq 0$, then a has an inverse w.r.t. $*$;

(F3) $*$ is commutative.

Let $K^* = K - \{0\}$. Observe that (F1) and (F2) are equivalent to the fact that K^* is a group w.r.t. $*$ with identity element 1. If $*$ is not commutative but (F1) and (F2) hold, we say that we have a *skew field* (or *noncommutative field*).

Note that we are assuming that the operation $*$ of a field is commutative. This convention is not universally adopted, but since $*$ will be commutative for most fields we will encounter, we may as well include this condition in the definition.

Example 2.9.

1. The rings $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields.

2. The set of (formal) fractions $f(X)/g(X)$ of polynomials $f(X), g(X) \in \mathbb{R}[X]$, where $g(X)$ is not the null polynomial, is a field.

3. The ring $\mathcal{C}(a, b)$ of continuous functions $f\colon (a, b) \to \mathbb{R}$ such that $f(x) \neq 0$ for all $x \in (a, b)$ is a field.

4. Using Proposition 2.17, it is easy to see that the ring $\mathbb{Z}/p\mathbb{Z}$ is a field iff p is prime.

5. If d is a square-free positive integer and if $d \geq 2$, the set
$$\mathbb{Q}(\sqrt{d}) = \{a + b\sqrt{d} \in \mathbb{R} \mid a, b \in \mathbb{Q}\}$$
is a field. If $z = a + b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ and $\overline{z} = a - b\sqrt{d}$, then it is easy to check that if $z \neq 0$, then $z^{-1} = \overline{z}/(z\overline{z})$.

6. Similarly, If $d \geq 1$ is a square-free positive integer, the set of complex numbers
$$\mathbb{Q}(\sqrt{-d}) = \{a + ib\sqrt{d} \in \mathbb{C} \mid a, b \in \mathbb{Q}\}$$
is a field. If $z = a + ib\sqrt{d} \in \mathbb{Q}(\sqrt{-d})$ and $\overline{z} = a - ib\sqrt{d}$, then it is easy to check that if $z \neq 0$, then $z^{-1} = \overline{z}/(z\overline{z})$.

Definition 2.23. A homomorphism $h\colon K_1 \to K_2$ between two fields K_1 and K_2 is just a homomorphism between the rings K_1 and K_2.

However, because K_1^* and K_2^* are groups under multiplication, a homomorphism of fields must be injective.

Proof. First, observe that for any $x \neq 0$,

$$1 = h(1) = h(xx^{-1}) = h(x)h(x^{-1})$$

and

$$1 = h(1) = h(x^{-1}x) = h(x^{-1})h(x),$$

so $h(x) \neq 0$ and

$$h(x^{-1}) = h(x)^{-1}.$$

But then, if $h(x) = 0$, we must have $x = 0$. Consequently, h is injective. $\square$

Definition 2.24. A field homomorphism $h\colon K_1 \to K_2$ is an *isomorphism* iff there is a homomorphism $g\colon K_2 \to K_1$ such that $g \circ f = \mathrm{id}_{K_1}$ and $f \circ g = \mathrm{id}_{K_2}$. An isomorphism from a field to itself is called an *automorphism*.

Then, just as in the case of rings, g is unique and denoted by h^{-1}, and a bijective field homomorphism $h\colon K_1 \to K_2$ is an isomorphism.

Definition 2.25. Since every homomorphism $h\colon K_1 \to K_2$ between two fields is injective, the image $f(K_1)$ of K_1 is a subfield of K_2. We say that K_2 is an *extension* of K_1.

For example, $\mathbb{R}$ is an extension of $\mathbb{Q}$ and $\mathbb{C}$ is an extension of $\mathbb{R}$. The fields $\mathbb{Q}(\sqrt{d})$ and $\mathbb{Q}(\sqrt{-d})$ are extensions of $\mathbb{Q}$, the field $\mathbb{R}$ is an extension of $\mathbb{Q}(\sqrt{d})$ and the field $\mathbb{C}$ is an extension of $\mathbb{Q}(\sqrt{-d})$.

Definition 2.26. A field K is said to be *algebraically closed* if every polynomial $p(X)$ with coefficients in K has some root in K; that is, there is some $a \in K$ such that $p(a) = 0$.

It can be shown that every field K has some minimal extension Ω which is algebraically closed, called an *algebraic closure* of K. For example, $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$. The algebraic closure of $\mathbb{Q}$ is called the *field of algebraic numbers*. This field consists of all complex numbers that are zeros of a polynomial with coefficients in $\mathbb{Q}$.

Definition 2.27. Given a field K and an automorphism $h\colon K \to K$ of K, it is easy to check that the set

$$\mathrm{Fix}(h) = \{a \in K \mid h(a) = a\}$$

of elements of K fixed by h is a subfield of K called the *field fixed by h*.

For example, if $d \geq 2$ is square-free, then the map $c\colon \mathbb{Q}(\sqrt{d}) \to \mathbb{Q}(\sqrt{d})$ given by

$$c(a + b\sqrt{d}) = a - b\sqrt{d}$$

is an automorphism of $\mathbb{Q}(\sqrt{d})$, and $\mathrm{Fix}(c) = \mathbb{Q}$.

If K is a field, we have the ring homomorphism $h\colon \mathbb{Z} \to K$ given by $h(n) = n \cdot 1$. If h is injective, then K contains a copy of $\mathbb{Z}$, and since it is a field, it contains a copy of $\mathbb{Q}$. In this case, we say that K has *characteristic* 0. If h is not injective, then $h(\mathbb{Z})$ is a subring of K, and thus an integral domain, the kernel of h is a subgroup of $\mathbb{Z}$, which by Proposition 2.15 must be of the form $p\mathbb{Z}$ for some $p \geq 1$. By the first isomorphism theorem, $h(\mathbb{Z})$ is isomorphic to $\mathbb{Z}/p\mathbb{Z}$ for some $p \geq 1$. But then, p must be prime since $\mathbb{Z}/p\mathbb{Z}$ is an integral domain iff it is a field iff p is prime. The prime p is called the *characteristic* of K, and we also says that K is of *finite characteristic*.

Definition 2.28. If K is a field, then either

(1) $n \cdot 1 \neq 0$ for all integer $n \geq 1$, in which case we say that K has *characteristic* 0, or

(2) There is some smallest prime number p such that $p \cdot 1 = 0$ called the *characteristic* of K, and we say K is of *finite characteristic*.

A field K of characteristic 0 contains a copy of $\mathbb{Q}$, thus is infinite. As we will see in Section 8.10, a finite field has nonzero characteristic p. However, there are infinite fields of nonzero characteristic.

Part I

Linear Algebra

Chapter 3

Vector Spaces, Bases, Linear Maps

3.1 Motivations: Linear Combinations, Linear Independence and Rank

In linear optimization problems, we often encounter systems of linear equations. For example, consider the problem of solving the following system of three linear equations in the three variables $x_1, x_2, x_3 \in \mathbb{R}$:

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 1 \\ 2x_1 + x_2 + x_3 &= 2 \\ x_1 - 2x_2 - 2x_3 &= 3. \end{aligned}$$

One way to approach this problem is introduce the "vectors" u, v, w, and b, given by

$$u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad v = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix} \qquad w = \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

and to write our linear system as

$$x_1 u + x_2 v + x_3 w = b.$$

In the above equation, we used implicitly the fact that a vector z can be multiplied by a scalar $\lambda \in \mathbb{R}$, where

$$\lambda z = \lambda \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \lambda z_1 \\ \lambda z_2 \\ \lambda z_3 \end{pmatrix},$$

and two vectors y and and z can be added, where

$$y + z = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} y_1 + z_1 \\ y_2 + z_2 \\ y_3 + z_3 \end{pmatrix}.$$

Also, given a vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

we define the *additive inverse* $-x$ of x (pronounced minus x) as

$$-x = \begin{pmatrix} -x_1 \\ -x_2 \\ -x_3 \end{pmatrix}.$$

Observe that $-x = (-1)x$, the scalar multiplication of x by -1.

The set of all vectors with three components is denoted by $\mathbb{R}^{3\times 1}$. The reason for using the notation $\mathbb{R}^{3\times 1}$ rather than the more conventional notation $\mathbb{R}^3$ is that the elements of $\mathbb{R}^{3\times 1}$ are *column vectors*; they consist of three rows and a single column, which explains the superscript 3×1. On the other hand, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ consists of all triples of the form (x_1, x_2, x_3), with $x_1, x_2, x_3 \in \mathbb{R}$, and these are *row vectors*. However, there is an obvious bijection between $\mathbb{R}^{3\times 1}$ and $\mathbb{R}^3$ and they are usually identified. For the sake of clarity, in this introduction, we will denote the set of column vectors with n components by $\mathbb{R}^{n\times 1}$.

An expression such as

$$x_1 u + x_2 v + x_3 w$$

where u, v, w are vectors and the x_is are scalars (in $\mathbb{R}$) is called a *linear combination*. Using this notion, the problem of solving our linear system

$$x_1 u + x_2 v + x_3 w = b.$$

is equivalent to *determining whether b can be expressed as a linear combination of u, v, w.*

Now if the vectors u, v, w are *linearly independent*, which means that there is *no* triple $(x_1, x_2, x_3) \neq (0, 0, 0)$ such that

$$x_1 u + x_2 v + x_3 w = 0_3,$$

it can be shown that *every* vector in $\mathbb{R}^{3\times 1}$ can be written as a linear combination of u, v, w. Here, 0_3 is the *zero vector*

$$0_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

It is customary to abuse notation and to write 0 instead of 0_3. This rarely causes a problem because in most cases, whether 0 denotes the scalar zero or the zero vector can be inferred from the context.

In fact, every vector $z \in \mathbb{R}^{3\times 1}$ can be written *in a unique way* as a linear combination

$$z = x_1 u + x_2 v + x_3 w.$$

This is because if

$$z = x_1u + x_2v + x_3w = y_1u + y_2v + y_3w,$$

then by using our (linear!) operations on vectors, we get

$$(y_1 - x_1)u + (y_2 - x_2)v + (y_3 - x_3)w = 0,$$

which implies that

$$y_1 - x_1 = y_2 - x_2 = y_3 - x_3 = 0,$$

by linear independence. Thus,

$$y_1 = x_1, \quad y_2 = x_2, \quad y_3 = x_3,$$

which shows that z has a unique expression as a linear combination, as claimed. Then our equation

$$x_1u + x_2v + x_3w = b$$

has a *unique solution*, and indeed, we can check that

$$\begin{aligned} x_1 &= 1.4 \\ x_2 &= -0.4 \\ x_3 &= -0.4 \end{aligned}$$

is the solution.

But then, *how do we determine that some vectors are linearly independent?*

One answer is to compute a numerical quantity $\det(u, v, w)$, called the *determinant* of (u, v, w), and to check that it is nonzero. In our case, it turns out that

$$\det(u, v, w) = \begin{vmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{vmatrix} = 15,$$

which confirms that u, v, w are linearly independent.

Other methods, which are much better for systems with a large number of variables, consist of computing an LU-decomposition or a QR-decomposition, or an SVD of the *matrix* consisting of the three columns u, v, w,

$$A = \begin{pmatrix} u & v & w \end{pmatrix} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix}.$$

If we form the vector of unknowns

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

then our linear combination $x_1u + x_2v + x_3w$ can be written in matrix form as

$$x_1u + x_2v + x_3w = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

so our linear system is expressed by

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

or more concisely as

$$Ax = b.$$

Now what if the vectors u, v, w are *linearly dependent*? For example, if we consider the vectors

$$u = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad v = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} \qquad w = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix},$$

we see that

$$u - v = w,$$

a nontrivial *linear dependence*. It can be verified that u and v are still linearly independent. Now for our problem

$$x_1u + x_2v + x_3w = b$$

it must be the case that b can be expressed as linear combination of u and v. However, it turns out that u, v, b are linearly independent (one way to see this is to compute the determinant $\det(u, v, b) = -6$), so b cannot be expressed as a linear combination of u and v and thus, our system has *no* solution.

If we change the vector b to

$$b = \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix},$$

then

$$b = u + v,$$

and so the system

$$x_1u + x_2v + x_3w = b$$

has the solution

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 0.$$

Actually, since $w = u - v$, the above system is equivalent to

$$(x_1 + x_3)u + (x_2 - x_3)v = b,$$

and because u and v are linearly independent, the unique solution in $x_1 + x_3$ and $x_2 - x_3$ is

$$\begin{aligned} x_1 + x_3 &= 1 \\ x_2 - x_3 &= 1, \end{aligned}$$

which yields an infinite number of solutions parameterized by x_3, namely

$$\begin{aligned} x_1 &= 1 - x_3 \\ x_2 &= 1 + x_3. \end{aligned}$$

In summary, a 3×3 linear system may have a unique solution, no solution, or an infinite number of solutions, depending on the linear independence (and dependence) or the vectors u, v, w, b. This situation can be generalized to any $n \times n$ system, and even to any $n \times m$ system (n equations in m variables), as we will see later.

The point of view where our linear system is expressed in matrix form as $Ax = b$ stresses the fact that the map $x \mapsto Ax$ is a *linear transformation*. This means that

$$A(\lambda x) = \lambda(Ax)$$

for all $x \in \mathbb{R}^{3\times 1}$ and all $\lambda \in \mathbb{R}$ and that

$$A(u + v) = Au + Av,$$

for all $u, v \in \mathbb{R}^{3\times 1}$. We can view the matrix A as a way of expressing a linear map from $\mathbb{R}^{3\times 1}$ to $\mathbb{R}^{3\times 1}$ and solving the system $Ax = b$ amounts to determining whether b belongs to the image of this linear map.

Given a 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

whose columns are three vectors denoted A^1, A^2, A^3, and given any vector $x = (x_1, x_2, x_3)$, we defined the product Ax as the linear combination

$$Ax = x_1 A^1 + x_2 A^2 + x_3 A^3 = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix}.$$

The common pattern is that the ith coordinate of Ax is given by a certain kind of product called an *inner product*, of a *row vector*, the ith row of A, times the *column vector* x:

$$\begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3.$$

More generally, given any two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, their *inner product* denoted $x \cdot y$, or $\langle x, y \rangle$, is the number

$$x \cdot y = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n x_i y_i.$$

Inner products play a very important role. First, we quantity

$$\|x\|_2 = \sqrt{x \cdot x} = (x_1^2 + \cdots + x_n^2)^{1/2}$$

is a generalization of the length of a vector, called the *Euclidean norm*, or *ℓ^2-norm*. Second, it can be shown that we have the inequality

$$|x \cdot y| \leq \|x\| \, \|y\| \, ,$$

so if $x, y \neq 0$, the ratio $(x \cdot y)/(\|x\| \, \|y\|)$ can be viewed as the cosine of an angle, the angle between x and y. In particular, if $x \cdot y = 0$ then the vectors x and y make the angle $\pi/2$, that is, they are *orthogonal*. The (square) matrices Q that preserve the inner product, in the sense that $\langle Qx, Qy \rangle = \langle x, y \rangle$ for all $x, y \in \mathbb{R}^n$, also play a very important role. They can be thought of as generalized rotations.

Returning to matrices, if A is an $m \times n$ matrix consisting of n columns $A^1, \ldots, A^n$ (in $\mathbb{R}^m$), and B is a $n \times p$ matrix consisting of p columns $B^1, \ldots, B^p$ (in $\mathbb{R}^n$) we can form the p vectors (in $\mathbb{R}^m$)

$$AB^1, \ldots, AB^p.$$

These p vectors constitute the $m \times p$ matrix denoted AB, whose jth column is AB^j. But we know that the ith coordinate of AB^j is the inner product of the ith row of A by the jth column of B,

$$\begin{pmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{pmatrix} \cdot \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{pmatrix} = \sum_{k=1}^n a_{ik} b_{kj}.$$

Thus we have defined a multiplication operation on matrices, namely if $A = (a_{ik})$ is a $m \times n$ matrix and if $B = (b_{jk})$ if $n \times p$ matrix, then their product AB is the $m \times n$ matrix whose entry on the ith row and the jth column is given by the inner product of the ith row of A by the jth column of B,

$$(AB)_{ij} = \sum_{k=1}^n a_{ik} b_{kj}.$$

Beware that unlike the multiplication of real (or complex) numbers, if A and B are two $n \times n$ matrices, in general, $AB \neq BA$.

Suppose that A is an $n \times n$ matrix and that we are trying to solve the linear system

$$Ax = b,$$

with $b \in \mathbb{R}^n$. Suppose we can find an $n \times n$ matrix B such that

$$BA^i = e_i, \quad i = 1, \ldots, n,$$

with $e_i = (0, \ldots, 0, 1, 0 \ldots, 0)$, where the only nonzero entry is 1 in the ith slot. If we form the $n \times n$ matrix

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

called the *identity matrix*, whose ith column is e_i, then the above is equivalent to

$$BA = I_n.$$

If $Ax = b$, then multiplying both sides on the left by B, we get

$$B(Ax) = Bb.$$

But is is easy to see that $B(Ax) = (BA)x = I_n x = x$, so we must have

$$x = Bb.$$

We can verify that $x = Bb$ is indeed a solution, because it can be shown that

$$A(Bb) = (AB)b = I_n b = b.$$

What is not obvious is that $BA = I_n$ implies $AB = I_n$, but this is indeed provable. The matrix B is usually denoted A^{-1} and called the *inverse* of A. It can be shown that it is the unique matrix such that

$$AA^{-1} = A^{-1}A = I_n.$$

If a square matrix A has an inverse, then we say that it is *invertible* or *nonsingular*, otherwise we say that it is *singular*. We will show later that a square matrix is invertible iff its columns are linearly independent iff its determinant is nonzero.

In summary, if A is a square invertible matrix, then the linear system $Ax = b$ has the unique solution $x = A^{-1}b$. In practice, this is not a good way to solve a linear system because computing A^{-1} is too expensive. A practical method for solving a linear system is Gaussian elimination, discussed in Chapter 8. Other practical methods for solving a linear system

$Ax = b$ make use of a factorization of A (QR decomposition, SVD decomposition), using orthogonal matrices defined next.

Given an $m \times n$ matrix $A = (a_{kl})$, the $n \times m$ matrix $A^\top = (a^\top_{ij})$ whose ith row is the ith column of A, which means that $a^\top_{ij} = a_{ji}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, is called the *transpose* of A. An $n \times n$ matrix Q such that

$$QQ^\top = Q^\top Q = I_n$$

is called an *orthogonal matrix*. Equivalently, the inverse Q^{-1} of an orthogonal matrix Q is equal to its transpose $Q^\top$. Orthogonal matrices play an important role. Geometrically, they correspond to linear transformation that preserve length. A major result of linear algebra states that every $m \times n$ matrix A can be written as

$$A = V\Sigma U^\top,$$

where V is an $m\times m$ orthogonal matrix, U is an $n\times n$ orthogonal matrix, and Σ is an $m\times n$ matrix whose only nonzero entries are nonnegative diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, where $p = \min(m, n)$, called the *singular values* of A. The factorization $A = V\Sigma U^\top$ is called a *singular decomposition* of A, or *SVD*.

The SVD can be used to "solve" a linear system $Ax = b$ where A is an $m \times n$ matrix, even when this system has no solution. This may happen when there are more equations that variables $(m > n)$, in which case the system is overdetermined.

Of course, there is no miracle, an unsolvable system has no solution. But we can look for a *good approximate solution*, namely a vector x that minimizes some measure of the error $Ax - b$. Legendre and Gauss used $\|Ax - b\|_2^2$, which is the squared Euclidean norm of the error. This quantity is differentiable, and it turns out that there is a unique vector x^+ of minimum Euclidean norm that minimizes $\|Ax - b\|_2^2$. Furthermore, x^+ is given by the expression $x^+ = A^+b$, where A^+ is the *pseudo-inverse* of A, and A^+ can be computed from an SVD $A = V\Sigma U^\top$ of A. Indeed, $A^+ = U\Sigma^+V^\top$, where Σ^+ is the matrix obtained from Σ by replacing every positive singular value σ_i by its inverse σ_i^{-1}, leaving all zero entries intact, and transposing.

Instead of searching for the vector of least Euclidean norm minimizing $\|Ax - b\|_2^2$, we can add the penalty term $K\|x\|_2^2$ (for some positive $K > 0$) to $\|Ax - b\|_2^2$ and minimize the quantity $\|Ax - b\|_2^2 + K\|x\|_2^2$. This approach is called *ridge regression*. It turns out that there is a unique minimizer x^+ given by $x^+ = (A^\top A + KI_n)^{-1}A^\top b$, as shown in the second volume.

Another approach is to replace the penalty term $K\|x\|_2^2$ by $K\|x\|_1$, where $\|x\|_1 = |x_1| + \cdots + |x_n|$ (the ℓ^1-norm of x). The remarkable fact is that the minimizers x of $\|Ax - b\|_2^2 + K\|x\|_1$ tend to be *sparse*, which means that many components of x are equal to zero. This approach known as *lasso* is popular in machine learning and will be discussed in the second volume.

Another important application of the SVD is *principal component analysis* (or *PCA*), an important tool in data analysis.

Yet another fruitful way of interpreting the resolution of the system $Ax = b$ is to view this problem as an intersection problem. Indeed, each of the equations

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 1 \\ 2x_1 + x_2 + x_3 &= 2 \\ x_1 - 2x_2 - 2x_3 &= 3 \end{aligned}$$

defines a subset of $\mathbb{R}^3$ which is actually a *plane*. The first equation

$$x_1 + 2x_2 - x_3 = 1$$

defines the plane H_1 passing through the three points $(1, 0, 0), (0, 1/2, 0), (0, 0, -1)$, on the coordinate axes, the second equation

$$2x_1 + x_2 + x_3 = 2$$

defines the plane H_2 passing through the three points $(1, 0, 0), (0, 2, 0), (0, 0, 2)$, on the coordinate axes, and the third equation

$$x_1 - 2x_2 - 2x_3 = 3$$

defines the plane H_3 passing through the three points $(3, 0, 0), (0, -3/2, 0), (0, 0, -3/2)$, on the coordinate axes. See Figure 3.1.

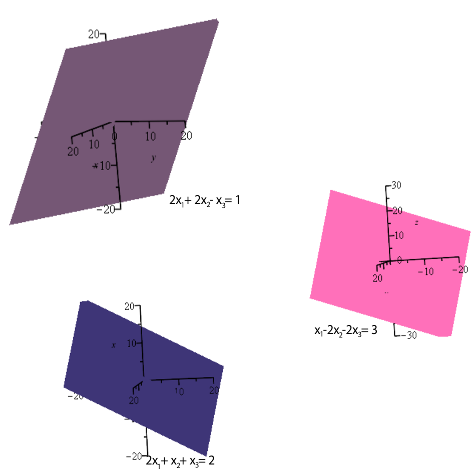


Figure 3.1: The planes defined by the preceding linear equations.

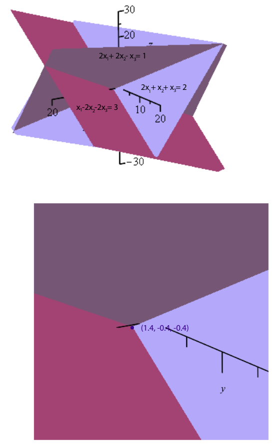


Figure 3.2: The solution of the system is the point in common with each of the three planes.

The intersection $H_i \cap H_j$ of any two distinct planes H_i and H_j is a line, and the intersection $H_1 \cap H_2 \cap H_3$ of the three planes consists of the single point $(1.4, -0.4, -0.4)$, as illustrated in Figure 3.2.

The planes corresponding to the system

$$\begin{aligned}
x_1 + 2x_2 - x_3 &= 1 \\
2x_1 + x_2 + x_3 &= 2 \\
x_1 - x_2 + 2x_3 &= 3,
\end{aligned}$$

are illustrated in Figure 3.3.

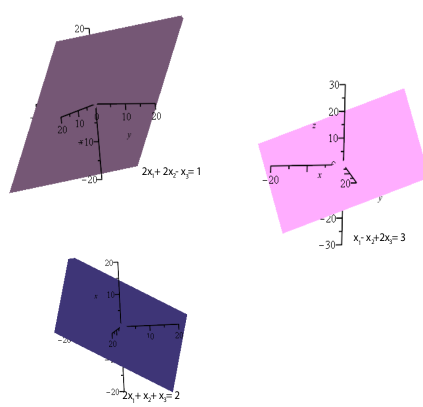


Figure 3.3: The planes defined by the equations $x_1 + 2x_2 - x_3 = 1$, $2x_1 + x_2 + x_3 = 2$, and $x_1 - x_2 + 2x_3 = 3$.

This system has no solution since there is no point simultaneously contained in all three planes; see Figure 3.4.

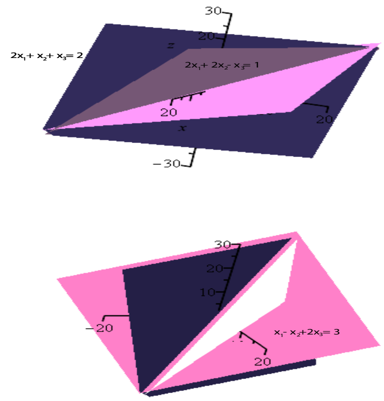


Figure 3.4: The linear system $x_1 + 2x_2 - x_3 = 1$, $2x_1 + x_2 + x_3 = 2$, $x_1 - x_2 + 2x_3 = 3$ has no solution.

Finally, the planes corresponding to the system

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 3 \\ 2x_1 + x_2 + x_3 &= 3 \\ x_1 - x_2 + 2x_3 &= 0, \end{aligned}$$

are illustrated in Figure 3.5.

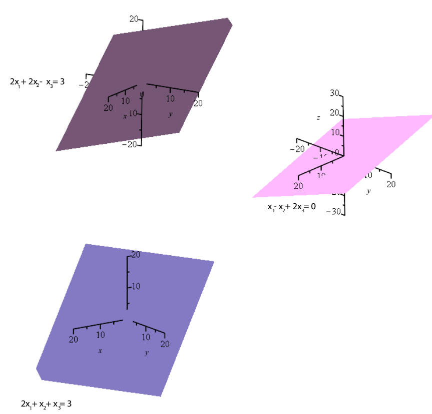


Figure 3.5: The planes defined by the equations $x_1 + 2x_2 - x_3 = 3$, $2x_1 + x_2 + x_3 = 3$, and $x_1 - x_2 + 2x_3 = 0$.

This system has infinitely many solutions, given parametrically by $(1 - x_3, 1 + x_3, x_3)$. Geometrically, this is a line common to all three planes; see Figure 3.6.

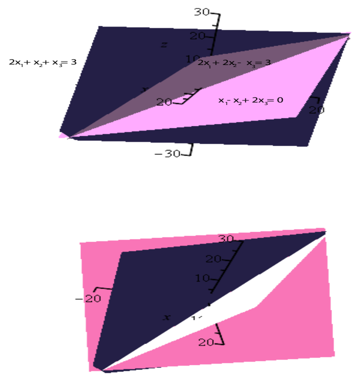


Figure 3.6: The linear system $x_1 + 2x_2 - x_3 = 3$, $2x_1 + x_2 + x_3 = 3$, $x_1 - x_2 + 2x_3 = 0$ has the red line common to all three planes.

Under the above interpretation, observe that we are focusing on the *rows* of the matrix A, rather than on its *columns*, as in the previous interpretations.

Another great example of a real-world problem where linear algebra proves to be very effective is the problem of *data compression*, that is, of representing a very large data set using a much smaller amount of storage.

Typically the data set is represented as an $m \times n$ matrix A where each row corresponds to an n-dimensional data point and typically, $m \geq n$. In most applications, the data are not independent so the rank of A is a lot smaller than $\min\{m, n\}$, and the the goal of *low-rank decomposition* is to factor A as the product of two matrices B and C, where B is a $m \times k$ matrix and C is a $k \times n$ matrix, with $k \ll \min\{m, n\}$ (here, $\ll$ means "much smaller than"):

$$\begin{pmatrix} \\ \\ \quad\quad A \quad\quad \\ m \times n \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ B \\ m \times k \\ \\ \\ \end{pmatrix} \begin{pmatrix} \quad\quad C \quad\quad \\ k \times n \end{pmatrix}$$

Now it is generally too costly to find an exact factorization as above, so we look for a low-rank matrix A' which is a "good" *approximation* of A. In order to make this statement precise, we need to define a mechanism to determine how close two matrices are. This can be done using *matrix norms*, a notion discussed in Chapter 9. The norm of a matrix A is a nonnegative real number $\|A\|$ which behaves a lot like the absolute value $|x|$ of a real number x. Then our goal is to find some low-rank matrix A' that minimizes the norm

$$\|A - A'\|^2,$$

over all matrices A' of rank at most k, for some given $k \ll \min\{m, n\}$.

Some advantages of a low-rank approximation are:

1. Fewer elements are required to represent A; namely, $k(m+n)$ instead of mn. Thus less storage and fewer operations are needed to reconstruct A.

2. Often, the process for obtaining the decomposition exposes the underlying structure of the data. Thus, it may turn out that "most" of the significant data are concentrated along some directions called *principal directions*.

Low-rank decompositions of a set of data have a multitude of applications in engineering, including computer science (especially computer vision), statistics, and machine learning. As we will see later in Chapter 23, the *singular value decomposition* (SVD) provides a very satisfactory solution to the low-rank approximation problem. Still, in many cases, the data sets are so large that another ingredient is needed: *randomization*. However, as a first step, linear algebra often yields a good initial solution.

We will now be more precise as to what kinds of operations are allowed on vectors. In the early 1900, the notion of a *vector space* emerged as a convenient and unifying framework for working with "linear" objects and we will discuss this notion in the next few sections.

3.2 Vector Spaces

For every $n \geq 1$, let $\mathbb{R}^n$ be the set of n-tuples $x = (x_1, \ldots, x_n)$. Addition can be extended to $\mathbb{R}^n$ as follows:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

We can also define an operation $\cdot \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ as follows:

$$\lambda \cdot (x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n).$$

The resulting algebraic structure has some interesting properties, those of a vector space.

However, keep in mind that vector spaces are not just algebraic objects; they are also geometric objects.

Vector spaces are defined as follows.

Definition 3.1. Given a field K (with addition $+$ and multiplication $*$), a *vector space over K* (or *K-vector space*) is a set E (of vectors) together with two operations $+_E \colon E \times E \to E$ (called *vector addition*),and $\cdot_E \colon K \times E \to E$ (called *scalar multiplication*) satisfying the following conditions for all $\alpha, \beta \in K$ and all $u, v \in E$;

(V0) E is an abelian group w.r.t. $+_E$, with identity element 0_E;

(V1) $\alpha \cdot_E (u +_E v) = (\alpha \cdot_E u) +_E (\alpha \cdot_E v)$;

(V2) $(\alpha + \beta) \cdot_E u = (\alpha \cdot_E u) +_E (\beta \cdot_E u)$;

(V3) $(\alpha * \beta) \cdot_E u = \alpha \cdot_E (\beta \cdot_E u)$;

(V4) $1 \cdot_E u = u$.

In (V1), $+$ denotes addition in the field K, and in (V3), $*$ denotes multiplication in K.

Given $\alpha, \beta \in K$ and $v \in E$, for simplicity of notation, the element $\alpha \cdot_E v$ is denoted by αv and $\alpha * \beta$ by $\alpha\beta$. The field K is often called the field of scalars.

Unless specified otherwise or unless we are dealing with several different fields, in the rest of this chapter, we assume that all K-vector spaces are defined with respect to a fixed field K. Thus, we will refer to a K-vector space simply as a vector space. In most cases, the field K will be the field $\mathbb{R}$ of reals.

From (V0), a vector space always contains the null vector 0_E, and thus is nonempty. From (V1), we get $\alpha \cdot_E 0_E = 0_E$, and $\alpha \cdot_E (-_E v) = -_E(\alpha \cdot_E v)$. From (V2), we get $0 \cdot_E v = 0_E$, and $(-\alpha) \cdot_E v = -_E(\alpha \cdot_E v)$.

Another important consequence of the axioms is the following fact:

Proposition 3.1. *For any $u \in E$ and any $\lambda \in K$, if $\lambda \neq 0$ and $\lambda \cdot_E u = 0_E$, then $u = 0_E$.*

Proof. Indeed, since $\lambda \neq 0$, it has a multiplicative inverse λ^{-1}, so from $\lambda \cdot_E u = 0_E$, we get

$$\lambda^{-1} \cdot_E (\lambda \cdot_E u) = \lambda^{-1} \cdot_E 0_E.$$

However, we just observed that $\lambda^{-1} \cdot_E 0_E = 0_E$, and from (V3) and (V4), we have

$$\lambda^{-1} \cdot_E (\lambda \cdot_E u) = (\lambda^{-1}\lambda) \cdot_E u = 1 \cdot_E u = u,$$

and we deduce that $u = 0_E$. $\square$

Remark: One may wonder whether axiom (V4) is really needed. Could it be derived from the other axioms? The answer is **no**. For example, one can take $E = \mathbb{R}^n$ and define $\cdot \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$\lambda \cdot (x_1, \ldots, x_n) = (0, \ldots, 0)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. Axioms (V0)–(V3) are all satisfied, but (V4) fails. Less trivial examples can be given using the notion of a basis, which has not been defined yet.

The field K itself can be viewed as a vector space over itself, addition of vectors being addition in the field, and multiplication by a scalar being multiplication in the field.

One quickly realizes that although it is of course important to distinguish between addition $+_E$ in E with addition $+$ in the field K, the zero vector 0_E in E with zero 0 in the field, and multiplication $\cdot_E$ by a scalar (in the field K) with multiplication $*$ (or $\cdot$) in the field, using all these subscripts becomes a notational burden. Thus, unless confusion may arise, we denote addition in E and K by the same symbol $+$, the zero in E and in K by 0, and scalar multiplication by $\cdot$ (we even write αu instead of $\alpha \cdot u$). Systematic use of different symbols for denoting scalars (say $\alpha, \beta, \gamma, \delta, \lambda$) and vectors say (u, v, x, y, z) is usually enough to avoid confusion. To avoid confusing zero in E and zero in K, we can use $\mathbf{0}$ for the zero vector.

Example 3.1.

1. The fields $\mathbb{R}$ and $\mathbb{C}$ are vector spaces over $\mathbb{R}$.

2. The groups $\mathbb{R}^n$ and $\mathbb{C}^n$ are vector spaces over $\mathbb{R}$, with scalar multiplication given by

$$\lambda(x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n),$$

for any $\lambda \in \mathbb{R}$ and with $(x_1, \ldots, x_n) \in \mathbb{R}^n$ or $(x_1, \ldots, x_n) \in \mathbb{C}^n$, and $\mathbb{C}^n$ is a vector space over $\mathbb{C}$ with scalar multiplication as above, but with $\lambda \in \mathbb{C}$.

3. The ring $\mathbb{R}[X]_n$ of polynomials of degree at most n with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]_n$ of polynomials of degree at most n with complex coefficients is a vector space over $\mathbb{C}$, with scalar multiplication $\lambda \cdot P(X)$ of a polynomial

$$P(X) = a_0 X^m + a_1 X^{m-1} + \cdots + a_{m-1} X + a_m$$

(with $a_i \in \mathbb{R}$ or $a_i \in \mathbb{C}$) by the scalar λ (in $\mathbb{R}$ or $\mathbb{C}$), with $m \leq n$, given by

$$\lambda \cdot P(X) = \lambda a_0 X^m + \lambda a_1 X^{m-1} + \cdots + \lambda a_{m-1} X + \lambda a_m.$$

4. The ring $\mathbb{R}[X]$ of all polynomials with real coefficients is a vector space over $\mathbb{R}$, and the ring $\mathbb{C}[X]$ of all polynomials with complex coefficients is a vector space over $\mathbb{C}$, with the same scalar multiplication as above.

5. The ring of $n \times n$ matrices $\mathrm{M}_n(\mathbb{R})$ is a vector space over $\mathbb{R}$.

6. The ring of $m \times n$ matrices $\mathrm{M}_{m,n}(\mathbb{R})$ is a vector space over $\mathbb{R}$.

7. The ring $\mathcal{C}(a, b)$ of continuous functions $f \colon (a, b) \to \mathbb{R}$ is a vector space over $\mathbb{R}$, with the scalar multiplication λf of a function $f \colon (a, b) \to \mathbb{R}$ by a scalar $\lambda \in \mathbb{R}$ given by

$$(\lambda f)(x) = \lambda f(x), \qquad \text{for all } x \in (a, b).$$

8. A very important example of vector space is the set of linear maps between two vector spaces to be defined in Section 11.1. Here is an example that will prepare us for the vector space of linear maps. Let X be any nonempty set and let E be a vector space. The set of all functions $f\colon X \to E$ can be made into a vector space as follows: Given any two functions $f\colon X \to E$ and $g\colon X \to E$, let $(f+g)\colon X \to E$ be defined such that

$$(f+g)(x) = f(x) + g(x)$$

for all $x \in X$, and for every $\lambda \in \mathbb{R}$, let $\lambda f\colon X \to E$ be defined such that

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in X$. The axioms of a vector space are easily verified.

Let E be a vector space. We would like to define the important notions of linear combination and linear independence.

Before defining these notions, we need to discuss a strategic choice which, depending how it is settled, may reduce or increase headaches in dealing with notions such as linear combinations and linear dependence (or independence). The issue has to do with using sets of vectors versus sequences of vectors.

3.3 Indexed Families; the Sum Notation $\sum_{i\in I} a_i$

Our experience tells us that *it is preferable to use sequences of vectors*; even better, indexed families of vectors. (We are not alone in having opted for sequences over sets, and we are in good company; for example, Artin [7], Axler [10], and Lang [109] use sequences. Nevertheless, some prominent authors such as Lax [113] use sets. We leave it to the reader to conduct a survey on this issue.)

Given a set A, recall that a *sequence* is an ordered n-tuple $(a_1, \ldots, a_n) \in A^n$ of elements from A, for some natural number n. The elements of a sequence need not be distinct and the order is important. For example, (a_1, a_2, a_1) and (a_2, a_1, a_1) are two distinct sequences in A^3. Their underlying set is $\{a_1, a_2\}$.

What we just defined are *finite* sequences, which can also be viewed as functions from $\{1, 2, \ldots, n\}$ to the set A; the ith element of the sequence $(a_1, \ldots, a_n)$ is the image of i under the function. This viewpoint is fruitful, because it allows us to define (countably) infinite sequences as functions $s\colon \mathbb{N} \to A$. But then, why limit ourselves to ordered sets such as $\{1, \ldots, n\}$ or $\mathbb{N}$ as index sets?

The main role of the index set is to tag each element uniquely, and the order of the tags is not crucial, although convenient. Thus, it is natural to define the notion of indexed family.

Definition 3.2. Given a set A, an *I-indexed family* of elements of A, for short a *family*, is a function $a\colon I \to A$ where I is any set viewed as an index set. Since the function a is determined by its graph

$$\{(i, a(i)) \mid i \in I\},$$

the family a can be viewed as the set of pairs $a = \{(i, a(i)) \mid i \in I\}$. For notational simplicity, we write a_i instead of $a(i)$, and denote the family $a = \{(i, a(i)) \mid i \in I\}$ by $(a_i)_{i\in I}$.

For example, if $I = \{r, g, b, y\}$ and $A = \mathbb{N}$, the set of pairs

$$a = \{(r, 2), (g, 3), (b, 2), (y, 11)\}$$

is an indexed family. The element 2 appears twice in the family with the two distinct tags r and b.

When the indexed set I is totally ordered, a family $(a_i)_{i\in I}$ is often called an *I-sequence*. Interestingly, sets can be viewed as special cases of families. Indeed, a set A can be viewed as the A-indexed family $\{(a, a) \mid a \in I\}$ corresponding to the identity function.

Remark: An indexed family should not be confused with a multiset. Given any set A, a *multiset* is a similar to a set, except that elements of A may occur more than once. For example, if $A = \{a, b, c, d\}$, then $\{a, a, a, b, c, c, d, d\}$ is a multiset. Each element appears with a certain multiplicity, but the order of the elements does not matter. For example, a has multiplicity 3. Formally, a multiset is a function $s\colon A \to \mathbb{N}$, or equivalently a set of pairs $\{(a, i) \mid a \in A\}$. Thus, a multiset is an A-indexed family of elements from $\mathbb{N}$, but not a $\mathbb{N}$-indexed family, since distinct elements may have the same multiplicity (such as c an d in the example above). *An indexed family is a generalization of a sequence, but a multiset is a generalization of a set.*

We also need to take care of an annoying technicality, which is to define sums of the form $\sum_{i\in I} a_i$, where I is any *finite* index set and $(a_i)_{i\in I}$ is a family of elements in some set A equiped with a binary operation $+\colon A \times A \to A$ which is associative (Axiom (G1)) and commutative. This will come up when we define linear combinations.

The issue is that the binary operation $+$ only tells us how to compute $a_1 + a_2$ for two elements of A, but it does not tell us what is the sum of three of more elements. For example, how should $a_1 + a_2 + a_3$ be defined?

What we have to do is to define $a_1+a_2+a_3$ by using a sequence of steps each involving two elements, and there are two possible ways to do this: $a_1 + (a_2 + a_3)$ and $(a_1 + a_2) + a_3$. If our operation $+$ is not associative, these are different values. If it associative, then $a_1+(a_2+a_3) = (a_1 + a_2) + a_3$, but then there are still six possible permutations of the indices $1, 2, 3$, and if $+$ is not commutative, these values are generally different. If our operation is commutative, then all six permutations have the same value. *Thus, if $+$ is associative and commutative, it seems intuitively clear that a sum of the form $\sum_{i\in I} a_i$ does not depend on the order of the operations used to compute it.*

This is indeed the case, but a rigorous proof requires induction, and such a proof is surprisingly involved. Readers may accept without proof the fact that sums of the form $\sum_{i\in I} a_i$ are indeed well defined, and jump directly to Definition 3.3. For those who want to see the gory details, here we go.

First, we define sums $\sum_{i\in I} a_i$, where I is a finite sequence of distinct natural numbers, say $I = (i_1, \ldots, i_m)$. If $I = (i_1, \ldots, i_m)$ with $m \geq 2$, we denote the sequence $(i_2, \ldots, i_m)$ by $I - \{i_1\}$. We proceed by induction on the size m of I. Let

$$\begin{aligned}
\sum_{i\in I} a_i &= a_{i_1}, \quad \text{if } m = 1,\\
\sum_{i\in I} a_i &= a_{i_1} + \Big(\sum_{i\in I-\{i_1\}} a_i\Big), \quad \text{if } m > 1.
\end{aligned}$$

For example, if $I = (1, 2, 3, 4)$, we have

$$\sum_{i\in I} a_i = a_1 + (a_2 + (a_3 + a_4)).$$

If the operation $+$ is not associative, the grouping of the terms matters. For instance, in general

$$a_1 + (a_2 + (a_3 + a_4)) \neq (a_1 + a_2) + (a_3 + a_4).$$

However, if the operation $+$ is associative, the sum $\sum_{i\in I} a_i$ should not depend on the grouping of the elements in I, as long as their order is preserved. For example, if $I = (1, 2, 3, 4, 5)$, $J_1 = (1, 2)$, and $J_2 = (3, 4, 5)$, we expect that

$$\sum_{i\in I} a_i = \Big(\sum_{j\in J_1} a_j\Big) + \Big(\sum_{j\in J_2} a_j\Big).$$

This indeed the case, as we have the following proposition.

Proposition 3.2. *Given any nonempty set A equipped with an associative binary operation $+\colon A \times A \to A$, for any nonempty finite sequence I of distinct natural numbers and for any partition of I into p nonempty sequences $I_{k_1}, \ldots, I_{k_p}$, for some nonempty sequence $K = (k_1, \ldots, k_p)$ of distinct natural numbers such that $k_i < k_j$ implies that $\alpha < \beta$ for all $\alpha \in I_{k_i}$ and all $\beta \in I_{k_j}$, for every sequence $(a_i)_{i\in I}$ of elements in A, we have*

$$\sum_{\alpha\in I} a_\alpha = \sum_{k\in K}\Big(\sum_{\alpha\in I_k} a_\alpha\Big).$$

Proof. We proceed by induction on the size n of I.

If $n = 1$, then we must have $p = 1$ and $I_{k_1} = I$, so the proposition holds trivially.

Next, assume $n > 1$. If $p = 1$, then $I_{k_1} = I$ and the formula is trivial, so assume that $p \geq 2$ and write $J = (k_2, \ldots, k_p)$. There are two cases.

Case 1. The sequence I_{k_1} has a single element, say β, which is the first element of I. In this case, write C for the sequence obtained from I by deleting its first element β. By definition,

$$\sum_{\alpha\in I} a_\alpha = a_\beta + \left(\sum_{\alpha\in C} a_\alpha\right),$$

and

$$\sum_{k\in K}\left(\sum_{\alpha\in I_k} a_\alpha\right) = a_\beta + \left(\sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right)\right).$$

Since $|C| = n - 1$, by the induction hypothesis, we have

$$\left(\sum_{\alpha\in C} a_\alpha\right) = \sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right),$$

which yields our identity.

Case 2. The sequence I_{k_1} has at least two elements. In this case, let β be the first element of I (and thus of I_{k_1}), let I' be the sequence obtained from I by deleting its first element β, let I'_{k_1} be the sequence obtained from I_{k_1} by deleting its first element β, and let $I'_{k_i} = I_{k_i}$ for $i = 2, \ldots, p$. Recall that $J = (k_2, \ldots, k_p)$ and $K = (k_1, \ldots, k_p)$. The sequence I' has $n - 1$ elements, so by the induction hypothesis applied to I' and the I'_{k_i}, we get

$$\sum_{\alpha\in I'} a_\alpha = \sum_{k\in K}\left(\sum_{\alpha\in I'_k} a_\alpha\right) = \left(\sum_{\alpha\in I'_{k_1}} a_\alpha\right) + \left(\sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right)\right).$$

If we add the lefthand side to a_β, by definition we get

$$\sum_{\alpha\in I} a_\alpha.$$

If we add the righthand side to a_β, using associativity and the definition of an indexed sum, we get

$$\begin{aligned}
a_\beta + \left(\left(\sum_{\alpha\in I'_{k_1}} a_\alpha\right) + \left(\sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right)\right)\right) &= \left(a_\beta + \left(\sum_{\alpha\in I'_{k_1}} a_\alpha\right)\right) + \left(\sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right)\right)\\
&= \left(\sum_{\alpha\in I_{k_1}} a_\alpha\right) + \left(\sum_{j\in J}\left(\sum_{\alpha\in I_j} a_\alpha\right)\right)\\
&= \sum_{k\in K}\left(\sum_{\alpha\in I_k} a_\alpha\right),
\end{aligned}$$

as claimed. $\square$

If $I = (1, \ldots, n)$, we also write $\sum_{i=1}^{n} a_i$ instead of $\sum_{i \in I} a_i$. Since $+$ is associative, Proposition 3.2 shows that the sum $\sum_{i=1}^{n} a_i$ is independent of the grouping of its elements, which justifies the use the notation $a_1 + \cdots + a_n$ (without any parentheses).

If we also assume that our associative binary operation on A is commutative, then we can show that the sum $\sum_{i \in I} a_i$ does not depend on the ordering of the index set I.

Proposition 3.3. *Given any nonempty set A equipped with an associative and commutative binary operation $+\colon A \times A \to A$, for any two nonempty finite sequences I and J of distinct natural numbers such that J is a permutation of I (in other words, the underlying sets of I and J are identical), for every sequence $(a_i)_{i \in I}$ of elements in A, we have*

$$\sum_{\alpha \in I} a_\alpha = \sum_{\alpha \in J} a_\alpha.$$

Proof. We proceed by induction on the number p of elements in I. If $p = 1$, we have $I = J$ and the proposition holds trivially.

If $p > 1$, to simplify notation, assume that $I = (1, \ldots, p)$ and that J is a permutation $(i_1, \ldots, i_p)$ of I. First, assume that $2 \leq i_1 \leq p-1$, let J' be the sequence obtained from J by deleting i_1, I' be the sequence obtained from I by deleting i_1, and let $P = (1, 2, \ldots, i_1 - 1)$ and $Q = (i_1 + 1, \ldots, p-1, p)$. Observe that the sequence I' is the concatenation of the sequences P and Q. By the induction hypothesis applied to J' and I', and then by Proposition 3.2 applied to I' and its partition (P, Q), we have

$$\sum_{\alpha \in J'} a_\alpha = \sum_{\alpha \in I'} a_\alpha = \left(\sum_{i=1}^{i_1 - 1} a_i \right) + \left(\sum_{i = i_1 + 1}^{p} a_i \right).$$

If we add the lefthand side to a_{i_1}, by definition we get

$$\sum_{\alpha \in J} a_\alpha.$$

If we add the righthand side to a_{i_1}, we get

$$a_{i_1} + \left(\left(\sum_{i=1}^{i_1 - 1} a_i \right) + \left(\sum_{i = i_1 + 1}^{p} a_i \right) \right).$$

Using associativity, we get

$$a_{i_1} + \left(\left(\sum_{i=1}^{i_1 - 1} a_i \right) + \left(\sum_{i = i_1 + 1}^{p} a_i \right) \right) = \left(a_{i_1} + \left(\sum_{i=1}^{i_1 - 1} a_i \right) \right) + \left(\sum_{i = i_1 + 1}^{p} a_i \right),$$

then using associativity and commutativity several times (more rigorously, using induction on $i_1 - 1$), we get

$$\begin{aligned}\left(a_{i_1} + \left(\sum_{i=1}^{i_1-1} a_i\right)\right) + \left(\sum_{i=i_1+1}^{p} a_i\right) &= \left(\sum_{i=1}^{i_1-1} a_i\right) + a_{i_1} + \left(\sum_{i=i_1+1}^{p} a_i\right)\\ &= \sum_{i=1}^{p} a_i,\end{aligned}$$

as claimed.

The cases where $i_1 = 1$ or $i_1 = p$ are treated similarly, but in a simpler manner since either $P = ()$ or $Q = ()$ (where $()$ denotes the empty sequence). $\square$

Having done all this, we can now make sense of sums of the form $\sum_{i\in I} a_i$, for any finite indexed set I and any family $a = (a_i)_{i\in I}$ of elements in A, where A is a set equipped with a binary operation $+$ which is associative and commutative.

Indeed, since I is finite, it is in bijection with the set $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$, and any total ordering $\preceq$ on I corresponds to a permutation $I_{\preceq}$ of $\{1, \ldots, n\}$ (where we identify a permutation with its image). For any total ordering $\preceq$ on I, we define $\sum_{i\in I,\preceq} a_i$ as

$$\sum_{i\in I,\preceq} a_i = \sum_{j\in I_{\preceq}} a_j.$$

Then for any other total ordering $\preceq'$ on I, we have

$$\sum_{i\in I,\preceq'} a_i = \sum_{j\in I_{\preceq'}} a_j,$$

and since $I_{\preceq}$ and $I_{\preceq'}$ are different permutations of $\{1, \ldots, n\}$, by Proposition 3.3, we have

$$\sum_{j\in I_{\preceq}} a_j = \sum_{j\in I_{\preceq'}} a_j.$$

Therefore, the sum $\sum_{i\in I,\preceq} a_i$ does not depend on the total ordering on I. We define *the* sum $\sum_{i\in I} a_i$ as the common value $\sum_{i\in I,\preceq} a_i$ for all total orderings $\preceq$ of I.

Here are some examples with $A = \mathbb{R}$:

1. If $I = \{1, 2, 3\}$, $a = \{(1, 2), (2, -3), (3, \sqrt{2})\}$, then $\sum_{i\in I} a_i = 2 - 3 + \sqrt{2} = -1 + \sqrt{2}$.
2. If $I = \{2, 5, 7\}$, $a = \{(2, 2), (5, -3), (7, \sqrt{2})\}$, then $\sum_{i\in I} a_i = 2 - 3 + \sqrt{2} = -1 + \sqrt{2}$.
3. If $I = \{r, g, b\}$, $a = \{(r, 2), (g, -3), (b, 1)\}$, then $\sum_{i\in I} a_i = 2 - 3 + 1 = 0$.

3.4 Linear Independence, Subspaces

One of the most useful properties of vector spaces is that they possess bases. What this means is that in every vector space E, there is some set of vectors, $\{e_1, \ldots, e_n\}$, such that *every* vector $v \in E$ can be written as a linear combination,

$$v = \lambda_1 e_1 + \cdots + \lambda_n e_n,$$

of the e_i, for some scalars, $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Furthermore, the n-tuple, $(\lambda_1, \ldots, \lambda_n)$, as above is unique.

This description is fine when E has a finite basis, $\{e_1, \ldots, e_n\}$, but this is not always the case! For example, the vector space of real polynomials, $\mathbb{R}[X]$, does not have a finite basis but instead it has an infinite basis, namely

$$1,\ X,\ X^2,\ \ldots, X^n,\ \ldots$$

One might wonder if it is possible for a vector space to have bases of different sizes, or even to have a finite basis as well as an infinite basis. We will see later on that this is not possible; all bases of a vector space have the same number of elements (cardinality), which is called the *dimension* of the space. However, we have the following problem: If a vector space has an infinite basis, $\{e_1, e_2, \ldots,\}$, how do we define linear combinations? Do we allow linear combinations

$$\lambda_1 e_1 + \lambda_2 e_2 + \cdots$$

with infinitely many nonzero coefficients?

If we allow linear combinations with infinitely many nonzero coefficients, then we have to make sense of these sums and this can only be done reasonably if we define such a sum as the limit of the sequence of vectors, $s_1, s_2, \ldots, s_n, \ldots$, with $s_1 = \lambda_1 e_1$ and

$$s_{n+1} = s_n + \lambda_{n+1} e_{n+1}.$$

But then, how do we define such limits? Well, we have to define some topology on our space, by means of a norm, a metric or some other mechanism. This can indeed be done and this is what Banach spaces and Hilbert spaces are all about but this seems to require a lot of machinery.

A way to avoid limits is to restrict our attention to linear combinations involving only *finitely many* vectors. We may have an infinite supply of vectors but we only form linear combinations involving finitely many nonzero coefficients. Technically, this can be done by introducing *families of finite support*. This gives us the ability to manipulate families of scalars indexed by some fixed infinite set and yet to be treat these families as if they were finite.

With these motivations in mind, given a set A, recall that an I*-indexed family* $(a_i)_{i \in I}$ of elements of A (for short, a *family*) is a function $a\colon I \to A$, or equivalently a set of pairs $\{(i, a_i) \mid i \in I\}$. We agree that when $I = \emptyset$, $(a_i)_{i \in I} = \emptyset$. A family $(a_i)_{i \in I}$ is finite if I is finite.

Remark: When considering a family $(a_i)_{i\in I}$, there is no reason to assume that I is ordered. The crucial point is that every element of the family is uniquely indexed by an element of I. Thus, unless specified otherwise, we do not assume that the elements of an index set are ordered.

If A is an abelian group with identity 0, we say that a family $(a_i)_{i\in I}$ has *finite support* if $a_i = 0$ for all $i \in I - J$, where J is a finite subset of I (the support of the family).

Given two disjoint sets I and J, the union of two families $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$, denoted as $(u_i)_{i\in I} \cup (v_j)_{j\in J}$, is the family $(w_k)_{k\in(I\cup J)}$ defined such that $w_k = u_k$ if $k \in I$, and $w_k = v_k$ if $k \in J$. Given a family $(u_i)_{i\in I}$ and any element v, we denote by $(u_i)_{i\in I} \cup_k (v)$ the family $(w_i)_{i\in I\cup\{k\}}$ defined such that, $w_i = u_i$ if $i \in I$, and $w_k = v$, where k is any index such that $k \notin I$. Given a family $(u_i)_{i\in I}$, a *subfamily* of $(u_i)_{i\in I}$ is a family $(u_j)_{j\in J}$ where J is any subset of I.

In this chapter, unless specified otherwise, *is assumed that all families of scalars have finite support*.

Definition 3.3. Let E be a vector space. A vector $v \in E$ is a *linear combination of a family $(u_i)_{i\in I}$ of elements of E* iff there is a family $(\lambda_i)_{i\in I}$ of scalars in K such that

$$v = \sum_{i\in I} \lambda_i u_i.$$

When $I = \emptyset$, we stipulate that $v = 0$. (By Proposition 3.3, sums of the form $\sum_{i\in I} \lambda_i u_i$ are well defined.) We say that a family $(u_i)_{i\in I}$ is *linearly independent* iff for every family $(\lambda_i)_{i\in I}$ of scalars in K,

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{implies that} \quad \lambda_i = 0 \ \text{for all } i \in I.$$

Equivalently, a family $(u_i)_{i\in I}$ is *linearly dependent* iff there is some family $(\lambda_i)_{i\in I}$ of scalars in K such that

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{and} \quad \lambda_j \neq 0 \ \text{for some } j \in I.$$

We agree that when $I = \emptyset$, the family $\emptyset$ is linearly independent.

Observe that defining linear combinations for families of vectors rather than for sets of vectors has the advantage that *the vectors being combined need not be distinct*. For example, for $I = \{1, 2, 3\}$ and the families (u, v, u) and $(\lambda_1, \lambda_2, \lambda_1)$, the linear combination

$$\sum_{i\in I} \lambda_i u_i = \lambda_1 u + \lambda_2 v + \lambda_1 u$$

makes sense. Using sets of vectors in the definition of a linear combination does not allow such linear combinations; this is too restrictive.

Unravelling Definition 3.3, a family $(u_i)_{i\in I}$ is linearly dependent iff either I consists of a single element, say i, and $u_i = 0$, or $|I| \geq 2$ and some u_j in the family can be expressed as a linear combination of the other vectors in the family. Indeed, in the second case, there is some family $(\lambda_i)_{i\in I}$ of scalars in K such that

$$\sum_{i\in I} \lambda_i u_i = 0 \quad \text{and} \quad \lambda_j \neq 0 \text{ for some } j \in I,$$

and since $|I| \geq 2$, the set $I - \{j\}$ is nonempty and we get

$$u_j = \sum_{i\in(I-\{j\})} -\lambda_j^{-1}\lambda_i u_i.$$

Observe that one of the reasons for defining linear dependence for families of vectors rather than for sets of vectors is that our definition allows multiple occurrences of a vector. This is important because a matrix may contain identical columns, and we would like to say that these columns are linearly dependent. The definition of linear dependence for sets does not allow us to do that.

The above also shows that a family $(u_i)_{i\in I}$ is linearly independent iff either $I = \emptyset$, or I consists of a single element i and $u_i \neq 0$, or $|I| \geq 2$ and no vector u_j in the family can be expressed as a linear combination of the other vectors in the family.

When I is nonempty, if the family $(u_i)_{i\in I}$ is linearly independent, note that $u_i \neq 0$ for all $i \in I$. Otherwise, if $u_i = 0$ for some $i \in I$, then we get a nontrivial linear dependence $\sum_{i\in I} \lambda_i u_i = 0$ by picking any nonzero λ_i and letting $\lambda_k = 0$ for all $k \in I$ with $k \neq i$, since $\lambda_i 0 = 0$. If $|I| \geq 2$, we must also have $u_i \neq u_j$ for all $i, j \in I$ with $i \neq j$, since otherwise we get a nontrivial linear dependence by picking $\lambda_i = \lambda$ and $\lambda_j = -\lambda$ for any nonzero λ, and letting $\lambda_k = 0$ for all $k \in I$ with $k \neq i, j$.

Thus, the definition of linear independence implies that *a nontrivial linearly independent family is actually a set*. This explains why certain authors choose to define linear independence for sets of vectors. The problem with this approach is that linear dependence, which is the logical negation of linear independence, is then only defined for sets of vectors. However, as we pointed out earlier, it is really desirable to define linear dependence for families allowing multiple occurrences of the same vector.

Example 3.2.

1. Any two distinct scalars $\lambda, \mu \neq 0$ in K are linearly dependent.

2. In $\mathbb{R}^3$, the vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ are linearly independent. See Figure 3.7.

3. In $\mathbb{R}^4$, the vectors $(1, 1, 1, 1)$, $(0, 1, 1, 1)$, $(0, 0, 1, 1)$, and $(0, 0, 0, 1)$ are linearly independent.

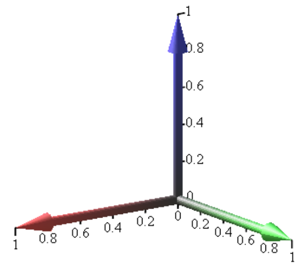


Figure 3.7: A visual (arrow) depiction of the red vector $(1, 0, 0)$, the green vector $(0, 1, 0)$, and the blue vector $(0, 0, 1)$ in $\mathbb{R}^3$.

4. In $\mathbb{R}^2$, the vectors $u = (1, 1)$, $v = (0, 1)$ and $w = (2, 3)$ are linearly dependent, since
$$w = 2u + v.$$
See Figure 3.8.

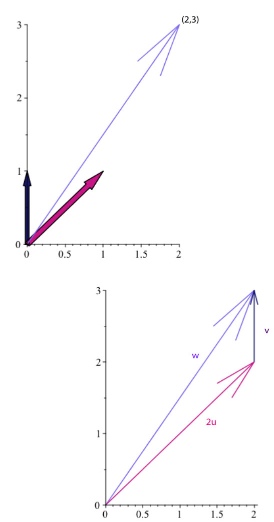


Figure 3.8: A visual (arrow) depiction of the pink vector $u = (1, 1)$, the dark purple vector $v = (0, 1)$, and the vector sum $w = 2u + v$.

When I is finite, we often assume that it is the set $I = \{1, 2, \ldots, n\}$. In this case, we denote the family $(u_i)_{i\in I}$ as $(u_1, \ldots, u_n)$.

The notion of a subspace of a vector space is defined as follows.

Definition 3.4. Given a vector space E, a subset F of E is a *linear subspace* (or *subspace*) of E iff F is nonempty and $\lambda u + \mu v \in F$ for all $u, v \in F$, and all $\lambda, \mu \in K$.

It is easy to see that a subspace F of E is indeed a vector space, since the restriction of $+_E\colon E \times E \to E$ to $F \times F$ is indeed a function $+_F\colon F \times F \to F$, and the restriction of $\cdot_E\colon K \times E \to E$ to $K \times F$ is indeed a function $\cdot_F\colon K \times F \to F$.

Since a subspace F is nonempty, if we pick any vector $u \in F$ and if we let $\lambda = \mu = 0$, then $\lambda u + \mu u = 0u + 0u = 0_E + 0_E = 0_E$, so *every subspace contains the zero vector* 0_E.

The following facts also hold. The proof is left as an exercise.

Proposition 3.4.

(1) The intersection of any family (even infinite) of subspaces of a vector space E is a subspace.

(2) Let F be any subspace of a vector space E. For any nonempty finite index set I, if $(u_i)_{i\in I}$ is any family of vectors $u_i \in F$ and $(\lambda_i)_{i\in I}$ is any family of scalars, then $\sum_{i\in I} \lambda_i u_i \in F$.

The subspace $\{0_E\}$ will be denoted by (0_E), or even (0) (with a mild abuse of notation).

Example 3.3.

1. In $\mathbb{R}^2$, the set of vectors $u = (x, y)$ such that

$$x + y = 0$$

is the subspace illustrated by Figure 3.9.

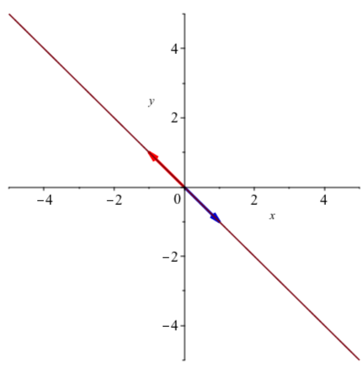


Figure 3.9: The subspace $x + y = 0$ is the line through the origin with slope -1. It consists of all vectors of the form $\lambda(-1, 1)$.

2. In $\mathbb{R}^3$, the set of vectors $u = (x, y, z)$ such that

$$x + y + z = 0$$

is the subspace illustrated by Figure 3.10.

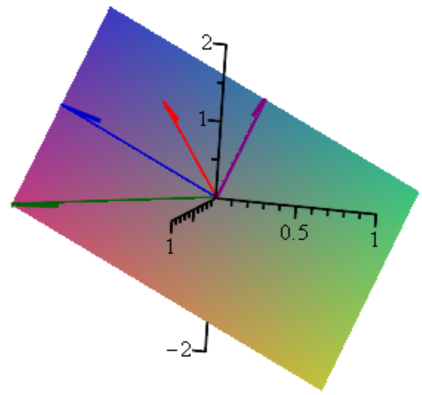


Figure 3.10: The subspace $x+y+z=0$ is the plane through the origin with normal $(1,1,1)$.

3. For any $n \geq 0$, the set of polynomials $f(X) \in \mathbb{R}[X]$ of degree at most n is a subspace of $\mathbb{R}[X]$.

4. The set of upper triangular $n \times n$ matrices is a subspace of the space of $n \times n$ matrices.

Proposition 3.5. *Given any vector space E, if S is any nonempty subset of E, then the smallest subspace $\langle S \rangle$ (or* Span(S)*) of E containing S is the set of all (finite) linear combinations of elements from S.*

Proof. We prove that the set $\mathrm{Span}(S)$ of all linear combinations of elements of S is a subspace of E, leaving as an exercise the verification that every subspace containing S also contains $\mathrm{Span}(S)$.

First, $\mathrm{Span}(S)$ is nonempty since it contains S (which is nonempty). If $u = \sum_{i\in I} \lambda_i u_i$ and $v = \sum_{j\in J} \mu_j v_j$ are any two linear combinations in $\mathrm{Span}(S)$, for any two scalars $\lambda, \mu \in K$,

$$\begin{aligned}
\lambda u + \mu v &= \lambda \sum_{i\in I} \lambda_i u_i + \mu \sum_{j\in J} \mu_j v_j \\
&= \sum_{i\in I} \lambda\lambda_i u_i + \sum_{j\in J} \mu\mu_j v_j \\
&= \sum_{i\in I-J} \lambda\lambda_i u_i + \sum_{i\in I\cap J} (\lambda\lambda_i + \mu\mu_i) u_i + \sum_{j\in J-I} \mu\mu_j v_j,
\end{aligned}$$

which is a linear combination with index set $I \cup J$, and thus $\lambda u + \mu v \in \mathrm{Span}(S)$, which proves that $\mathrm{Span}(S)$ is a subspace. $\square$

One might wonder what happens if we add extra conditions to the coefficients involved in forming linear combinations. Here are three natural restrictions which turn out to be important (as usual, we assume that our index sets are finite):

(1) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which

$$\sum_{i\in I} \lambda_i = 1.$$

These are called *affine combinations*. One should realize that every linear combination $\sum_{i\in I} \lambda_i u_i$ can be viewed as an affine combination. For example, if k is an index not in I, if we let $J = I \cup \{k\}$, $u_k = 0$, and $\lambda_k = 1 - \sum_{i\in I} \lambda_i$, then $\sum_{j\in J} \lambda_j u_j$ is an affine combination and

$$\sum_{i\in I} \lambda_i u_i = \sum_{j\in J} \lambda_j u_j.$$

However, we get new spaces. For example, in $\mathbb{R}^3$, the set of all affine combinations of the three vectors $e_1 = (1, 0, 0), e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$, is the plane passing through these three points. Since it does not contain $0 = (0, 0, 0)$, it is not a linear subspace.

(2) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which

$$\lambda_i \geq 0, \quad \text{for all } i \in I.$$

These are called *positive* (or *conic*) *combinations*. It turns out that positive combinations of families of vectors are *cones*. They show up naturally in convex optimization.

(3) Consider combinations $\sum_{i\in I} \lambda_i u_i$ for which we require (1) *and* (2), that is

$$\sum_{i\in I} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0 \quad \text{for all } i \in I.$$

These are called *convex combinations*. Given any finite family of vectors, the set of all convex combinations of these vectors is a *convex polyhedron*. Convex polyhedra play a very important role in convex optimization.

Remark: The notion of linear combination can also be defined for infinite index sets I. To ensure that a sum $\sum_{i\in I} \lambda_i u_i$ makes sense, we restrict our attention to families of finite support.

Definition 3.5. Given any field K, a family of scalars $(\lambda_i)_{i\in I}$ has *finite support* if $\lambda_i = 0$ for all $i \in I - J$, for some finite subset J of I.

If $(\lambda_i)_{i\in I}$ is a family of scalars of finite support, for any vector space E over K, for any (possibly infinite) family $(u_i)_{i\in I}$ of vectors $u_i \in E$, we define the linear combination $\sum_{i\in I} \lambda_i u_i$ as the finite linear combination $\sum_{j\in J} \lambda_j u_j$, where J is any finite subset of I such that $\lambda_i = 0$ for all $i \in I - J$. In general, results stated for finite families also hold for families of finite support.

3.5 Bases of a Vector Space

Given a vector space E, given a family $(v_i)_{i\in I}$, the subset V of E consisting of the null vector 0 and of all linear combinations of $(v_i)_{i\in I}$ is easily seen to be a subspace of E. The family $(v_i)_{i\in I}$ is an economical way of representing the entire subspace V, but such a family would be even nicer if it was not redundant. Subspaces having such an "efficient" generating family (called a basis) play an important role and motivate the following definition.

Definition 3.6. Given a vector space E and a subspace V of E, a family $(v_i)_{i\in I}$ of vectors $v_i \in V$ *spans* V or *generates* V iff for every $v \in V$, there is some family $(\lambda_i)_{i\in I}$ of scalars in K such that

$$v = \sum_{i\in I} \lambda_i v_i.$$

We also say that the elements of $(v_i)_{i\in I}$ are *generators* of V and that V is *spanned by* $(v_i)_{i\in I}$, or *generated by* $(v_i)_{i\in I}$. If a subspace V of E is generated by a finite family $(v_i)_{i\in I}$, we say that V is *finitely generated*. A family $(u_i)_{i\in I}$ that spans V and is linearly independent is called a *basis* of V.

Example 3.4.

1. In $\mathbb{R}^3$, the vectors $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$, illustrated in Figure 3.9, form a basis.

2. The vectors $(1,1,1,1), (1,1,-1,-1), (1,-1,0,0), (0,0,1,-1)$ form a basis of $\mathbb{R}^4$ known as the *Haar basis*. This basis and its generalization to dimension 2^n are crucial in wavelet theory.

3. In the subspace of polynomials in $\mathbb{R}[X]$ of degree at most n, the polynomials $1, X, X^2, \ldots, X^n$ form a basis.

4. The *Bernstein polynomials* $\binom{n}{k}(1-X)^{n-k}X^k$ for $k = 0,\ldots,n$, also form a basis of that space. These polynomials play a major role in the theory of *spline curves*.

The first key result of linear algebra is that *every vector space E has a basis.* We begin with a crucial lemma which formalizes the mechanism for building a basis incrementally.

Lemma 3.6. *Given a linearly independent family $(u_i)_{i\in I}$ of elements of a vector space E, if $v \in E$ is not a linear combination of $(u_i)_{i\in I}$, then the family $(u_i)_{i\in I} \cup_k (v)$ obtained by adding v to the family $(u_i)_{i\in I}$ is linearly independent (where $k \notin I$).*

Proof. Assume that $\mu v + \sum_{i\in I} \lambda_i u_i = 0$, for any family $(\lambda_i)_{i\in I}$ of scalars in K. If $\mu \neq 0$, then μ has an inverse (because K is a field), and thus we have $v = -\sum_{i\in I}(\mu^{-1}\lambda_i)u_i$, showing that v is a linear combination of $(u_i)_{i\in I}$ and contradicting the hypothesis. Thus, $\mu = 0$. But then, we have $\sum_{i\in I} \lambda_i u_i = 0$, and since the family $(u_i)_{i\in I}$ is linearly independent, we have $\lambda_i = 0$ for all $i \in I$. $\square$

The next theorem holds in general, but the proof is more sophisticated for vector spaces that do not have a finite set of generators. *Thus, in this chapter, we only prove the theorem for finitely generated vector spaces.*

Theorem 3.7. *Given any finite family $S = (u_i)_{i \in I}$ generating a vector space E and any linearly independent subfamily $L = (u_j)_{j \in J}$ of S (where $J \subseteq I$), there is a basis B of E such that $L \subseteq B \subseteq S$.*

Proof. Consider the set of linearly independent families B such that $L \subseteq B \subseteq S$. Since this set is nonempty and finite, it has some maximal element (that is, a subfamily $B = (u_h)_{h \in H}$ of S with $H \subseteq I$ of maximum cardinality), say $B = (u_h)_{h \in H}$. We claim that B generates E. Indeed, if B does not generate E, then there is some $u_p \in S$ that is not a linear combination of vectors in B (since S generates E), with $p \notin H$. Then by Lemma 3.6, the family $B' = (u_h)_{h \in H \cup \{p\}}$ is linearly independent, and since $L \subseteq B \subset B' \subseteq S$, this contradicts the maximality of B. Thus, B is a basis of E such that $L \subseteq B \subseteq S$. $\square$

Remark: Theorem 3.7 also holds for vector spaces that are not finitely generated. In this case, the problem is to guarantee the existence of a maximal linearly independent family B such that $L \subseteq B \subseteq S$. The existence of such a maximal family can be shown using Zorn's lemma, see Appendix C and the references given there.

A situation where the full generality of Theorem 3.7 is needed is the case of the vector space $\mathbb{R}$ over the field of coefficients $\mathbb{Q}$. The numbers 1 and $\sqrt{2}$ are linearly independent over $\mathbb{Q}$, so according to Theorem 3.7, the linearly independent family $L = (1, \sqrt{2})$ can be extended to a basis B of $\mathbb{R}$. Since $\mathbb{R}$ is uncountable and $\mathbb{Q}$ is countable, such a basis must be uncountable!

The notion of a basis can also be defined in terms of the notion of maximal linearly independent family and minimal generating family.

Definition 3.7. Let $(v_i)_{i \in I}$ be a family of vectors in a vector space E. We say that $(v_i)_{i \in I}$ a *maximal linearly independent family of E* if it is linearly independent, and if for any vector $w \in E$, the family $(v_i)_{i \in I} \cup_k \{w\}$ obtained by adding w to the family $(v_i)_{i \in I}$ is linearly dependent. We say that $(v_i)_{i \in I}$ a *minimal generating family of E* if it spans E, and if for any index $p \in I$, the family $(v_i)_{i \in I - \{p\}}$ obtained by removing v_p from the family $(v_i)_{i \in I}$ does not span E.

The following proposition giving useful properties characterizing a basis is an immediate consequence of Lemma 3.6.

Proposition 3.8. *Given a vector space E, for any family $B = (v_i)_{i \in I}$ of vectors of E, the following properties are equivalent:*

(1) B is a basis of E.

(2) B is a maximal linearly independent family of E.

(3) B is a minimal generating family of E.

Proof. We will first prove the equivalence of (1) and (2). Assume (1). Since B is a basis, it is a linearly independent family. We claim that B is a maximal linearly independent family. If B is not a maximal linearly independent family, then there is some vector $w \in E$ such that the family B' obtained by adding w to B is linearly independent. However, since B is a basis of E, the vector w can be expressed as a linear combination of vectors in B, contradicting the fact that B' is linearly independent.

Conversely, assume (2). We claim that B spans E. If B does not span E, then there is some vector $w \in E$ which is not a linear combination of vectors in B. By Lemma 3.6, the family B' obtained by adding w to B is linearly independent. Since B is a proper subfamily of B', this contradicts the assumption that B is a maximal linearly independent family. Therefore, B must span E, and since B is also linearly independent, it is a basis of E.

Now we will prove the equivalence of (1) and (3). Again, assume (1). Since B is a basis, it is a generating family of E. We claim that B is a minimal generating family. If B is not a minimal generating family, then there is a proper subfamily B' of B that spans E. Then, every $w \in B - B'$ can be expressed as a linear combination of vectors from B', contradicting the fact that B is linearly independent.

Conversely, assume (3). We claim that B is linearly independent. If B is not linearly independent, then some vector $w \in B$ can be expressed as a linear combination of vectors in $B' = B - \{w\}$. Since B generates E, the family B' also generates E, but B' is a proper subfamily of B, contradicting the minimality of B. Since B spans E and is linearly independent, it is a basis of E. $\square$

The second key result of linear algebra is that *for any two bases $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$ of a vector space E, the index sets I and J have the same cardinality.* In particular, if E has a finite basis of n elements, every basis of E has n elements, and the integer n is called the *dimension* of the vector space E.

To prove the second key result, we can use the following *replacement lemma* due to Steinitz. This result shows the relationship between finite linearly independent families and finite families of generators of a vector space. We begin with a version of the lemma which is a bit informal, but easier to understand than the precise and more formal formulation given in Proposition 3.10. The technical difficulty has to do with the fact that some of the indices need to be renamed.

Proposition 3.9. *(Replacement lemma, version 1) Given a vector space E, let $(u_1, \ldots, u_m)$ be any finite linearly independent family in E, and let $(v_1, \ldots, v_n)$ be any finite family such that every u_i is a linear combination of $(v_1, \ldots, v_n)$. Then we must have $m \leq n$, and there is a replacement of m of the vectors v_j by $(u_1, \ldots, u_m)$, such that after renaming some of the indices of the v_js, the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace of E.*

Proof. We proceed by induction on m. When $m = 0$, the family $(u_1, \ldots, u_m)$ is empty, and the proposition holds trivially. For the induction step, we have a linearly independent family $(u_1, \ldots, u_m, u_{m+1})$. Consider the linearly independent family $(u_1, \ldots, u_m)$. By the induction hypothesis, $m \leq n$, and there is a replacement of m of the vectors v_j by $(u_1, \ldots, u_m)$, such that after renaming some of the indices of the vs, the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace of E. The vector u_{m+1} can also be expressed as a linear combination of $(v_1, \ldots, v_n)$, and since $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace, u_{m+1} can be expressed as a linear combination of $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$, say

$$u_{m+1} = \sum_{i=1}^{m} \lambda_i u_i + \sum_{j=m+1}^{n} \lambda_j v_j.$$

We claim that $\lambda_j \neq 0$ for some j with $m + 1 \leq j \leq n$, which implies that $m + 1 \leq n$.

Otherwise, we would have

$$u_{m+1} = \sum_{i=1}^{m} \lambda_i u_i,$$

a nontrivial linear dependence of the u_i, which is impossible since $(u_1, \ldots, u_{m+1})$ are linearly independent.

Therefore, $m + 1 \leq n$, and after renaming indices if necessary, we may assume that $\lambda_{m+1} \neq 0$, so we get

$$v_{m+1} = -\sum_{i=1}^{m} (\lambda_{m+1}^{-1} \lambda_i) u_i - \lambda_{m+1}^{-1} u_{m+1} - \sum_{j=m+2}^{n} (\lambda_{m+1}^{-1} \lambda_j) v_j.$$

Observe that the families $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$ generate the same subspace, since u_{m+1} is a linear combination of $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and v_{m+1} is a linear combination of $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$. Since $(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$ and $(v_1, \ldots, v_n)$ generate the same subspace, we conclude that $(u_1, \ldots, u_{m+1}, v_{m+2}, \ldots, v_n)$ and and $(v_1, \ldots, v_n)$ generate the same subspace, which concludes the induction hypothesis. $\square$

Here is an example illustrating the replacement lemma. Consider sequences (u_1, u_2, u_3) and $(v_1, v_2, v_3, v_4, v_5)$, where (u_1, u_2, u_3) is a linearly independent family and with the u_is expressed in terms of the v_js as follows:

$$\begin{aligned}
u_1 &= v_4 + v_5 \\
u_2 &= v_3 + v_4 - v_5 \\
u_3 &= v_1 + v_2 + v_3.
\end{aligned}$$

From the first equation we get

$$v_4 = u_1 - v_5,$$

and by substituting in the second equation we have

$$u_2 = v_3 + v_4 - v_5 = v_3 + u_1 - v_5 - v_5 = u_1 + v_3 - 2v_5.$$

From the above equation we get

$$v_3 = -u_1 + u_2 + 2v_5,$$

and so

$$u_3 = v_1 + v_2 + v_3 = v_1 + v_2 - u_1 + u_2 + 2v_5.$$

Finally, we get

$$v_1 = u_1 - u_2 + u_3 - v_2 - 2v_5$$

Therefore we have

$$\begin{aligned}
v_1 &= u_1 - u_2 + u_3 - v_2 - 2v_5 \\
v_3 &= -u_1 + u_2 + 2v_5 \\
v_4 &= u_1 - v_5,
\end{aligned}$$

which shows that $(u_1, u_2, u_3, v_2, v_5)$ spans the same subspace as $(v_1, v_2, v_3, v_4, v_5)$. The vectors (v_1, v_3, v_4) have been replaced by (u_1, u_2, u_3), and the vectors left over are (v_2, v_5). We can rename them (v_4, v_5).

For the sake of completeness, here is a more formal statement of the replacement lemma (and its proof).

Proposition 3.10. *(Replacement lemma, version 2) Given a vector space E, let $(u_i)_{i\in I}$ be any finite linearly independent family in E, where $|I| = m$, and let $(v_j)_{j\in J}$ be any finite family such that every u_i is a linear combination of $(v_j)_{j\in J}$, where $|J| = n$. Then there exists a set L and an injection $\rho\colon L \to J$ (a relabeling function) such that $L \cap I = \emptyset$, $|L| = n - m$, and the families $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E. In particular, $m \leq n$.*

Proof. We proceed by induction on $|I| = m$. When $m = 0$, the family $(u_i)_{i\in I}$ is empty, and the proposition holds trivially with $L = J$ (ρ is the identity). Assume $|I| = m+1$. Consider the linearly independent family $(u_i)_{i\in(I-\{p\})}$, where p is any member of I. By the induction hypothesis, there exists a set L and an injection $\rho\colon L \to J$ such that $L \cap (I - \{p\}) = \emptyset$, $|L| = n - m$, and the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E. If $p \in L$, we can replace L by $(L - \{p\}) \cup \{p'\}$ where p' does not belong to $I \cup L$, and replace ρ by the injection ρ' which agrees with ρ on $L - \{p\}$ and such that $\rho'(p') = \rho(p)$. Thus, we can always assume that $L \cap I = \emptyset$. Since u_p is a linear combination of $(v_j)_{j\in J}$ and the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(v_j)_{j\in J}$ generate the same subspace of E, u_p is a linear combination of $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$. Let

$$u_p = \sum_{i\in(I-\{p\})} \lambda_i u_i + \sum_{l\in L} \lambda_l v_{\rho(l)}. \tag{1}$$

If $\lambda_l = 0$ for all $l \in L$, we have

$$\sum_{i\in(I-\{p\})} \lambda_i u_i - u_p = 0,$$

contradicting the fact that $(u_i)_{i\in I}$ is linearly independent. Thus, $\lambda_l \neq 0$ for some $l \in L$, say $l = q$. Since $\lambda_q \neq 0$, we have

$$v_{\rho(q)} = \sum_{i\in(I-\{p\})} (-\lambda_q^{-1}\lambda_i)u_i + \lambda_q^{-1}u_p + \sum_{l\in(L-\{q\})} (-\lambda_q^{-1}\lambda_l)v_{\rho(l)}. \tag{2}$$

We claim that the families $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$ and $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$ generate the same subset of E. Indeed, the second family is obtained from the first by replacing $v_{\rho(q)}$ by u_p, and vice-versa, and u_p is a linear combination of $(u_i)_{i\in(I-\{p\})} \cup (v_{\rho(l)})_{l\in L}$, by (1), and $v_{\rho(q)}$ is a linear combination of $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$, by (2). Thus, the families $(u_i)_{i\in I} \cup (v_{\rho(l)})_{l\in(L-\{q\})}$ and $(v_j)_{j\in J}$ generate the same subspace of E, and the proposition holds for $L-\{q\}$ and the restriction of the injection $\rho\colon L \to J$ to $L - \{q\}$, since $L \cap I = \emptyset$ and $|L| = n - m$ imply that $(L - \{q\}) \cap I = \emptyset$ and $|L - \{q\}| = n - (m + 1)$. $\square$

The idea is that m of the vectors v_j can be *replaced* by the linearly independent u_is in such a way that the same subspace is still generated. The purpose of the function $\rho\colon L \to J$ is to pick $n - m$ elements $j_1, \ldots, j_{n-m}$ of J and to relabel them $l_1, \ldots, l_{n-m}$ in such a way that these new indices do not clash with the indices in I; this way, the vectors $v_{j_1}, \ldots, v_{j_{n-m}}$ who "survive" (i.e. are not replaced) are relabeled $v_{l_1}, \ldots, v_{l_{n-m}}$, and the other m vectors v_j with $j \in J - \{j_1, \ldots, j_{n-m}\}$ are replaced by the u_i. The index set of this new family is $I \cup L$.

Actually, one can prove that Proposition 3.10 implies Theorem 3.7 when the vector space is finitely generated. Putting Theorem 3.7 and Proposition 3.10 together, we obtain the following fundamental theorem.

Theorem 3.11. *Let E be a finitely generated vector space. Any family $(u_i)_{i\in I}$ generating E contains a subfamily $(u_j)_{j\in J}$ which is a basis of E. Any linearly independent family $(u_i)_{i\in I}$ can be extended to a family $(u_j)_{j\in J}$ which is a basis of E (with $I \subseteq J$). Furthermore, for every two bases $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$ of E, we have $|I| = |J| = n$ for some fixed integer $n \geq 0$.*

Proof. The first part follows immediately by applying Theorem 3.7 with $L = \emptyset$ and $S = (u_i)_{i\in I}$. For the second part, consider the family $S' = (u_i)_{i\in I} \cup (v_h)_{h\in H}$, where $(v_h)_{h\in H}$ is any finitely generated family generating E, and with $I \cap H = \emptyset$. Then apply Theorem 3.7 to $L = (u_i)_{i\in I}$ and to S'. For the last statement, assume that $(u_i)_{i\in I}$ and $(v_j)_{j\in J}$ are bases of E. Since $(u_i)_{i\in I}$ is linearly independent and $(v_j)_{j\in J}$ spans E, Proposition 3.10 implies that $|I| \leq |J|$. A symmetric argument yields $|J| \leq |I|$. $\square$

Remark: Theorem 3.11 also holds for vector spaces that are not finitely generated. This can be shown as follows. Let $(u_i)_{i\in I}$ be a basis of E, let $(v_j)_{j\in J}$ be a generating family of E, and assume that I is infinite. For every $j \in J$, let $L_j \subseteq I$ be the finite set

$$L_j = \{i \in I \mid v_j = \sum_{i \in I} \lambda_i u_i,\ \lambda_i \neq 0\}.$$

Let $L = \bigcup_{j \in J} L_j$. By definition $L \subseteq I$, and since $(u_i)_{i \in I}$ is a basis of E, we must have $I = L$, since otherwise $(u_i)_{i \in L}$ would be another basis of E, and this would contradict the fact that $(u_i)_{i \in I}$ is linearly independent. Furthermore, J must be infinite, since otherwise, because the L_j are finite, I would be finite. But then, since $I = \bigcup_{j \in J} L_j$ with J infinite and the L_j finite, by a standard result of set theory, $|I| \leq |J|$. If $(v_j)_{j \in J}$ is also a basis, by a symmetric argument, we obtain $|J| \leq |I|$, and thus, $|I| = |J|$ for any two bases $(u_i)_{i \in I}$ and $(v_j)_{j \in J}$ of E.

Definition 3.8. When a vector space E is not finitely generated, we say that E is of infinite dimension. The *dimension* of a finitely generated vector space E is the common dimension n of all of its bases and is denoted by $\dim(E)$.

Clearly, if the field K itself is viewed as a vector space, then every family (a) where $a \in K$ and $a \neq 0$ is a basis. Thus $\dim(K) = 1$. Note that $\dim(\{0\}) = 0$.

Definition 3.9. If E is a vector space of dimension $n \geq 1$, for any subspace U of E, if $\dim(U) = 1$, then U is called a *line*; if $\dim(U) = 2$, then U is called a *plane*; if $\dim(U) = n-1$, then U is called a *hyperplane*. If $\dim(U) = k$, then U is sometimes called a *k-plane*.

Let $(u_i)_{i \in I}$ be a basis of a vector space E. For any vector $v \in E$, since the family $(u_i)_{i \in I}$ generates E, there is a family $(\lambda_i)_{i \in I}$ of scalars in K, such that

$$v = \sum_{i \in I} \lambda_i u_i.$$

A very important fact is that the family $(\lambda_i)_{i \in I}$ is **unique**.

Proposition 3.12. *Given a vector space E, let $(u_i)_{i \in I}$ be a family of vectors in E. Let $v \in E$, and assume that $v = \sum_{i \in I} \lambda_i u_i$. Then the family $(\lambda_i)_{i \in I}$ of scalars such that $v = \sum_{i \in I} \lambda_i u_i$ is unique iff $(u_i)_{i \in I}$ is linearly independent.*

Proof. First, assume that $(u_i)_{i \in I}$ is linearly independent. If $(\mu_i)_{i \in I}$ is another family of scalars in K such that $v = \sum_{i \in I} \mu_i u_i$, then we have

$$\sum_{i \in I} (\lambda_i - \mu_i) u_i = 0,$$

and since $(u_i)_{i \in I}$ is linearly independent, we must have $\lambda_i - \mu_i = 0$ for all $i \in I$, that is, $\lambda_i = \mu_i$ for all $i \in I$. The converse is shown by contradiction. If $(u_i)_{i \in I}$ was linearly dependent, there would be a family $(\mu_i)_{i \in I}$ of scalars not all null such that

$$\sum_{i \in I} \mu_i u_i = 0$$

and $\mu_j \neq 0$ for some $j \in I$. But then,

$$v = \sum_{i \in I} \lambda_i u_i + 0 = \sum_{i \in I} \lambda_i u_i + \sum_{i \in I} \mu_i u_i = \sum_{i \in I} (\lambda_i + \mu_i) u_i,$$

with $\lambda_j \neq \lambda_j + \mu_j$ since $\mu_j \neq 0$, contradicting the assumption that $(\lambda_i)_{i \in I}$ is the unique family such that $v = \sum_{i \in I} \lambda_i u_i$. $\square$

Definition 3.10. If $(u_i)_{i \in I}$ is a basis of a vector space E, for any vector $v \in E$, if $(x_i)_{i \in I}$ is the unique family of scalars in K such that

$$v = \sum_{i \in I} x_i u_i,$$

each x_i is called the *component (or coordinate) of index i of v with respect to the basis $(u_i)_{i \in I}$*.

Given a field K and any (nonempty) set I, we can form a vector space $K^{(I)}$ which, in some sense, is the standard vector space of dimension $|I|$.

Definition 3.11. Given a field K and any (nonempty) set I, let $K^{(I)}$ be the subset of the cartesian product K^I consisting of all families $(\lambda_i)_{i \in I}$ with finite support of scalars in K.[1] We define addition and multiplication by a scalar as follows:

$$(\lambda_i)_{i \in I} + (\mu_i)_{i \in I} = (\lambda_i + \mu_i)_{i \in I},$$

and

$$\lambda \cdot (\mu_i)_{i \in I} = (\lambda \mu_i)_{i \in I}.$$

It is immediately verified that addition and multiplication by a scalar are well defined. Thus, $K^{(I)}$ is a vector space. Furthermore, because families with finite support are considered, the family $(e_i)_{i \in I}$ of vectors e_i, defined such that $(e_i)_j = 0$ if $j \neq i$ and $(e_i)_i = 1$, is clearly a basis of the vector space $K^{(I)}$. When $I = \{1, \ldots, n\}$, we denote $K^{(I)}$ by K^n. The function $\iota \colon I \to K^{(I)}$, such that $\iota(i) = e_i$ for every $i \in I$, is clearly an injection.

When I is a finite set, $K^{(I)} = K^I$, but this is false when I is infinite. In fact, $\dim(K^{(I)}) = |I|$, but $\dim(K^I)$ is strictly greater when I is infinite.

3.6 Matrices

In Section 2.1 we introduced informally the notion of a matrix. In this section we define matrices precisely, and also introduce some operations on matrices. It turns out that matrices form a vector space equipped with a multiplication operation which is associative, but noncommutative. We will explain in Section 4.1 how matrices can be used to represent linear maps, defined in the next section.

[1]Where K^I denotes the set of all functions from I to K.

Definition 3.12. If $K = \mathbb{R}$ or $K = \mathbb{C}$, an *$m \times n$-matrix over K* is a family $(a_{i\,j})_{1\leq i\leq m,\ 1\leq j\leq n}$ of scalars in K, represented by an array

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix}$$

In the special case where $m = 1$, we have a *row vector*, represented by

$$(a_{1\,1} \cdots a_{1\,n})$$

and in the special case where $n = 1$, we have a *column vector*, represented by

$$\begin{pmatrix} a_{1\,1} \\ \vdots \\ a_{m\,1} \end{pmatrix}.$$

In these last two cases, we usually omit the constant index 1 (first index in case of a row, second index in case of a column). The set of all $m \times n$-matrices is denoted by $\mathrm{M}_{m,n}(K)$ or $\mathrm{M}_{m,n}$. An $n \times n$-matrix is called a *square matrix of dimension n*. The set of all square matrices of dimension n is denoted by $\mathrm{M}_n(K)$, or M_n.

Remark: As defined, a matrix $A = (a_{i\,j})_{1\leq i\leq m,\ 1\leq j\leq n}$ is a *family*, that is, a function from $\{1, 2, \dots, m\} \times \{1, 2, \dots, n\}$ to K. As such, there is no reason to assume an ordering on the indices. Thus, the matrix A can be represented in many different ways as an array, by adopting different orders for the rows or the columns. However, it is customary (and usually convenient) to assume the natural ordering on the sets $\{1, 2, \dots, m\}$ and $\{1, 2, \dots, n\}$, and to represent A as an array according to this ordering of the rows and columns.

We define some operations on matrices as follows.

Definition 3.13. Given two $m \times n$ matrices $A = (a_{i\,j})$ and $B = (b_{i\,j})$, we define their *sum* $A + B$ as the matrix $C = (c_{i\,j})$ such that $c_{i\,j} = a_{i\,j} + b_{i\,j}$; that is,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} + \begin{pmatrix} b_{1\,1} & b_{1\,2} & \dots & b_{1\,n} \\ b_{2\,1} & b_{2\,2} & \dots & b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m\,1} & b_{m\,2} & \dots & b_{m\,n} \end{pmatrix}$$
$$= \begin{pmatrix} a_{1\,1} + b_{1\,1} & a_{1\,2} + b_{1\,2} & \dots & a_{1\,n} + b_{1\,n} \\ a_{2\,1} + b_{2\,1} & a_{2\,2} + b_{2\,2} & \dots & a_{2\,n} + b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} + b_{m\,1} & a_{m\,2} + b_{m\,2} & \dots & a_{m\,n} + b_{m\,n} \end{pmatrix}.$$

For any matrix $A = (a_{i\,j})$, we let $-A$ be the matrix $(-a_{i\,j})$. Given a scalar $\lambda \in K$, we define the matrix λA as the matrix $C = (c_{i\,j})$ such that $c_{i\,j} = \lambda a_{i\,j}$; that is

$$\lambda \begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} = \begin{pmatrix} \lambda a_{1\,1} & \lambda a_{1\,2} & \dots & \lambda a_{1\,n} \\ \lambda a_{2\,1} & \lambda a_{2\,2} & \dots & \lambda a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m\,1} & \lambda a_{m\,2} & \dots & \lambda a_{m\,n} \end{pmatrix}.$$

Given an $m \times n$ matrices $A = (a_{i\,k})$ and an $n \times p$ matrices $B = (b_{k\,j})$, we define their *product* AB as the $m \times p$ matrix $C = (c_{i\,j})$ such that

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j},$$

for $1 \leq i \leq m$, and $1 \leq j \leq p$. In the product $AB = C$ shown below

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \dots & a_{m\,n} \end{pmatrix} \begin{pmatrix} b_{1\,1} & b_{1\,2} & \dots & b_{1\,p} \\ b_{2\,1} & b_{2\,2} & \dots & b_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n\,1} & b_{n\,2} & \dots & b_{n\,p} \end{pmatrix} = \begin{pmatrix} c_{1\,1} & c_{1\,2} & \dots & c_{1\,p} \\ c_{2\,1} & c_{2\,2} & \dots & c_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m\,1} & c_{m\,2} & \dots & c_{m\,p} \end{pmatrix},$$

note that the entry of index i and j of the matrix AB obtained by multiplying the matrices A and B can be identified with the product of the row matrix corresponding to the i-th row of A with the column matrix corresponding to the j-column of B:

$$(a_{i\,1} \cdots a_{i\,n}) \begin{pmatrix} b_{1\,j} \\ \vdots \\ b_{n\,j} \end{pmatrix} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j}.$$

Definition 3.14. The square matrix I_n of dimension n containing 1 on the diagonal and 0 everywhere else is called the *identity matrix*. It is denoted by

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Definition 3.15. Given an $m \times n$ matrix $A = (a_{i\,j})$, its *transpose* $A^{\top} = (a^{\top}_{j\,i})$, is the $n \times m$-matrix such that $a^{\top}_{j\,i} = a_{i\,j}$, for all i, $1 \leq i \leq m$, and all j, $1 \leq j \leq n$.

The transpose of a matrix A is sometimes denoted by A^t, or even by ${}^t A$. Note that the transpose $A^{\top}$ of a matrix A has the property that the j-th row of $A^{\top}$ is the j-th column of

A. In other words, transposition exchanges the rows and the columns of a matrix. Here is an example. If A is the 5×6 matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 1 & 2 & 3 & 4 & 5 \\ 8 & 7 & 1 & 2 & 3 & 4 \\ 9 & 8 & 7 & 1 & 2 & 3 \\ 10 & 9 & 8 & 7 & 1 & 2 \end{pmatrix},$$

then $A^\top$ is the 6×5 matrix

$$A^\top = \begin{pmatrix} 1 & 7 & 8 & 9 & 10 \\ 2 & 1 & 7 & 8 & 9 \\ 3 & 2 & 1 & 7 & 8 \\ 4 & 3 & 2 & 1 & 7 \\ 5 & 4 & 3 & 2 & 1 \\ 6 & 5 & 4 & 3 & 2 \end{pmatrix}.$$

The following observation will be useful later on when we discuss the SVD. Given any $m \times n$ matrix A and any $n \times p$ matrix B, if we denote the columns of A by $A^1, \ldots, A^n$ and the rows of B by $B_1, \ldots, B_n$, then we have

$$AB = A^1 B_1 + \cdots + A^n B_n.$$

For every square matrix A of dimension n, it is immediately verified that $AI_n = I_n A = A$.

Definition 3.16. For any square matrix A of dimension n, if a matrix B such that $AB = BA = I_n$ exists, then it is unique, and it is called the *inverse* of A. The matrix B is also denoted by A^{-1}. An invertible matrix is also called a *nonsingular* matrix, and a matrix that is not invertible is called a *singular* matrix.

The following result is a matrix analog of Proposition 3.21.

Proposition 3.13. *If a square matrix $A \in \mathrm{M}_n(K)$ has a left inverse, that is a matrix B such that $BA = I_n$, or a right inverse, that is a matrix C such that $AC = I_n$, then A is actually invertible. Furthermore, $B = A^{-1}$ and $C = A^{-1}$.*

Proof. Proposition 3.13 follows from Proposition 3.21 and the fact that matrices represent linear maps. We can also give a direct proof as follows. Suppose that there is a matrix B such that $BA = I_n$. This implies that the columns $A^1, \ldots, A^n$ of A are linearly independent, because if

$$A\lambda = \lambda_1 A^1 + \cdots + \lambda_n A^n = 0,$$

where $\lambda \in K^n$ is the column vector

$$\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix},$$

for some $\lambda_1, \ldots, \lambda_n \in K$, by multiplying both sides of the equation $A\lambda = 0$ by B we get

$$BA\lambda = I_n\lambda = \lambda = B0 = 0,$$

so $\lambda = 0$. Then since $(A^1, \ldots, A^n)$ are n linearly independent vectors in K^n, they form a basis of K^n. Consequently, for every vector $b \in K^n$, there is a unique column vector $(x_1, \ldots, x_n) \in K^n$ such that

$$Ax = x_1A^1 + \cdots + x_nA^n = b,$$

where x is the column vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Thus we can solve the n equations

$$Ax^j = e_j, \quad 1 \leq j \leq n,$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the jth canonical basis vector in K^n. These equations yield the matrix equation

$$AX = I_n,$$

where $X = (x^1 \, \cdots \, x^n)$ is the $n \times n$ matrix whose jth column is x^j. Consequently, X is a right inverse of A. Now A has a left inverse B and a right inverse X, so by Proposition 2.2, we have $X = B$, so A is invertible and its inverse is equal to B.

Let us now assume that there is a matrix C such that $AC = I_n$. We can repeat the previous proof with C playing the role of A and A playing the role of B to conclude that C is invertible and that $C^{-1} = A$. But then C^{-1} is invertible with inverse C, and since $C = (C^{-1})^{-1} = A^{-1}$, we conclude that A is invertible and that its inverse is equal to C. $\square$

Using Proposition 2.3 (or mimicking the computations in its proof), we note that if A and B are two $n \times n$ invertible matrices, then AB is also invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

An important criterion for a square matrix to be invertible is stated next. Another proof is provided in Proposition 4.4 .

Proposition 3.14. *A square matrix $A \in \mathrm{M}_n(K)$ is invertible iff its columns $(A^1, \ldots, A^n)$ are linearly independent.*

Proof. If A is invertible, then in particular it has a left inverse A^{-1}, so the first part of the proof of Proposition 3.13 with $B = A^{-1}$ proves that the columns $(A^1, \ldots, A^n)$ of A are linearly independent. This fact is also proven as part of the proof of Proposition 4.4. Conversely, assume that the columns $(A^1, \ldots, A^n)$ of A are linearly independent. The second part of the proof of Proposition 3.13 shows that A is invertible. $\square$

Another useful criterion for a square matrix to be invertible is stated next.

Proposition 3.15. *A square matrix $A \in \mathrm{M}_n(K)$ is invertible iff for any $x \in K^n$, the equation $Ax = 0$ implies that $x = 0$.*

Proof. If A is invertible and if $Ax = 0$, then by multiplying both sides of the equation $x = 0$ by A^{-1}, we get

$$A^{-1}Ax = I_n x = x = A^{-1}0 = 0.$$

Conversely, for any $x = (x_1, \ldots, x_n) \in K^n$, since

$$Ax = x_1 A^1 + \cdots + x_n A^n,$$

the condition $Ax = 0$ implies $x = 0$ is equivalent to the linear independence of the columns $(A^1, \ldots, A^n)$ of A. By Proposition 3.14, the matrix A is invertible. $\square$

It is immediately verified that the set $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices is a *vector space* under addition of matrices and multiplication of a matrix by a scalar.

Definition 3.17. The $m \times n$-matrices $E_{ij} = (e_{h\,k})$, are defined such that $e_{i\,j} = 1$, and $e_{h\,k} = 0$, if $h \neq i$ or $k \neq j$; in other words, the (i, j)-entry is equal to 1 and all other entries are 0.

Here are the E_{ij} matrices for $m = 2$ and $n = 3$:

$$E_{11} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad E_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad E_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$
$$E_{21} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad E_{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad E_{23} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is clear that every matrix $A = (a_{i\,j}) \in \mathrm{M}_{m,n}(K)$ can be written in a unique way as

$$A = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i\,j} E_{ij}.$$

Thus, the family $(E_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$ is a basis of the vector space $\mathrm{M}_{m,n}(K)$, which has dimension mn.

Remark: Definition 3.12 and Definition 3.13 also make perfect sense when K is a (commutative) ring rather than a field. In this more general setting, the framework of vector spaces is too narrow, but we can consider structures over a commutative ring A satisfying all the axioms of Definition 3.1. Such structures are called *modules*. The theory of modules is (much) more complicated than that of vector spaces. For example, modules do not always have a basis, and other properties holding for vector spaces usually fail for modules. When a module has a basis, it is called a *free module*. For example, when A is a commutative ring, the structure A^n is a module such that the vectors e_i, with $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$, form a basis of A^n. Many properties of vector spaces still hold for A^n. Thus, A^n is a free module. As another example, when A is a commutative ring, $\mathrm{M}_{m,n}(A)$ is a free module with basis $(E_{i,j})_{1\leq i\leq m, 1\leq j\leq n}$. Polynomials over a commutative ring also form a free module of infinite dimension.

The properties listed in Proposition 3.16 are easily verified, although some of the computations are a bit tedious. A more conceptual proof is given in Proposition 4.1.

Proposition 3.16. *(1) Given any matrices $A \in \mathrm{M}_{m,n}(K)$, $B \in \mathrm{M}_{n,p}(K)$, and $C \in \mathrm{M}_{p,q}(K)$, we have*

$$(AB)C = A(BC);$$

that is, matrix multiplication is associative.

(2) Given any matrices $A, B \in \mathrm{M}_{m,n}(K)$, and $C, D \in \mathrm{M}_{n,p}(K)$, for all $\lambda \in K$, we have

$$\begin{aligned}(A + B)C &= AC + BC\\ A(C + D) &= AC + AD\\ (\lambda A)C &= \lambda(AC)\\ A(\lambda C) &= \lambda(AC),\end{aligned}$$

so that matrix multiplication $\cdot \colon \mathrm{M}_{m,n}(K) \times \mathrm{M}_{n,p}(K) \to \mathrm{M}_{m,p}(K)$ is bilinear.

The properties of Proposition 3.16 together with the fact that $AI_n = I_nA = A$ for all square $n\times n$ matrices show that $\mathrm{M}_n(K)$ is a ring with unit I_n (in fact, an associative algebra). This is a noncommutative ring with zero divisors, as shown by the following example.

Example 3.5. For example, letting A, B be the 2×2-matrices

$$A = \begin{pmatrix} 1 & 0\\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0\\ 1 & 0 \end{pmatrix},$$

then

$$AB = \begin{pmatrix} 1 & 0\\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0\\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0\\ 0 & 0 \end{pmatrix},$$

and

$$BA = \begin{pmatrix} 0 & 0\\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0\\ 1 & 0 \end{pmatrix}.$$

Thus $AB \neq BA$, and $AB = 0$, even though both $A, B \neq 0$.

3.7 Linear Maps

Now that we understand vector spaces and how to generate them, we would like to be able to transform one vector space E into another vector space F. A function between two vector spaces that preserves the vector space structure is called a homomorphism of vector spaces, or *linear map*. Linear maps formalize the concept of linearity of a function.

Keep in mind that linear maps, which are transformations of space, are usually far more important than the spaces themselves.

In the rest of this section, we assume that all vector spaces are over a given field K (say $\mathbb{R}$).

Definition 3.18. Given two vector spaces E and F, a *linear map* between E and F is a function $f\colon E \to F$ satisfying the following two conditions:

$$\begin{aligned} f(x +_E y) &= f(x) +_F f(y) && \text{for all } x, y \in E; \\ f(\lambda \cdot_E x) &= \lambda \cdot_F f(x) && \text{for all } \lambda \in K,\ x \in E. \end{aligned}$$

Setting $x = y = 0_E$ in the first identity, we get $f(0_E) = 0_F$. *The basic property of linear maps is that they transform linear combinations into linear combinations.* Given any finite family $(u_i)_{i\in I}$ of vectors in E, given any family $(\lambda_i)_{i\in I}$ of scalars in K, we have

$$f\left(\sum_{i\in I} \lambda_i \cdot_E u_i\right) = \sum_{i\in I} \lambda_i \cdot_F f(u_i).$$

The sum $\sum_{i\in I} \lambda_i \cdot_E u_i$ uses addition $+_E$ in E, and the sum $\sum_{i\in I} \lambda_i \cdot_F f(u_i)$ uses addition in F. The above identity is shown by induction on $|I|$ using the properties of Definition 3.18.

As we did when defining vector spaces, for notational simplicity, we usually drop the subscripts E and F in the notations $+_E, +_F$ and $\cdot_E, \cdot_F$.

Example 3.6.

1. The map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined such that

$$\begin{aligned} x' &= x - y \\ y' &= x + y \end{aligned}$$

is a linear map. The reader should check that it is the composition of a rotation by $\pi/4$ with a magnification of ratio $\sqrt{2}$.

2. For any vector space E, the *identity map* $\mathrm{id}\colon E \to E$ given by
$$\mathrm{id}(u) = u \quad \text{for all } u \in E$$
is a linear map. When we want to be more precise, we write id_E instead of id.

3. The map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ defined such that
$$(DP)(X) = P'(X),$$
where $P'(X)$ is the derivative of the polynomial $P(X)$, is a linear map. More explicitly, if
$$P(X) = a_0X^n + a_1X^{n-1} + \cdots + a_{n-2}X^2 + a_{n-1}X + a_n, \quad a_i \in \mathbb{R}$$
is a polynomial of degree n, then
$$(DP)(X) = na_0X^{n-1} + (n-1)a_1X^{n-2} + \cdots + 2a_{n-2}X + a_{n-1}.$$

4. The map $\Phi\colon \mathcal{C}([a,b]) \to \mathbb{R}$ given by
$$\Phi(f) = \int_a^b f(t)dt,$$
where $\mathcal{C}([a,b])$ is the set of continuous functions defined on the interval $[a,b]$, is a linear map.

5. The function $\langle -,-\rangle\colon \mathcal{C}([a,b]) \times \mathcal{C}([a,b]) \to \mathbb{R}$ given by
$$\langle f, g\rangle = \int_a^b f(t)g(t)dt,$$
is linear in each of the variable f, g. It also satisfies the properties $\langle f, g\rangle = \langle g, f\rangle$ and $\langle f, f\rangle = 0$ iff $f = 0$. It is an example of an *inner product*.

Definition 3.19. Given a linear map $f\colon E \to F$, we define its *image (or range)* $\mathrm{Im}\, f = f(E)$, as the set
$$\mathrm{Im}\, f = \{y \in F \mid (\exists x \in E)(y = f(x))\},$$
and its *Kernel (or nullspace)* $\mathrm{Ker}\, f = f^{-1}(0_F)$, as the set
$$\mathrm{Ker}\, f = \{x \in E \mid f(x) = 0_F\}.$$

The derivative map $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$ from Example 3.6(3) has kernel the constant polynomials, so $\mathrm{Ker}\, D = \mathbb{R}$. If we consider the second derivative $D \circ D\colon \mathbb{R}[X] \to \mathbb{R}[X]$, then the kernel of $D \circ D$ consists of all polynomials of degree ≤ 1. The image of $D\colon \mathbb{R}[X] \to \mathbb{R}[X]$

is actually $\mathbb{R}[X]$ itself, because every polynomial $P(X) = a_0X^n + \cdots + a_{n-1}X + a_n$ of degree n is the derivative of the polynomial $Q(X)$ of degree $n+1$ given by

$$Q(X) = a_0\frac{X^{n+1}}{n+1} + \cdots + a_{n-1}\frac{X^2}{2} + a_nX.$$

On the other hand, if we consider the restriction of D to the vector space $\mathbb{R}[X]_n$ of polynomials of degree $\leq n$, then the kernel of D is still $\mathbb{R}$, but the image of D is the $\mathbb{R}[X]_{n-1}$, the vector space of polynomials of degree $\leq n-1$.

Proposition 3.17. *Given a linear map $f\colon E \to F$, the set $\mathrm{Im}\, f$ is a subspace of F and the set $\mathrm{Ker}\, f$ is a subspace of E. The linear map $f\colon E \to F$ is injective iff $\mathrm{Ker}\, f = (0_E)$ (where (0_E) is the trivial subspace $\{0_E\}$).*

Proof. Given any $x, y \in \mathrm{Im}\, f$, there are some $u, v \in E$ such that $x = f(u)$ and $y = f(v)$, and for all $\lambda, \mu \in \mathbb{R}$, we have

$$f(\lambda u +_E \mu v) = \lambda f(u) +_F \mu f(v) = \lambda x +_F \mu y,$$

and thus, $\lambda x +_F \mu y \in \mathrm{Im}\, f$, showing that $\mathrm{Im}\, f$ is a subspace of F.

Given any $x, y \in \mathrm{Ker}\, f$, we have $f(x) = 0_F$ and $f(y) = 0_F$, and thus,

$$f(\lambda x +_E \mu y) = \lambda f(x) +_F \mu f(y) = 0_F,$$

that is, $\lambda x +_E \mu y \in \mathrm{Ker}\, f$, showing that $\mathrm{Ker}\, f$ is a subspace of E.

First, assume that $\mathrm{Ker}\, f = (0_E)$. We need to prove that $f(x) = f(y)$ implies that $x = y$. However, if $f(x) = f(y)$, then $f(x) -_F f(y) = 0_F$, and by linearity of f we get $f(x -_E y) = 0_F$. Because $\mathrm{Ker}\, f = (0_E)$, we must have $x -_E y = 0_E$, that is $x = y$, so f is injective. Conversely, assume that f is injective. If $x \in \mathrm{Ker}\, f$, that is $f(x) = 0_F$, since $f(0_E) = 0_F$ we have $f(x) = f(0_E)$, and by injectivity, $x = 0_E$, which proves that $\mathrm{Ker}\, f = (0_E)$. Therefore, f is injective iff $\mathrm{Ker}\, f = (0_E)$. $\square$

Since by Proposition 3.17, the image $\mathrm{Im}\, f$ of a linear map f is a subspace of F, we can define the *rank* $\mathrm{rk}(f)$ of f as the dimension of $\mathrm{Im}\, f$.

Definition 3.20. Given a linear map $f\colon E \to F$, the *rank* $\mathrm{rk}(f)$ of f is the dimension of the image $\mathrm{Im}\, f$ of f.

A fundamental property of bases in a vector space is that they allow the definition of linear maps as unique homomorphic extensions, as shown in the following proposition.

Proposition 3.18. *Given any two vector spaces E and F, given any basis $(u_i)_{i\in I}$ of E, given any other family of vectors $(v_i)_{i\in I}$ in F, there is a unique linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$. Furthermore, f is injective iff $(v_i)_{i\in I}$ is linearly independent, and f is surjective iff $(v_i)_{i\in I}$ generates F.*

Proof. If such a linear map $f\colon E \to F$ exists, since $(u_i)_{i\in I}$ is a basis of E, every vector $x \in E$ can written uniquely as a linear combination

$$x = \sum_{i\in I} x_i u_i,$$

and by linearity, we must have

$$f(x) = \sum_{i\in I} x_i f(u_i) = \sum_{i\in I} x_i v_i.$$

Define the function $f\colon E \to F$, by letting

$$f(x) = \sum_{i\in I} x_i v_i, \quad x = \sum_{i\in I} x_i u_i, \tag{$\dagger$}$$

for some unique coordinates $(x_i)_{i\in I}$ of x.

To prove that f as defined by ($\dagger$) is linear it suffices to prove that

$$f(\lambda x + \mu y) = \lambda f(x) + \mu f(y)$$

for all $x, y \in E$ and all $\lambda, \mu \in \mathbb{R}$. Since $(u_i)_{i\in I}$ is a basis of E, we have

$$x = \sum_{i\in I} x_i u_i, \quad y = \sum_{i\in I} y_i u_i,$$

for some unique coordinates $(x_i)_{i\in I}$ of x and $(y_i)_{i\in I}$ of y, and by ($\dagger$) we have

$$f(x) = \sum_{i\in I} x_i v_i, \quad f(y) = \sum_{i\in I} y_i v_i,$$

and since

$$\lambda x + \mu y = \lambda\Biggl(\sum_{i\in I} x_i u_i\Biggr) + \mu\Biggl(\sum_{i\in I} y_i u_i\Biggr) = \sum_{i\in I}(\lambda x_i + \mu y_i) u_i,$$

by ($\dagger$),

$$\begin{aligned} f(\lambda x + \mu y) &= f\Biggl(\sum_{i\in I}(\lambda x_i + \mu y_i) u_i\Biggr) = \sum_{i\in I}(\lambda x_i + \mu y_i) v_i \\ &= \lambda\Biggl(\sum_{i\in I} x_i v_i\Biggr) + \mu\Biggl(\sum_{i\in I} y_i v_i\Biggr) = \lambda f(x) + \mu f(y), \end{aligned}$$

proving that f is linear. The map f is unique by ($\dagger$), and obviously, $f(u_i) = v_i$.

Now assume that f is injective. Let $(\lambda_i)_{i\in I}$ be any family of scalars, and assume that

$$\sum_{i\in I} \lambda_i v_i = 0.$$

Since $v_i = f(u_i)$ for every $i \in I$, we have

$$f\left(\sum_{i\in I} \lambda_i u_i\right) = \sum_{i\in I} \lambda_i f(u_i) = \sum_{i\in I} \lambda_i v_i = 0.$$

Since f is injective iff $\mathrm{Ker}\, f = (0)$, we have

$$\sum_{i\in I} \lambda_i u_i = 0,$$

and since $(u_i)_{i\in I}$ is a basis, we have $\lambda_i = 0$ for all $i \in I$, which shows that $(v_i)_{i\in I}$ is linearly independent. Conversely, assume that $(v_i)_{i\in I}$ is linearly independent. Since $(u_i)_{i\in I}$ is a basis of E, every vector $x \in E$ is a linear combination $x = \sum_{i\in I} \lambda_i u_i$ of $(u_i)_{i\in I}$. If

$$f(x) = f\left(\sum_{i\in I} \lambda_i u_i\right) = 0,$$

then

$$\sum_{i\in I} \lambda_i v_i = \sum_{i\in I} \lambda_i f(u_i) = f\left(\sum_{i\in I} \lambda_i u_i\right) = 0,$$

and $\lambda_i = 0$ for all $i \in I$ because $(v_i)_{i\in I}$ is linearly independent, which means that $x = 0$. Therefore, $\mathrm{Ker}\, f = (0)$, which implies that f is injective. The part where f is surjective is left as a simple exercise. $\square$

Figure 3.11 provides an illustration of Proposition 3.18 when $E = \mathbb{R}^3$ and $V = \mathbb{R}^2$

By the second part of Proposition 3.18, an injective linear map $f\colon E \to F$ sends a basis $(u_i)_{i\in I}$ to a linearly independent family $(f(u_i))_{i\in I}$ of F, which is also a basis when f is bijective. Also, when E and F have the same finite dimension n, $(u_i)_{i\in I}$ is a basis of E, and $f\colon E \to F$ is injective, then $(f(u_i))_{i\in I}$ is a basis of F (by Proposition 3.8).

We can now show that the vector space $K^{(I)}$ of Definition 3.11 has a universal property that amounts to saying that $K^{(I)}$ is the vector space freely generated by I. Recall that $\iota\colon I \to K^{(I)}$, such that $\iota(i) = e_i$ for every $i \in I$, is an injection from I to $K^{(I)}$.

Proposition 3.19. *Given any set I, for any vector space F, and for any function $f\colon I \to F$, there is a unique linear map $\overline{f}\colon K^{(I)} \to F$, such that*

$$f = \overline{f} \circ \iota,$$

as in the following diagram:

$$\begin{tikzcd} I \arrow[r, "\iota"] \arrow[dr, "f"'] & K^{(I)} \arrow[d, "\overline{f}"] \\ & F \end{tikzcd}$$

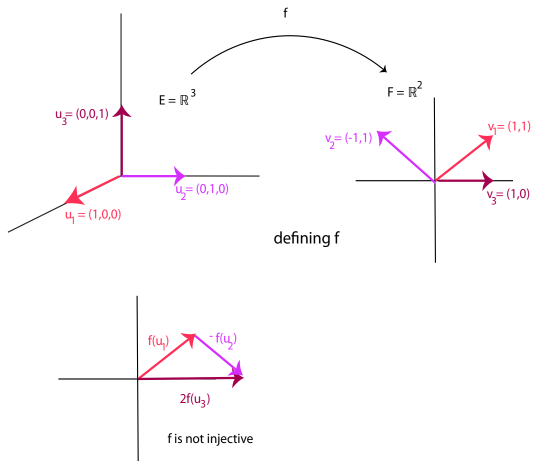


Figure 3.11: Given $u_1 = (1, 0, 0)$, $u_2 = (0, 1, 0)$, $u_3 = (0, 0, 1)$ and $v_1 = (1, 1)$, $v_2 = (-1, 1)$, $v_3 = (1, 0)$, define the unique linear map $f\colon \mathbb{R}^3 \to \mathbb{R}^2$ by $f(u_1) = v_1$, $f(u_2) = v_2$, and $f(u_3) = v_3$. This map is surjective but not injective since $f(u_1 - u_2) = f(u_1) - f(u_2) = (1, 1) - (-1, 1) = (2, 0) = 2f(u_3) = f(2u_3)$.

Proof. If such a linear map $\overline{f}\colon K^{(I)} \to F$ exists, since $f = \overline{f} \circ \iota$, we must have

$$f(i) = \overline{f}(\iota(i)) = \overline{f}(e_i),$$

for every $i \in I$. However, the family $(e_i)_{i\in I}$ is a basis of $K^{(I)}$, and $(f(i))_{i\in I}$ is a family of vectors in F, and by Proposition 3.18, there is a unique linear map $\overline{f}\colon K^{(I)} \to F$ such that $\overline{f}(e_i) = f(i)$ for every $i \in I$, which proves the existence and uniqueness of a linear map $\overline{f}$ such that $f = \overline{f} \circ \iota$. $\square$

The following simple proposition is also useful.

Proposition 3.20. *Given any two vector spaces E and F, with F nontrivial, given any family $(u_i)_{i\in I}$ of vectors in E, the following properties hold:*

(1) The family $(u_i)_{i\in I}$ generates E iff for every family of vectors $(v_i)_{i\in I}$ in F, there is at most one linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

(2) The family $(u_i)_{i\in I}$ is linearly independent iff for every family of vectors $(v_i)_{i\in I}$ in F, there is some linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$.

Proof. (1) If there is any linear map $f\colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$, since $(u_i)_{i\in I}$ generates E, every vector $x \in E$ can be written as some linear combination

$$x = \sum_{i\in I} x_i u_i,$$

and by linearity, we must have

$$f(x) = \sum_{i \in I} x_i f(u_i) = \sum_{i \in I} x_i v_i.$$

This shows that f is unique if it exists. Conversely, assume that $(u_i)_{i \in I}$ does not generate E. Since F is nontrivial, there is some some vector $y \in F$ such that $y \neq 0$. Since $(u_i)_{i \in I}$ does not generate E, there is some vector $w \in E$ that is not in the subspace generated by $(u_i)_{i \in I}$. By Theorem 3.11, there is a linearly independent subfamily $(u_i)_{i \in I_0}$ of $(u_i)_{i \in I}$ generating the same subspace. Since by hypothesis, $w \in E$ is not in the subspace generated by $(u_i)_{i \in I_0}$, by Lemma 3.6 and by Theorem 3.11 again, there is a basis $(e_j)_{j \in I_0 \cup J}$ of E, such that $e_i = u_i$ for all $i \in I_0$, and $w = e_{j_0}$ for some $j_0 \in J$. Letting $(v_i)_{i \in I}$ be the family in F such that $v_i = 0$ for all $i \in I$, defining $f \colon E \to F$ to be the constant linear map with value 0, we have a linear map such that $f(u_i) = 0$ for all $i \in I$. By Proposition 3.18, there is a unique linear map $g \colon E \to F$ such that $g(w) = y$, and $g(e_j) = 0$ for all $j \in (I_0 \cup J) - \{j_0\}$. By definition of the basis $(e_j)_{j \in I_0 \cup J}$ of E, we have $g(u_i) = 0$ for all $i \in I$, and since $f \neq g$, this contradicts the fact that there is at most one such map. See Figure 3.12.

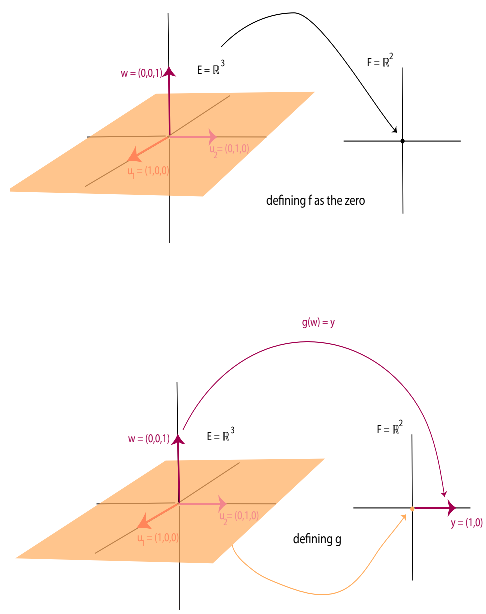


Figure 3.12: Let $E = \mathbb{R}^3$ and $F = \mathbb{R}^2$. The vectors $u_1 = (1, 0, 0)$, $u_2 = (0, 1, 0)$ do not generate $\mathbb{R}^3$ since both the zero map and the map g, where $g(0, 0, 1) = (1, 0)$, send the peach xy-plane to the origin.

(2) If the family $(u_i)_{i \in I}$ is linearly independent, then by Theorem 3.11, $(u_i)_{i \in I}$ can be extended to a basis of E, and the conclusion follows by Proposition 3.18. Conversely, assume

that $(u_i)_{i\in I}$ is linearly dependent. Then there is some family $(\lambda_i)_{i\in I}$ of scalars (not all zero) such that

$$\sum_{i\in I} \lambda_i u_i = 0.$$

By the assumption, for any nonzero vector $y \in F$, for every $i \in I$, there is some linear map $f_i\colon E \to F$, such that $f_i(u_i) = y$, and $f_i(u_j) = 0$, for $j \in I - \{i\}$. Then we would get

$$0 = f_i(\sum_{i\in I} \lambda_i u_i) = \sum_{i\in I} \lambda_i f_i(u_i) = \lambda_i y,$$

and since $y \neq 0$, this implies $\lambda_i = 0$ for every $i \in I$. Thus, $(u_i)_{i\in I}$ is linearly independent. $\square$

Given vector spaces E, F, and G, and linear maps $f\colon E \to F$ and $g\colon F \to G$, it is easily verified that the composition $g \circ f\colon E \to G$ of f and g is a linear map.

Definition 3.21. A linear map $f\colon E \to F$ is an *isomorphism* iff there is a linear map $g\colon F \to E$, such that

$$g \circ f = \mathrm{id}_E \quad \text{and} \quad f \circ g = \mathrm{id}_F. \tag{$*$}$$

The map g in Definition 3.21 is unique. This is because if g and h both satisfy $g \circ f = \mathrm{id}_E$, $f \circ g = \mathrm{id}_F$, $h \circ f = \mathrm{id}_E$, and $f \circ h = \mathrm{id}_F$, then

$$g = g \circ \mathrm{id}_F = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_E \circ h = h.$$

The map g satisfying $(*)$ above is called the *inverse* of f and it is also denoted by f^{-1}.

Observe that Proposition 3.18 shows that if $F = \mathbb{R}^n$, then we get an isomorphism between any vector space E of dimension $|J| = n$ and $\mathbb{R}^n$. Proposition 3.18 also implies that if E and F are two vector spaces, $(u_i)_{i\in I}$ is a basis of E, and $f\colon E \to F$ is a linear map which is an isomorphism, then the family $(f(u_i))_{i\in I}$ is a basis of F.

One can verify that if $f\colon E \to F$ is a bijective linear map, then its inverse $f^{-1}\colon F \to E$, as a function, is also a linear map, and thus f is an isomorphism.

Another useful corollary of Proposition 3.18 is this:

Proposition 3.21. *Let E be a vector space of finite dimension $n \geq 1$ and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) If f has a left inverse g, that is, if g is a linear map such that $g \circ f = \mathrm{id}$, then f is an isomorphism and $f^{-1} = g$.

(2) If f has a right inverse h, that is, if h is a linear map such that $f \circ h = \mathrm{id}$, then f is an isomorphism and $f^{-1} = h$.

Proof. (1) The equation $g \circ f = \mathrm{id}$ implies that f is injective; this is a standard result about functions (if $f(x) = f(y)$, then $g(f(x)) = g(f(y))$, which implies that $x = y$ since $g \circ f = \mathrm{id}$). Let $(u_1, \ldots, u_n)$ be any basis of E. By Proposition 3.18, since f is injective, $(f(u_1), \ldots, f(u_n))$ is linearly independent, and since E has dimension n, it is a basis of E (if $(f(u_1), \ldots, f(u_n))$ doesn't span E, then it can be extended to a basis of dimension strictly greater than n, contradicting Theorem 3.11). Then f is bijective, and by a previous observation its inverse is a linear map. We also have

$$g = g \circ \mathrm{id} = g \circ (f \circ f^{-1}) = (g \circ f) \circ f^{-1} = \mathrm{id} \circ f^{-1} = f^{-1}.$$

(2) The equation $f \circ h = \mathrm{id}$ implies that f is surjective; this is a standard result about functions (for any $y \in E$, we have $f(h(y)) = y$). Let $(u_1, \ldots, u_n)$ be any basis of E. By Proposition 3.18, since f is surjective, $(f(u_1), \ldots, f(u_n))$ spans E, and since E has dimension n, it is a basis of E (if $(f(u_1), \ldots, f(u_n))$ is not linearly independent, then because it spans E, it contains a basis of dimension strictly smaller than n, contradicting Theorem 3.11). Then f is bijective, and by a previous observation its inverse is a linear map. We also have

$$h = \mathrm{id} \circ h = (f^{-1} \circ f) \circ h = f^{-1} \circ (f \circ h) = f^{-1} \circ \mathrm{id} = f^{-1}.$$

This completes the proof. $\square$

Definition 3.22. The set of all linear maps between two vector spaces E and F is denoted by $\mathrm{Hom}(E, F)$ or by $\mathcal{L}(E; F)$ (the notation $\mathcal{L}(E; F)$ is usually reserved to the set of continuous linear maps, where E and F are normed vector spaces). When we wish to be more precise and specify the field K over which the vector spaces E and F are defined we write $\mathrm{Hom}_K(E, F)$.

The set $\mathrm{Hom}(E, F)$ is a vector space under the operations defined in Example 3.1, namely

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in E$, and

$$(\lambda f)(x) = \lambda f(x)$$

for all $x \in E$. The point worth checking carefully is that λf is indeed a linear map, which uses the commutativity of $*$ in the field K (typically, $K = \mathbb{R}$ or $K = \mathbb{C}$). Indeed, we have

$$(\lambda f)(\mu x) = \lambda f(\mu x) = \lambda \mu f(x) = \mu \lambda f(x) = \mu (\lambda f)(x).$$

When E and F have finite dimensions, the vector space $\mathrm{Hom}(E, F)$ also has finite dimension, as we shall see shortly.

Definition 3.23. When $E = F$, a linear map $f\colon E \to E$ is also called an *endomorphism*. The space $\mathrm{Hom}(E, E)$ is also denoted by $\mathrm{End}(E)$.

It is also important to note that composition confers to $\mathrm{Hom}(E, E)$ a ring structure. Indeed, composition is an operation $\circ\colon \mathrm{Hom}(E, E) \times \mathrm{Hom}(E, E) \to \mathrm{Hom}(E, E)$, which is associative and has an identity id_E, and the distributivity properties hold:

$$\begin{aligned}(g_1 + g_2) \circ f &= g_1 \circ f + g_2 \circ f;\\ g \circ (f_1 + f_2) &= g \circ f_1 + g \circ f_2.\end{aligned}$$

The ring $\mathrm{Hom}(E, E)$ is an example of a noncommutative ring.

It is easily seen that the set of bijective linear maps $f\colon E \to E$ is a group under composition.

Definition 3.24. Bijective linear maps $f\colon E \to E$ are also called *automorphisms*. The group of automorphisms of E is called the *general linear group (of E)*, and it is denoted by $\mathbf{GL}(E)$, or by $\mathrm{Aut}(E)$, or when $E = \mathbb{R}^n$, by $\mathbf{GL}(n, \mathbb{R})$, or even by $\mathbf{GL}(n)$.

Although in this book, we will not have many occasions to use quotient spaces, they are fundamental in algebra. The next section may be omitted until needed.

3.8 Quotient Spaces

Let E be a vector space, and let M be any subspace of E. The subspace M induces a relation $\equiv_M$ on E, defined as follows: For all $u, v \in E$,

$u \equiv_M v$ iff $u - v \in M$.

We have the following simple proposition.

Proposition 3.22. *Given any vector space E and any subspace M of E, the relation $\equiv_M$ is an equivalence relation with the following two congruential properties:*

1. *If $u_1 \equiv_M v_1$ and $u_2 \equiv_M v_2$, then $u_1 + u_2 \equiv_M v_1 + v_2$, and*
2. *if $u \equiv_M v$, then $\lambda u \equiv_M \lambda v$.*

Proof. It is obvious that $\equiv_M$ is an equivalence relation. Note that $u_1 \equiv_M v_1$ and $u_2 \equiv_M v_2$ are equivalent to $u_1 - v_1 = w_1$ and $u_2 - v_2 = w_2$, with $w_1, w_2 \in M$, and thus,

$$(u_1 + u_2) - (v_1 + v_2) = w_1 + w_2,$$

and $w_1 + w_2 \in M$, since M is a subspace of E. Thus, we have $u_1 + u_2 \equiv_M v_1 + v_2$. If $u - v = w$, with $w \in M$, then

$$\lambda u - \lambda v = \lambda w,$$

and $\lambda w \in M$, since M is a subspace of E, and thus $\lambda u \equiv_M \lambda v$. $\square$

Proposition 3.22 shows that we can define addition and multiplication by a scalar on the set E/M of equivalence classes of the equivalence relation $\equiv_M$.

Definition 3.25. Given any vector space E and any subspace M of E, we define the following operations of addition and multiplication by a scalar on the set E/M of equivalence classes of the equivalence relation $\equiv_M$ as follows: for any two equivalence classes $[u], [v] \in E/M$, we have

$$\begin{aligned} [u] + [v] &= [u + v], \\ \lambda[u] &= [\lambda u]. \end{aligned}$$

By Proposition 3.22, the above operations do not depend on the specific choice of representatives in the equivalence classes $[u], [v] \in E/M$. It is also immediate to verify that E/M is a vector space. The function $\pi\colon E \to E/F$, defined such that $\pi(u) = [u]$ for every $u \in E$, is a surjective linear map called the *natural projection of E onto E/F*. The vector space E/M is called the *quotient space of E by the subspace M*.

Given any linear map $f\colon E \to F$, we know that $\mathrm{Ker}\, f$ is a subspace of E, and it is immediately verified that $\mathrm{Im}\, f$ is isomorphic to the quotient space $E/\mathrm{Ker}\, f$.

3.9 Linear Forms and the Dual Space

We already observed that the field K itself ($K = \mathbb{R}$ or $K = \mathbb{C}$) is a vector space (over itself). The vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K, the linear forms, plays a particular role. In this section, we only define linear forms and show that every finite-dimensional vector space has a dual basis. A more advanced presentation of dual spaces and duality is given in Chapter 11.

Definition 3.26. Given a vector space E, the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K is called the *dual space (or dual)* of E. The space $\mathrm{Hom}(E, K)$ is also denoted by E^*, and the linear maps in E^* are called *the linear forms*, or *covectors*. The dual space E^{**} of the space E^* is called the *bidual* of E.

As a matter of notation, linear forms $f\colon E \to K$ will also be denoted by starred symbol, such as u^*, x^*, *etc.*

If E is a vector space of finite dimension n and $(u_1, \ldots, u_n)$ is a basis of E, for any linear form $f^* \in E^*$, for every $x = x_1u_1 + \cdots + x_nu_n \in E$, by linearity we have

$$\begin{aligned} f^*(x) &= f^*(u_1)x_1 + \cdots + f^*(u_n)x_n \\ &= \lambda_1 x_1 + \cdots + \lambda_n x_n, \end{aligned}$$

with $\lambda_i = f^*(u_i) \in K$ for every i, $1 \leq i \leq n$. Thus, with respect to the basis $(u_1, \ldots, u_n)$, the linear form f^* is represented by the row vector

$$\begin{pmatrix} \lambda_1 & \cdots & \lambda_n \end{pmatrix},$$

we have

$$f^*(x) = \begin{pmatrix} \lambda_1 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

a linear combination of the coordinates of x, and we can view the linear form f^* as a *linear equation*. If we decide to use a column vector of coefficients

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

instead of a row vector, then the linear form f^* is defined by

$$f^*(x) = c^\top x.$$

The above notation is often used in machine learning.

Example 3.7. Given any differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$, by definition, for any $x \in \mathbb{R}^n$, the *total derivative* df_x of f at x is the linear form $df_x\colon \mathbb{R}^n \to \mathbb{R}$ defined so that for all $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$df_x(u) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x)\, u_i.$$

Example 3.8. Let $\mathcal{C}([0,1])$ be the vector space of continuous functions $f\colon [0,1] \to \mathbb{R}$. The map $\mathcal{I}\colon \mathcal{C}([0,1]) \to \mathbb{R}$ given by

$$\mathcal{I}(f) = \int_0^1 f(x)dx \quad \text{for any } f \in \mathcal{C}([0,1])$$

is a linear form (integration).

Example 3.9. Consider the vector space $\mathrm{M}_n(\mathbb{R})$ of real $n \times n$ matrices. Let $\mathrm{tr}\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn},$$

called the *trace* of A. It is a linear form. Let $s\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$s(A) = \sum_{i,j=1}^n a_{ij},$$

where $A = (a_{ij})$. It is immediately verified that s is a linear form.

Given a vector space E and any basis $(u_i)_{i\in I}$ for E, we can associate to each u_i a linear form $u_i^* \in E^*$, and the u_i^* have some remarkable properties.

Definition 3.27. Given a vector space E and any basis $(u_i)_{i\in I}$ for E, by Proposition 3.18, for every $i \in I$, there is a unique linear form u_i^* such that

$$u_i^*(u_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

for every $j \in I$. The linear form u_i^* is called the *coordinate form* of index i w.r.t. the basis $(u_i)_{i\in I}$.

Remark: Given an index set I, authors often define the so called "Kronecker symbol" $\delta_{i\,j}$ such that

$$\delta_{i\,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

for all $i, j \in I$. Then, $u_i^*(u_j) = \delta_{i\,j}$.

The reason for the terminology *coordinate form* is as follows: If E has finite dimension and if $(u_1, \ldots, u_n)$ is a basis of E, for any vector

$$v = \lambda_1 u_1 + \cdots + \lambda_n u_n,$$

we have

$$\begin{aligned} u_i^*(v) &= u_i^*(\lambda_1 u_1 + \cdots + \lambda_n u_n) \\ &= \lambda_1 u_i^*(u_1) + \cdots + \lambda_i u_i^*(u_i) + \cdots + \lambda_n u_i^*(u_n) \\ &= \lambda_i, \end{aligned}$$

since $u_i^*(u_j) = \delta_{i\,j}$. Therefore, u_i^* is the linear function that returns the ith coordinate of a vector expressed over the basis $(u_1, \ldots, u_n)$.

The following theorem shows that in finite-dimension, every basis $(u_1, \ldots, u_n)$ of a vector space E yields a basis $(u_1^*, \ldots, u_n^*)$ of the dual space E^*, called a *dual basis*.

Theorem 3.23. *(Existence of dual bases) Let E be a vector space of dimension n. The following properties hold: For every basis $(u_1, \ldots, u_n)$ of E, the family of coordinate forms $(u_1^*, \ldots, u_n^*)$ is a basis of E^* (called the dual basis of $(u_1, \ldots, u_n)$).*

Proof. (a) If $v^* \in E^*$ is any linear form, consider the linear form

$$f^* = v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*.$$

Observe that because $u_i^*(u_j) = \delta_{i\,j}$,

$$\begin{aligned} f^*(u_i) &= (v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*)(u_i) \\ &= v^*(u_1)u_1^*(u_i) + \cdots + v^*(u_i)u_i^*(u_i) + \cdots + v^*(u_n)u_n^*(u_i) \\ &= v^*(u_i), \end{aligned}$$

and so f^* and v^* agree on the basis $(u_1, \ldots, u_n)$, which implies that

$$v^* = f^* = v^*(u_1)u_1^* + \cdots + v^*(u_n)u_n^*.$$

Therefore, $(u_1^*, \ldots, u_n^*)$ spans E^*. We claim that the covectors $u_1^*, \ldots, u_n^*$ are linearly independent. If not, we have a nontrivial linear dependence

$$\lambda_1 u_1^* + \cdots + \lambda_n u_n^* = 0,$$

and if we apply the above linear form to each u_i, using a familar computation, we get

$$0 = \lambda_i u_i^*(u_i) = \lambda_i,$$

proving that $u_1^*, \ldots, u_n^*$ are indeed linearly independent. Therefore, $(u_1^*, \ldots, u_n^*)$ is a basis of E^*. $\square$

In particular, Theorem 3.23 shows a finite-dimensional vector space and its dual E^* have the same dimension.

Example 3.10. Consider the vector space $\mathbb{R}[X]_n$ of polynomials in the variable X of degree at most n with real coefficients. We know that it has the basis $(1, X, X^2, \ldots, X^n)$. If we write $e_i = X^i$, the dual basis vectors e_i^* are the linear forms defined on polynomials which are uniquely defined by the equations

$$e_i^*(e_j) = e_i^*(X^j) = \delta_{ij}.$$

We can identify explicitly the linear forms e_i^*. Indeed, for any i such that $0 \leq i \leq n$, if $D^{(i)}$ is the linear form defined by $D^{(i)} = \underbrace{D \circ \cdots \circ D}_{i}$, where D is the differentiation operator on polynomials, in other words, for any polynomial $P \in \mathbb{R}[X]_n$, $D^{(i)}P$ is the result of differentiating P i times (with $D^{(0)} = \mathrm{id}$), then e_i^* given by

$$e_i^*(P) = \frac{1}{i!}(D^{(i)}P)(0)$$

does the job, since

$$D^{(i)}X^j = \begin{cases} i!X^{j-i}, & \text{if } j > i \\ i!, & \text{if } j = i \\ 0, & \text{if } j < i, \end{cases}$$

so when we evaluate $D^{(i)}X^j$ at 0, we get δ_{ij}.

We explained just after Definition 3.26 that if the space E is finite-dimensional and has a finite basis $(u_1, \ldots, u_n)$, then a linear form $f^*\colon E \to K$ is represented by the *row vector* of coefficients

$$\begin{pmatrix} f^*(u_1) & \cdots & f^*(u_n) \end{pmatrix}. \tag{1}$$

The proof of Theorem 3.23 shows that over the dual basis $(u_1^*, \ldots, u_n^*)$ of E^*, the linear form f^* is represented by the same coefficients, but as the *column vector*

$$\begin{pmatrix} f^*(u_1) \\ \vdots \\ f^*(u_n) \end{pmatrix}, \tag{2}$$

which is the transpose of the row vector in (1).

3.10 Summary

The main concepts and results of this chapter are listed below:

- The notion of a *vector space*.
- *Families* of vectors.
- *Linear combinations* of vectors; *linear dependence* and *linear independence* of a family of vectors.
- Linear *subspaces*.
- *Spanning* (or *generating*) family; *generators*, *finitely generated subspace*; *basis of a subspace*.
- *Every linearly independent family can be extended to a basis* (Theorem 3.7).
- A family B of vectors is a basis iff it is a maximal linearly independent family iff it is a minimal generating family (Proposition 3.8).
- The replacement lemma (Proposition 3.10).
- Any two bases in a finitely generated vector space E have the *same number of elements*; this is the *dimension* of E (Theorem 3.11).
- *Hyperplanes*.
- Every vector has a *unique representation* over a basis (in terms of its coordinates).
- *Matrices*
- *Column vectors*, *row vectors*.
- *Matrix operations*: addition, scalar multiplication, multiplication.
- The vector space $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices over the field K; The ring $\mathrm{M}_n(K)$ of $n \times n$ matrices over the field K.

- The notion of a *linear map*.
- The *image* $\operatorname{Im} f$ (or *range*) of a linear map f.
- The *kernel* $\operatorname{Ker} f$ (or *nullspace*) of a linear map f.
- The *rank* $\operatorname{rk}(f)$ of a linear map f.
- The image and the kernel of a linear map are subspaces. A linear map is injective iff its kernel is the trivial space (0) (Proposition 3.17).
- The *unique homomorphic extension property* of linear maps with respect to bases (Proposition 3.18).
- *Quotient spaces*.
- The vector space of linear maps $\operatorname{Hom}_K(E, F)$.
- Linear forms (covectors) and the *dual space* E^*.
- Coordinate forms.
- The existence of *dual bases* (in finite dimension).

3.11 Problems

Problem 3.1. Let H be the set of 3×3 upper triangular matrices given by

$$H = \left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \mid a, b, c \in \mathbb{R} \right\}.$$

(1) Prove that H with the binary operation of matrix multiplication is a group; find explicitly the inverse of every matrix in H. Is H abelian (commutative)?

(2) Given two groups G_1 and G_2, recall that a *homomorphism* if a function $\varphi\colon G_1 \to G_2$ such that

$$\varphi(ab) = \varphi(a)\varphi(b), \quad a, b \in G_1.$$

Prove that $\varphi(e_1) = e_2$ (where e_i is the identity element of G_i) and that

$$\varphi(a^{-1}) = (\varphi(a))^{-1}, \quad a \in G_1.$$

(3) Let S^1 be the unit circle, that is

$$S^1 = \{e^{i\theta} = \cos\theta + i \sin\theta \mid 0 \leq \theta < 2\pi\},$$

and let φ be the function given by

$$\varphi\begin{pmatrix}1 & a & b\\ 0 & 1 & c\\ 0 & 0 & 1\end{pmatrix} = (a, c, e^{ib}).$$

Prove that φ is a surjective function onto $G = \mathbb{R} \times \mathbb{R} \times S^1$, and that if we define multiplication on this set by

$$(x_1, y_1, u_1) \cdot (x_2, y_2, u_2) = (x_1 + x_2, y_1 + y_2, e^{ix_1y_2}u_1u_2),$$

then G is a group and φ is a group homomorphism from H onto G.

(4) The *kernel* of a homomorphism $\varphi\colon G_1 \to G_2$ is defined as

$$\mathrm{Ker}\,(\varphi) = \{a \in G_1 \mid \varphi(a) = e_2\}.$$

Find explicitly the kernel of φ and show that it is a subgroup of H.

Problem 3.2. For any $m \in \mathbb{Z}$ with $m > 0$, the subset $m\mathbb{Z} = \{mk \mid k \in \mathbb{Z}\}$ is an abelian subgroup of $\mathbb{Z}$. Check this.

(1) Give a group isomorphism (an invertible homomorphism) from $m\mathbb{Z}$ to $\mathbb{Z}$.

(2) Check that the inclusion map $i\colon m\mathbb{Z} \to \mathbb{Z}$ given by $i(mk) = mk$ is a group homomorphism. Prove that if $m \geq 2$ then there is no group homomorphism $p\colon \mathbb{Z} \to m\mathbb{Z}$ such that $p \circ i = \mathrm{id}$.

Remark: The above shows that abelian groups fail to have some of the properties of vector spaces. We will show later that a linear map satisfying the condition $p \circ i = \mathrm{id}$ always exists.

Problem 3.3. Let $E = \mathbb{R} \times \mathbb{R}$, and define the addition operation

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1, +y_2), \quad x_1, x_2, y_1, y_2 \in \mathbb{R},$$

and the multiplication operation $\cdot\colon \mathbb{R} \times E \to E$ by

$$\lambda \cdot (x, y) = (\lambda x, y), \quad \lambda, x, y \in \mathbb{R}.$$

Show that E with the above operations $+$ and $\cdot$ is not a vector space. Which of the axioms is violated?

Problem 3.4. (1) Prove that the axioms of vector spaces imply that

$$\begin{aligned}\alpha \cdot 0 &= 0\\ 0 \cdot v &= 0\\ \alpha \cdot (-v) &= -(\alpha \cdot v)\\ (-\alpha) \cdot v &= -(\alpha \cdot v),\end{aligned}$$

for all $v \in E$ and all $\alpha \in K$, where E is a vector space over K.

(2) For every $\lambda \in \mathbb{R}$ and every $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, define λx by

$$\lambda x = \lambda(x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n).$$

Recall that every vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ can be written uniquely as

$$x = x_1 e_1 + \cdots + x_n e_n,$$

where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, with a single 1 in position i. For any operation $\cdot \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, if $\cdot$ satisfies the Axiom (V1) of a vector space, then prove that for any $\alpha \in \mathbb{R}$, we have

$$\alpha \cdot x = \alpha \cdot (x_1 e_1 + \cdots + x_n e_n) = \alpha \cdot (x_1 e_1) + \cdots + \alpha \cdot (x_n e_n).$$

Conclude that $\cdot$ is completely determined by its action on the one-dimensional subspaces of $\mathbb{R}^n$ spanned by $e_1, \ldots, e_n$.

(3) Use (2) to define operations $\cdot \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ that satisfy the Axioms (V1–V3), but for which Axiom V4 fails.

(4) For any operation $\cdot \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, prove that if $\cdot$ satisfies the Axioms (V2–V3), then for every rational number $r \in \mathbb{Q}$ and every vector $x \in \mathbb{R}^n$, we have

$$r \cdot x = r(1 \cdot x).$$

In the above equation, $1 \cdot x$ is some vector $(y_1, \ldots, y_n) \in \mathbb{R}^n$ not necessarily equal to $x = (x_1, \ldots, x_n)$, and

$$r(1 \cdot x) = (ry_1, \ldots, ry_n),$$

as in Part (2).

Use (4) to conclude that any operation $\cdot \colon \mathbb{Q} \times \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the Axioms (V1–V3) is completely determined by the action of 1 on the one-dimensional subspaces of $\mathbb{R}^n$ spanned by $e_1, \ldots, e_n$.

Problem 3.5. Let A_1 be the following matrix:

$$A_1 = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & -1 \\ -3 & -5 & 1 \end{pmatrix}.$$

Prove that the columns of A_1 are linearly independent. Find the coordinates of the vector $x = (6, 2, -7)$ over the basis consisting of the column vectors of A_1.

Problem 3.6. Let A_2 be the following matrix:

$$A_2 = \begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 3 & 0 \end{pmatrix}.$$

Express the fourth column of A_2 as a linear combination of the first three columns of A_2. Is the vector $x = (7, 14, -1, 2)$ a linear combination of the columns of A_2?

Problem 3.7. Let A_3 be the following matrix:

$$A_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix}.$$

Prove that the columns of A_1 are linearly independent. Find the coordinates of the vector $x = (6, 9, 14)$ over the basis consisting of the column vectors of A_3.

Problem 3.8. Let A_4 be the following matrix:

$$A_4 = \begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 4 & 0 \end{pmatrix}.$$

Prove that the columns of A_4 are linearly independent. Find the coordinates of the vector $x = (7, 14, -1, 2)$ over the basis consisting of the column vectors of A_4.

Problem 3.9. Consider the following Haar matrix

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Prove that the columns of H are linearly independent.

Hint. Compute the product $H^\top H$.

Problem 3.10. Consider the following Hadamard matrix

$$H_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Prove that the columns of H_4 are linearly independent.

Hint. Compute the product $H_4^\top H_4$.

Problem 3.11. In solving this problem, **do not use determinants**.

(1) Let $(u_1, \ldots, u_m)$ and $(v_1, \ldots, v_m)$ be two families of vectors in some vector space E. Assume that each v_i is a linear combination of the u_js, so that

$$v_i = a_{i\,1}u_1 + \cdots + a_{i\,m}u_m, \quad 1 \le i \le m,$$

and that the matrix $A = (a_{i\,j})$ is an upper-triangular matrix, which means that if $1 \leq j < i \leq m$, then $a_{i\,j} = 0$. Prove that if $(u_1, \ldots, u_m)$ are linearly independent and if all the diagonal entries of A are nonzero, then $(v_1, \ldots, v_m)$ are also linearly independent.

Hint. Use induction on m.

(2) Let $A = (a_{i\,j})$ be an upper-triangular matrix. Prove that if all the diagonal entries of A are nonzero, then A is invertible and the inverse A^{-1} of A is also upper-triangular.

Hint. Use induction on m.

Prove that if A is invertible, then all the diagonal entries of A are nonzero.

(3) Prove that if the families $(u_1, \ldots, u_m)$ and $(v_1, \ldots, v_m)$ are related as in (1), then $(u_1, \ldots, u_m)$ are linearly independent iff $(v_1, \ldots, v_m)$ are linearly independent.

Problem 3.12. In solving this problem, **do not use determinants**. Consider the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 2 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & 2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 & 2 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 & 2 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(1) Find the solution $x = (x_1, \ldots, x_n)$ of the linear system

$$Ax = b,$$

for

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

(2) Prove that the matrix A is invertible and find its inverse A^{-1}. Given that the number of atoms in the universe is estimated to be $\leq 10^{82}$, compare the size of the coefficients the inverse of A to 10^{82}, if $n \geq 300$.

(3) Assume b is perturbed by a small amount δb (note that δb is a vector). Find the new solution of the system

$$A(x + \delta x) = b + \delta b,$$

where δx is also a vector. In the case where $b = (0, \ldots, 0, 1)$, and $\delta b = (0, \ldots, 0, \epsilon)$, show that

$$|(\delta x)_1| = 2^{n-1}|\epsilon|.$$

(where $(\delta x)_1$ is the first component of δx).

(4) Prove that $(A - I)^n = 0$.

Problem 3.13. An $n \times n$ matrix N is *nilpotent* if there is some integer $r \geq 1$ such that $N^r = 0$.

(1) Prove that if N is a nilpotent matrix, then the matrix $I - N$ is invertible and

$$(I - N)^{-1} = I + N + N^2 + \cdots + N^{r-1}.$$

(2) Compute the inverse of the following matrix A using (1):

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Problem 3.14. (1) Let A be an $n \times n$ matrix. If A is invertible, prove that for any $x \in \mathbb{R}^n$, if $Ax = 0$, then $x = 0$.

(2) Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix. Prove that $I_m - AB$ is invertible iff $I_n - BA$ is invertible.

Hint. If for all $x \in \mathbb{R}^n$, $Mx = 0$ implies that $x = 0$, then M is invertible.

Problem 3.15. Consider the following $n \times n$ matrix, for $n \geq 3$:

$$B = \begin{pmatrix} 1 & -1 & -1 & -1 & \cdots & -1 & -1 \\ 1 & -1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & -1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & -1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & -1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & -1 \end{pmatrix}$$

(1) If we denote the columns of B by $b_1, \ldots, b_n$, prove that

$$\begin{aligned} (n-3)b_1 - (b_2 + \cdots + b_n) &= 2(n-2)e_1 \\ b_1 - b_2 &= 2(e_1 + e_2) \\ b_1 - b_3 &= 2(e_1 + e_3) \\ \vdots \quad &\quad \vdots \\ b_1 - b_n &= 2(e_1 + e_n), \end{aligned}$$

where $e_1, \ldots, e_n$ are the canonical basis vectors of $\mathbb{R}^n$.

(2) Prove that B is invertible and that its inverse $A = (a_{ij})$ is given by

$$a_{11} = \frac{(n-3)}{2(n-2)}, \quad a_{i1} = -\frac{1}{2(n-2)} \quad 2 \leq i \leq n$$

and

$$\begin{aligned} a_{ii} &= -\frac{(n-3)}{2(n-2)}, \quad 2 \leq i \leq n \\ a_{ji} &= \frac{1}{2(n-2)}, \quad 2 \leq i \leq n, j \neq i. \end{aligned}$$

(3) Show that the n diagonal $n \times n$ matrices D_i defined such that the diagonal entries of D_i are equal the entries (from top down) of the ith column of B form a basis of the space of $n \times n$ diagonal matrices (matrices with zeros everywhere except possibly on the diagonal). For example, when $n = 4$, we have

$$D_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad D_2 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$D_3 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad D_4 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Problem 3.16. Given any $m \times n$ matrix A and any $n \times p$ matrix B, if we denote the columns of A by $A^1, \ldots, A^n$ and the rows of B by $B_1, \ldots, B_n$, prove that

$$AB = A^1 B_1 + \cdots + A^n B_n.$$

Problem 3.17. Let $f \colon E \to F$ be a linear map which is also a bijection (it is injective and surjective). Prove that the inverse function $f^{-1} \colon F \to E$ is linear.

Problem 3.18. Given two vectors spaces E and F, let $(u_i)_{i \in I}$ be any basis of E and let $(v_i)_{i \in I}$ be any family of vectors in F. Prove that the unique linear map $f \colon E \to F$ such that $f(u_i) = v_i$ for all $i \in I$ is surjective iff $(v_i)_{i \in I}$ spans F.

Problem 3.19. Let $f \colon E \to F$ be a linear map with $\dim(E) = n$ and $\dim(F) = m$. Prove that f has rank 1 iff f is represented by an $m \times n$ matrix of the form

$$A = uv^\top$$

with u a nonzero column vector of dimension m and v a nonzero column vector of dimension n.

Problem 3.20. Find a nontrivial linear dependence among the linear forms

$$\varphi_1(x, y, z) = 2x - y + 3z, \quad \varphi_2(x, y, z) = 3x - 5y + z, \quad \varphi_3(x, y, z) = 4x - 7y + z.$$

Problem 3.21. Prove that the linear forms

$$\varphi_1(x, y, z) = x + 2y + z, \quad \varphi_2(x, y, z) = 2x + 3y + 3z, \quad \varphi_3(x, y, z) = 3x + 7y + z$$

are linearly independent. Express the linear form $\varphi(x, y, z) = x+y+z$ as a linear combination of $\varphi_1, \varphi_2, \varphi_3$.

Chapter 4

Matrices and Linear Maps

In this chapter, all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$.

4.1 Representation of Linear Maps by Matrices

Proposition 3.18 shows that given two vector spaces E and F and a basis $(u_j)_{j\in J}$ of E, every linear map $f\colon E \to F$ is uniquely determined by the family $(f(u_j))_{j\in J}$ of the images under f of the vectors in the basis $(u_j)_{j\in J}$.

If we also have a basis $(v_i)_{i\in I}$ of F, then every vector $f(u_j)$ can be written in a unique way as

$$f(u_j) = \sum_{i\in I} a_{i\,j} v_i,$$

where $j \in J$, for a family of scalars $(a_{i\,j})_{i\in I}$. Thus, with respect to the two bases $(u_j)_{j\in J}$ of E and $(v_i)_{i\in I}$ of F, the linear map f is completely determined by a "$I \times J$-matrix" $M(f) = (a_{i\,j})_{(i,j)\in I\times J}$.

Remark: Note that we intentionally assigned the index set J to the basis $(u_j)_{j\in J}$ of E, and the index set I to the basis $(v_i)_{i\in I}$ of F, so that the rows of the matrix $M(f)$ associated with $f\colon E \to F$ are indexed by I, and the columns of the matrix $M(f)$ are indexed by J. Obviously, this causes a mildly unpleasant reversal. If we had considered the bases $(u_i)_{i\in I}$ of E and $(v_j)_{j\in J}$ of F, we would obtain a $J \times I$-matrix $M(f) = (a_{j\,i})_{(j,i)\in J\times I}$. No matter what we do, there will be a reversal! We decided to stick to the bases $(u_j)_{j\in J}$ of E and $(v_i)_{i\in I}$ of F, so that we get an $I \times J$-matrix $M(f)$, knowing that we may occasionally suffer from this decision!

When I and J are finite, and say, when $|I| = m$ and $|J| = n$, the linear map f is determined by the matrix $M(f)$ whose entries in the j-th column are the components of the vector $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, that is, the matrix

$$M(f) = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix}$$

whose entry on Row i and Column j is $a_{i\,j}$ ($1 \leq i \leq m$, $1 \leq j \leq n$).

We will now show that when E *and* F *have finite dimension, linear maps can be very conveniently represented by matrices, and that composition of linear maps corresponds to matrix multiplication.* We will follow rather closely an elegant presentation method due to Emil Artin.

Let E and F be two vector spaces, and assume that E has a finite basis $(u_1, \ldots, u_n)$ and that F has a finite basis $(v_1, \ldots, v_m)$. Recall that we have shown that every vector $x \in E$ can be written in a unique way as

$$x = x_1 u_1 + \cdots + x_n u_n,$$

and similarly every vector $y \in F$ can be written in a unique way as

$$y = y_1 v_1 + \cdots + y_m v_m.$$

Let $f \colon E \to F$ be a linear map between E and F. Then for every $x = x_1 u_1 + \cdots + x_n u_n$ in E, by linearity, we have

$$f(x) = x_1 f(u_1) + \cdots + x_n f(u_n).$$

Let

$$f(u_j) = a_{1\,j} v_1 + \cdots + a_{m\,j} v_m,$$

or more concisely,

$$f(u_j) = \sum_{i=1}^{m} a_{i\,j} v_i,$$

for every j, $1 \leq j \leq n$. This can be expressed by writing the coefficients $a_{1j}, a_{2j}, \ldots, a_{mj}$ of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, as the jth column of a matrix, as shown below:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_m \end{array} \begin{array}{c} \begin{array}{cccc} f(u_1) & f(u_2) & \ldots & f(u_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \end{array}.$$

Then substituting the right-hand side of each $f(u_j)$ into the expression for $f(x)$, we get

$$f(x) = x_1 (\sum_{i=1}^{m} a_{i\,1} v_i) + \cdots + x_n (\sum_{i=1}^{m} a_{i\,n} v_i),$$

which, by regrouping terms to obtain a linear combination of the v_i, yields

$$f(x) = (\sum_{j=1}^{n} a_{1\,j}x_j)v_1 + \cdots + (\sum_{j=1}^{n} a_{m\,j}x_j)v_m.$$

Thus, letting $f(x) = y = y_1v_1 + \cdots + y_mv_m$, we have

$$y_i = \sum_{j=1}^{n} a_{i\,j}x_j \tag{1}$$

for all i, $1 \leq i \leq m$.

To make things more concrete, let us treat the case where $n = 3$ and $m = 2$. In this case,

$$\begin{aligned} f(u_1) &= a_{11}v_1 + a_{21}v_2 \\ f(u_2) &= a_{12}v_1 + a_{22}v_2 \\ f(u_3) &= a_{13}v_1 + a_{23}v_2, \end{aligned}$$

which in matrix form is expressed by

$$\begin{array}{c c} & \begin{array}{ccc} f(u_1) & f(u_2) & f(u_3) \end{array} \\ \begin{array}{c} v_1 \\ v_2 \end{array} & \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \end{array},$$

and for any $x = x_1u_1 + x_2u_2 + x_3u_3$, we have

$$\begin{aligned} f(x) &= f(x_1u_1 + x_2u_2 + x_3u_3) \\ &= x_1f(u_1) + x_2f(u_2) + x_3f(u_3) \\ &= x_1(a_{11}v_1 + a_{21}v_2) + x_2(a_{12}v_1 + a_{22}v_2) + x_3(a_{13}v_1 + a_{23}v_2) \\ &= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)v_1 + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)v_2. \end{aligned}$$

Consequently, since

$$y = y_1v_1 + y_2v_2,$$

we have

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3. \end{aligned}$$

This agrees with the matrix equation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

We now formalize the representation of linear maps by matrices.

Definition 4.1. Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis for E, and $(v_1, \ldots, v_m)$ be a basis for F. Each vector $x \in E$ expressed in the basis $(u_1, \ldots, u_n)$ as $x = x_1u_1 + \cdots + x_nu_n$ is represented by the column matrix

$$M(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and similarly for each vector $y \in F$ expressed in the basis $(v_1, \ldots, v_m)$.

Every linear map $f\colon E \to F$ is represented by the matrix $M(f) = (a_{i\,j})$, where $a_{i\,j}$ is the i-th component of the vector $f(u_j)$ over the basis $(v_1, \ldots, v_m)$, i.e., where

$$f(u_j) = \sum_{i=1}^{m} a_{i\,j} v_i, \quad \text{for every } j,\ 1 \leq j \leq n.$$

The coefficients $a_{1j}, a_{2j}, \ldots, a_{mj}$ of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$ form the jth column of the matrix $M(f)$ shown below:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_m \end{array} \begin{array}{c} \begin{array}{cccc} f(u_1) & f(u_2) & \ldots & f(u_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \end{array}.$$

The matrix $M(f)$ associated with the linear map $f\colon E \to F$ is called the *matrix of f with respect to the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$*. When $E = F$ and the basis $(v_1, \ldots, v_m)$ is identical to the basis $(u_1, \ldots, u_n)$ of E, the matrix $M(f)$ associated with $f\colon E \to E$ (as above) is called the *matrix of f with respect to the basis $(u_1, \ldots, u_n)$*.

Remark: As in the remark after Definition 3.12, there is no reason to assume that the vectors in the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ are ordered in any particular way. However, it is often convenient to assume the natural ordering. When this is so, authors sometimes refer to the matrix $M(f)$ as the matrix of f with respect to the *ordered bases* $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$.

Let us illustrate the representation of a linear map by a matrix in a concrete situation. Let E be the vector space $\mathbb{R}[X]_4$ of polynomials of degree at most 4, let F be the vector space $\mathbb{R}[X]_3$ of polynomials of degree at most 3, and let the linear map be the derivative map d: that is,

$$\begin{aligned} d(P + Q) &= dP + dQ \\ d(\lambda P) &= \lambda dP, \end{aligned}$$

with $\lambda \in \mathbb{R}$. We choose $(1, x, x^2, x^3, x^4)$ as a basis of E and $(1, x, x^2, x^3)$ as a basis of F. Then the 4×5 matrix D associated with d is obtained by expressing the derivative dx^i of each basis vector x^i for $i = 0, 1, 2, 3, 4$ over the basis $(1, x, x^2, x^3)$. We find

$$D = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

If P denotes the polynomial

$$P = 3x^4 - 5x^3 + x^2 - 7x + 5,$$

we have

$$dP = 12x^3 - 15x^2 + 2x - 7.$$

The polynomial P is represented by the vector $(5, -7, 1, -5, 3)$, the polynomial dP is represented by the vector $(-7, 2, -15, 12)$, and we have

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ -7 \\ 1 \\ -5 \\ 3 \end{pmatrix} = \begin{pmatrix} -7 \\ 2 \\ -15 \\ 12 \end{pmatrix},$$

as expected! The kernel (nullspace) of d consists of the polynomials of degree 0, that is, the constant polynomials. Therefore $\dim(\mathrm{Ker}\, d) = 1$, and from

$$\dim(E) = \dim(\mathrm{Ker}\, d) + \dim(\mathrm{Im}\, d)$$

(see Theorem 6.16), we get $\dim(\mathrm{Im}\, d) = 4$ (since $\dim(E) = 5$).

For fun, let us figure out the linear map from the vector space $\mathbb{R}[X]_3$ to the vector space $\mathbb{R}[X]_4$ given by integration (finding the primitive, or anti-derivative) of x^i, for $i = 0, 1, 2, 3$). The 5×4 matrix S representing $\int$ with respect to the same bases as before is

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}.$$

We verify that $DS = I_4$,

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This is to be expected by the fundamental theorem of calculus since the derivative of an integral returns the function. As we will shortly see, the above matrix product corresponds to this functional composition. The equation $DS = I_4$ shows that S is injective and has D as a left inverse. However, $SD \neq I_5$, and instead

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

because constant polynomials (polynomials of degree 0) belong to the kernel of D.

4.2 Composition of Linear Maps and Matrix Multiplication

Let us now consider how the composition of linear maps is expressed in terms of bases.

Let E, F, and G, be three vectors spaces with respective bases $(u_1, \ldots, u_p)$ for E, $(v_1, \ldots, v_n)$ for F, and $(w_1, \ldots, w_m)$ for G. Let $g\colon E \to F$ and $f\colon F \to G$ be linear maps. As explained earlier, $g\colon E \to F$ is determined by the images of the basis vectors u_j, and $f\colon F \to G$ is determined by the images of the basis vectors v_k. We would like to understand how $f \circ g\colon E \to G$ is determined by the images of the basis vectors u_j.

Remark: Note that we are considering linear maps $g\colon E \to F$ and $f\colon F \to G$, instead of $f\colon E \to F$ and $g\colon F \to G$, which yields the composition $f \circ g\colon E \to G$ instead of $g \circ f\colon E \to G$. Our perhaps unusual choice is motivated by the fact that if f is represented by a matrix $M(f) = (a_{i\,k})$ and g is represented by a matrix $M(g) = (b_{k\,j})$, then $f \circ g\colon E \to G$ is represented by the product AB of the matrices A and B. If we had adopted the other choice where $f\colon E \to F$ and $g\colon F \to G$, then $g \circ f\colon E \to G$ would be represented by the product BA. Personally, we find it easier to remember the formula for the entry in Row i and Column j of the product of two matrices when this product is written by AB, rather than BA. Obviously, this is a matter of taste! We will have to live with our perhaps unorthodox choice.

Thus, let

$$f(v_k) = \sum_{i=1}^{m} a_{i\,k} w_i,$$

for every k, $1 \leq k \leq n$, and let

$$g(u_j) = \sum_{k=1}^{n} b_{k\,j} v_k,$$

for every j, $1 \leq j \leq p$; in matrix form, we have

$$\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{cccc} f(v_1) & f(v_2) & \dots & f(v_n) \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \end{array}$$

and

$$\begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{array} \begin{array}{c} \begin{array}{cccc} g(u_1) & g(u_2) & \dots & g(u_p) \end{array} \\ \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix} \end{array}.$$

By previous considerations, for every

$$x = x_1u_1 + \cdots + x_pu_p,$$

letting $g(x) = y = y_1v_1 + \cdots + y_nv_n$, we have

$$y_k = \sum_{j=1}^{p} b_{k\,j}x_j \tag{2}$$

for all k, $1 \leq k \leq n$, and for every

$$y = y_1v_1 + \cdots + y_nv_n,$$

letting $f(y) = z = z_1w_1 + \cdots + z_mw_m$, we have

$$z_i = \sum_{k=1}^{n} a_{i\,k}y_k \tag{3}$$

for all i, $1 \leq i \leq m$. Then if $y = g(x)$ and $z = f(y)$, we have $z = f(g(x))$, and in view of (2) and (3), we have

$$\begin{aligned} z_i &= \sum_{k=1}^{n} a_{i\,k}(\sum_{j=1}^{p} b_{k\,j}x_j) \\ &= \sum_{k=1}^{n}\sum_{j=1}^{p} a_{i\,k}b_{k\,j}x_j \\ &= \sum_{j=1}^{p}\sum_{k=1}^{n} a_{i\,k}b_{k\,j}x_j \\ &= \sum_{j=1}^{p}(\sum_{k=1}^{n} a_{i\,k}b_{k\,j})x_j. \end{aligned}$$

Thus, defining $c_{i\,j}$ such that

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j},$$

for $1 \leq i \leq m$, and $1 \leq j \leq p$, we have

$$z_i = \sum_{j=1}^{p} c_{i\,j} x_j \tag{4}$$

Identity (4) shows that the composition of linear maps corresponds to the product of matrices.

Then given a linear map $f\colon E \to F$ represented by the matrix $M(f) = (a_{i\,j})$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$, by Equation (1), namely

$$y_i = \sum_{j=1}^{n} a_{i\,j} x_j \quad 1 \leq i \leq m,$$

and the definition of matrix multiplication, the equation $y = f(x)$ corresponds to the matrix equation $M(y) = M(f)M(x)$, that is,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{1\,1} & \ldots & a_{1\,n} \\ \vdots & \ddots & \vdots \\ a_{m\,1} & \ldots & a_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Recall that

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m\,1} & a_{m\,2} & \ldots & a_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} a_{1\,1} \\ a_{2\,1} \\ \vdots \\ a_{m\,1} \end{pmatrix} + x_2 \begin{pmatrix} a_{1\,2} \\ a_{2\,2} \\ \vdots \\ a_{m\,2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1\,n} \\ a_{2\,n} \\ \vdots \\ a_{m\,n} \end{pmatrix}.$$

Sometimes, it is necessary to incorporate the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ in the notation for the matrix $M(f)$ expressing f with respect to these bases. This turns out to be a messy enterprise!

We propose the following course of action:

Definition 4.2. Write $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_m)$ for the bases of E and F, and denote by $M_{\mathcal{U},\mathcal{V}}(f)$ the *matrix of f with respect to the bases $\mathcal{U}$ and $\mathcal{V}$*. Furthermore, write $x_{\mathcal{U}}$ for the coordinates $M(x) = (x_1, \ldots, x_n)$ of $x \in E$ w.r.t. the basis $\mathcal{U}$ and write $y_{\mathcal{V}}$ for the coordinates $M(y) = (y_1, \ldots, y_m)$ of $y \in F$ w.r.t. the basis $\mathcal{V}$. Then

$$y = f(x)$$

is expressed in matrix form by

$$y_{\mathcal{V}} = M_{\mathcal{U},\mathcal{V}}(f)\, x_{\mathcal{U}}.$$

When $\mathcal{U} = \mathcal{V}$, we abbreviate $M_{\mathcal{U},\mathcal{V}}(f)$ as $M_{\mathcal{U}}(f)$.

The above notation seems reasonable, but it has the slight disadvantage that in the expression $M_{\mathcal{U},\mathcal{V}}(f)x_{\mathcal{U}}$, the input argument $x_{\mathcal{U}}$ which is fed to the matrix $M_{\mathcal{U},\mathcal{V}}(f)$ does not appear next to the subscript $\mathcal{U}$ in $M_{\mathcal{U},\mathcal{V}}(f)$. We could have used the notation $M_{\mathcal{V},\mathcal{U}}(f)$, and some people do that. But then, we find a bit confusing that $\mathcal{V}$ comes before $\mathcal{U}$ when f maps from the space E with the basis $\mathcal{U}$ to the space F with the basis $\mathcal{V}$. So, we prefer to use the notation $M_{\mathcal{U},\mathcal{V}}(f)$.

Be aware that other authors such as Meyer [125] use the notation $[f]_{\mathcal{U},\mathcal{V}}$, and others such as Dummit and Foote [54] use the notation $M_{\mathcal{U}}^{\mathcal{V}}(f)$, instead of $M_{\mathcal{U},\mathcal{V}}(f)$. This gets worse! You may find the notation $M_{\mathcal{V}}^{\mathcal{U}}(f)$ (as in Lang [109]), or ${}_{\mathcal{U}}[f]_{\mathcal{V}}$, or other strange notations.

Definition 4.2 shows that the function which associates to a linear map $f\colon E \to F$ the matrix $M(f)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ has the property that matrix multiplication corresponds to composition of linear maps. This allows us to transfer properties of linear maps to matrices. Here is an illustration of this technique:

Proposition 4.1. *(1) Given any matrices $A \in \mathrm{M}_{m,n}(K)$, $B \in \mathrm{M}_{n,p}(K)$, and $C \in \mathrm{M}_{p,q}(K)$, we have*

$$(AB)C = A(BC);$$

that is, matrix multiplication is associative.

(2) Given any matrices $A, B \in \mathrm{M}_{m,n}(K)$, and $C, D \in \mathrm{M}_{n,p}(K)$, for all $\lambda \in K$, we have

$$\begin{aligned}(A + B)C &= AC + BC\\ A(C + D) &= AC + AD\\ (\lambda A)C &= \lambda(AC)\\ A(\lambda C) &= \lambda(AC),\end{aligned}$$

so that matrix multiplication $\cdot\colon \mathrm{M}_{m,n}(K) \times \mathrm{M}_{n,p}(K) \to \mathrm{M}_{m,p}(K)$ is bilinear.

Proof. (1) Every $m \times n$ matrix $A = (a_{i\,j})$ defines the function $f_A\colon K^n \to K^m$ given by

$$f_A(x) = Ax,$$

for all $x \in K^n$. It is immediately verified that f_A is linear and that the matrix $M(f_A)$ representing f_A over the canonical bases in K^n and K^m is equal to A. Then Formula (4) proves that

$$M(f_A \circ f_B) = M(f_A)M(f_B) = AB,$$

so we get

$$M((f_A \circ f_B) \circ f_C) = M(f_A \circ f_B)M(f_C) = (AB)C$$

and

$$M(f_A \circ (f_B \circ f_C)) = M(f_A)M(f_B \circ f_C) = A(BC),$$

and since composition of functions is associative, we have $(f_A \circ f_B) \circ f_C = f_A \circ (f_B \circ f_C)$, which implies that

$$(AB)C = A(BC).$$

(2) It is immediately verified that if $f_1, f_2 \in \mathrm{Hom}_K(E, F)$, $A, B \in \mathrm{M}_{m,n}(K)$, $(u_1, \ldots, u_n)$ is any basis of E, and $(v_1, \ldots, v_m)$ is any basis of F, then

$$\begin{aligned} M(f_1 + f_2) &= M(f_1) + M(f_2) \\ f_{A+B} &= f_A + f_B. \end{aligned}$$

Then we have

$$\begin{aligned} (A + B)C &= M(f_{A+B})M(f_C) \\ &= M(f_{A+B} \circ f_C) \\ &= M((f_A + f_B) \circ f_C)) \\ &= M((f_A \circ f_C) + (f_B \circ f_C)) \\ &= M(f_A \circ f_C) + M(f_B \circ f_C) \\ &= M(f_A)M(f_C) + M(f_B)M(f_C) \\ &= AC + BC. \end{aligned}$$

The equation $A(C + D) = AC + AD$ is proven in a similar fashion, and the last two equations are easily verified. We could also have verified all the identities by making matrix computations. □

Note that Proposition 4.1 implies that the vector space $\mathrm{M}_n(K)$ of square matrices is a (noncommutative) ring with unit I_n. (It even shows that $\mathrm{M}_n(K)$ is an associative *algebra*.)

The following proposition states the main properties of the mapping $f \mapsto M(f)$ between $\mathrm{Hom}(E, F)$ and $\mathrm{M}_{m,n}$. In short, it is an isomorphism of vector spaces.

Proposition 4.2. *Given three vector spaces E, F, G, with respective bases $(u_1, \ldots, u_p)$, $(v_1, \ldots, v_n)$, and $(w_1, \ldots, w_m)$, the mapping $M\colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ that associates the matrix $M(g)$ to a linear map $g\colon E \to F$ satisfies the following properties for all $x \in E$, all $g, h\colon E \to F$, and all $f\colon F \to G$:*

$$\begin{aligned} M(g(x)) &= M(g)M(x) \\ M(g + h) &= M(g) + M(h) \\ M(\lambda g) &= \lambda M(g) \\ M(f \circ g) &= M(f)M(g), \end{aligned}$$

where $M(x)$ is the column vector associated with the vector x and $M(g(x))$ is the column vector associated with $g(x)$, as explained in Definition 4.1.

Thus, $M\colon \mathrm{Hom}(E,F) \to \mathrm{M}_{n,p}$ is an isomorphism of vector spaces, and when $p = n$ and the basis $(v_1, \ldots, v_n)$ is identical to the basis $(u_1, \ldots, u_p)$, $M\colon \mathrm{Hom}(E,E) \to \mathrm{M}_n$ is an isomorphism of rings.

Proof. That $M(g(x)) = M(g)M(x)$ was shown by Definition 4.2 or equivalently by Formula (1). The identities $M(g+h) = M(g) + M(h)$ and $M(\lambda g) = \lambda M(g)$ are straightforward, and $M(f \circ g) = M(f)M(g)$ follows from Identity (4) and the definition of matrix multiplication. The mapping $M\colon \mathrm{Hom}(E,F) \to \mathrm{M}_{n,p}$ is clearly injective, and since every matrix defines a linear map (see Proposition 4.1), it is also surjective, and thus bijective. In view of the above identities, it is an isomorphism (and similarly for $M\colon \mathrm{Hom}(E,E) \to \mathrm{M}_n$, where Proposition 4.1 is used to show that M_n is a ring). $\square$

In view of Proposition 4.2, it seems preferable to represent vectors from a vector space of finite dimension as column vectors rather than row vectors. *Thus, from now on, we will denote vectors of $\mathbb{R}^n$ (or more generally, of K^n) as column vectors.*

We explained in Section 3.9 that if the space E is finite-dimensional and has a finite basis $(u_1, \ldots, u_n)$, then a linear form $f^*\colon E \to K$ is represented by the *row vector* of coefficients

$$\begin{pmatrix} f^*(u_1) & \cdots & f^*(u_n) \end{pmatrix}, \tag{1}$$

over the bases $(u_1, \ldots, u_n)$ and 1 (in K), and that over the dual basis $(u_1^*, \ldots, u_n^*)$ of E^*, the linear form f^* is represented by the same coefficients, but as the *column vector*

$$\begin{pmatrix} f^*(u_1) \\ \vdots \\ f^*(u_n) \end{pmatrix}, \tag{2}$$

which is the transpose of the row vector in (1).

This is a special case of a more general phenomenon. A linear map $f\colon E \to F$ induces a map $f^\top\colon F^* \to E^*$ called the *transpose* of f (note that $f^\top$ maps F^* to E^*, *not* E^* to F^*), and if $(u_1 \ldots, u_n)$ is a basis of E, $(v_1 \ldots, v_m)$ is a basis of F, and if f is represented by the $m \times n$ matrix A over these bases, then over the dual bases $(v_1^*, \ldots, v_m^*)$ and $(u_1^*, \ldots, u_n^*)$, the linear map $f^\top$ is represented by $A^\top$, the transpose of the matrix A.

This is because over the basis $(v_1, \ldots, v_m)$, a linear form $\varphi \in F^*$ is represented by the row vector

$$\lambda = \begin{pmatrix} \varphi(v_1) & \cdots & \varphi(v_m) \end{pmatrix},$$

and we define $f^\top(\varphi)$ as the linear form represented by the row vector

$$\lambda A$$

over the basis $(u_1, \ldots, u_n)$. Since φ is represented by the column vector $\lambda^\top$ over the dual basis $(v_1^*, \ldots, v_m^*)$, we see that $f^\top(\varphi)$ is represented by the column vector

$$(\lambda A)^\top = A^\top \lambda^\top$$

over the dual basis $(u_1^*, \ldots, u_n^*)$. The matrix defining $f^\top$ over the dual bases $(v_1^*, \ldots, v_m^*)$ and $(u_1^*, \ldots, u_n^*)$ is indeed $A^\top$.

Conceptually, we will show later (see Section 30.1) that the linear map $f^\top \colon F^* \to E^*$ is defined by

$$f^\top(\varphi) = \varphi \circ f,$$

for all $\varphi \in F^*$ (remember that $\varphi \colon F \to K$, so composing $f \colon E \to F$ and $\varphi \colon F \to K$ yields a linear form $\varphi \circ f \colon E \to K$).

4.3 Change of Basis Matrix

It is important to observe that the isomorphism $M \colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ given by Proposition 4.2 depends on the choice of the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$, and similarly for the isomorphism $M \colon \mathrm{Hom}(E, E) \to \mathrm{M}_n$, which depends on the choice of the basis $(u_1, \ldots, u_n)$. Thus, it would be useful to know how a change of basis affects the representation of a linear map $f \colon E \to F$ as a matrix. The following simple proposition is needed.

Proposition 4.3. *Let E be a vector space, and let $(u_1, \ldots, u_n)$ be a basis of E. For every family $(v_1, \ldots, v_n)$, let $P = (a_{i\,j})$ be the matrix defined such that $v_j = \sum_{i=1}^n a_{i\,j} u_i$. The matrix P is invertible iff $(v_1, \ldots, v_n)$ is a basis of E.*

Proof. Note that we have $P = M(f)$, the matrix (with respect to the basis $(u_1, \ldots, u_n)$) associated with the unique linear map $f \colon E \to E$ such that $f(u_i) = v_i$. By Proposition 3.18, f is bijective iff $(v_1, \ldots, v_n)$ is a basis of E. Furthermore, it is obvious that the identity matrix I_n is the matrix associated with the identity $\mathrm{id} \colon E \to E$ w.r.t. any basis. If f is an isomorphism, then $f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}$, and by Proposition 4.2, we get $M(f)M(f^{-1}) = M(f^{-1})M(f) = I_n$, showing that P is invertible and that $M(f^{-1}) = P^{-1}$. $\square$

An important corollary of Proposition 4.3 yields the following criterion for a square matrix to be invertible. This criterion was already proven in Proposition 3.14 but Proposition 4.3 yields a shorter proof.

Proposition 4.4. *A square matrix $A \in \mathrm{M}_n(K)$ is invertible iff its columns $(A^1, \ldots, A^n)$ are linearly independent.*

Proof. First assume that A is invertible. If $\lambda_1 A^1 + \cdots + \lambda_n A^n = 0$ for some $\lambda_1, \ldots, \lambda_n \in K$, then

$$A\lambda = \lambda_1 A^1 + \cdots + \lambda_n A^n = 0,$$

where λ is the column vector $\lambda = (\lambda_1, \ldots, \lambda_n)$. Since A has an inverse A^{-1}, by multiplying both sides of the equation $A\lambda = 0$ by A^{-1} we obtain

$$A^{-1}A\lambda = I_n\lambda = \lambda = A^{-1}0 = 0,$$

which shows that the columns $(A^1, \ldots, A^n)$ are linearly independent.

Conversely, assume that the columns $(A^1, \ldots, A^n)$ are linearly independent. Since the vector space $E = K^n$ has dimension n, the vectors $(v_1, \ldots, v_n) = (A^1, \ldots, A^n)$ form a basis of K^n. By definition, the matrix A is defined by expressing each vector $v_j = A^j$ as the linear combination $A^j = \sum_{i=1}^n a_{ij} e_i$, where $(e_1, \ldots, e_n)$ is the canonical basis of K^n, and since $(v_1, \ldots, v_n)$ is a basis, by Proposition 4.3, the matrix A is invertible. $\square$

Proposition 4.3 suggests the following definition.

Definition 4.3. Given a vector space E of dimension n, for any two bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ of E, let $P = (a_{i\,j})$ be the invertible matrix defined such that

$$v_j = \sum_{i=1}^n a_{i\,j} u_i,$$

which is also the matrix of the identity $\mathrm{id}\colon E \to E$ with respect to the bases $(v_1, \ldots, v_n)$ and $(u_1, \ldots, u_n)$, *in that order*. Indeed, we express each $\mathrm{id}(v_j) = v_j$ over the basis $(u_1, \ldots, u_n)$. The coefficients $a_{1j}, a_{2j}, \ldots, a_{nj}$ of v_j over the basis $(u_1, \ldots, u_n)$ form the jth column of the matrix P shown below:

$$\begin{array}{c} \\ u_1 \\ u_2 \\ \vdots \\ u_n \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & \ldots & v_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \end{array}.$$

The matrix P is called the *change of basis matrix from* $(u_1, \ldots, u_n)$ *to* $(v_1, \ldots, v_n)$.

Clearly, the change of basis matrix from $(v_1, \ldots, v_n)$ to $(u_1, \ldots, u_n)$ is P^{-1}. Since $P = (a_{i\,j})$ is the matrix of the identity $\mathrm{id}\colon E \to E$ with respect to the bases $(v_1, \ldots, v_n)$ and $(u_1, \ldots, u_n)$, given any vector $x \in E$, if $x = x_1 u_1 + \cdots + x_n u_n$ over the basis $(u_1, \ldots, u_n)$ and $x = x'_1 v_1 + \cdots + x'_n v_n$ over the basis $(v_1, \ldots, v_n)$, from Proposition 4.2, we have

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{1\,1} & \ldots & a_{1\,n} \\ \vdots & \ddots & \vdots \\ a_{n\,1} & \ldots & a_{n\,n} \end{pmatrix} \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix},$$

showing that the *old* coordinates (x_i) of x (over $(u_1, \ldots, u_n)$) are expressed in terms of the *new* coordinates (x'_i) of x (over $(v_1, \ldots, v_n)$). This fact may seem wrong, but it is correct as we can reassure ourselves by doing the following computation. Suppose that $n = 2$, so that

$$\begin{aligned} v_1 &= a_{11}u_1 + a_{21}u_2 \\ v_2 &= a_{12}u_1 + a_{22}u_2, \end{aligned}$$

and our matrix is

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

The same vector x is written as

$$x = x_1u_1 + x_2u_2 = x'_1v_1 + x'_2v_2,$$

so by substituting the expressions for v_1 and v_2 as linear combinations of u_1 and u_2, we obtain

$$\begin{aligned} x_1u_1 + x_2u_2 &= x'_1v_1 + x'_2v_2 \\ &= x'_1(a_{11}u_1 + a_{21}u_2) + x'_2(a_{12}u_1 + a_{22}u_2) \\ &= (a_{11}x'_1 + a_{12}x'_2)u_1 + (a_{21}x'_1 + a_{22}x'_2)u_2, \end{aligned}$$

and since u_1 and u_2 are linearly independent, we must have

$$\begin{aligned} x_1 &= a_{11}x'_1 + a_{12}x'_2 \\ x_2 &= a_{21}x'_1 + a_{22}x'_2, \end{aligned}$$

namely

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix},$$

as claimed.

If the vectors $u_1, \ldots, u_n$ and the vectors $v_1, \ldots, v_n$ are vectors in K^n, then we can form the $n \times n$ matrix $U = (u_1 \ \cdots \ u_n)$ whose columns are $u_1, \ldots, u_n$ and the $n \times n$ matrix $V = (v_1 \ \cdots \ v_n)$ whose columns are $v_1, \ldots, v_n$. Then we can express the change of basis P from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$ in terms of U and V. Indeeed, the equation

$$v_j = \sum_{i=1}^{n} a_{ij}u_i$$

can be expressed in matrix form as

$$v_j = UA^j,$$

where

$$A^j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{n1} \end{pmatrix}$$

is the jth column of P, so we get

$$V = UP,$$

which yields

$$P = U^{-1}V.$$

Now we face the painful task of assigning a "good" notation incorporating the bases $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ into the notation for the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$. Because the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ is the matrix of the identity map id_E *with respect to the bases* $\mathcal{V}$ *and* $\mathcal{U}$ *in that order*, we could denote it by $M_{\mathcal{V},\mathcal{U}}(\mathrm{id})$ (Meyer [125] uses the notation $[I]_{\mathcal{V},\mathcal{U}}$). We prefer to use an abbreviation for $M_{\mathcal{V},\mathcal{U}}(\mathrm{id})$.

Definition 4.4. The *change of basis matrix from* $\mathcal{U}$ *to* $\mathcal{V}$ is denoted

$$P_{\mathcal{V},\mathcal{U}}.$$

Note that

$$P_{\mathcal{U},\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}.$$

Then, if we write $x_{\mathcal{U}} = (x_1, \ldots, x_n)$ for the *old* coordinates of x with respect to the basis $\mathcal{U}$ and $x_{\mathcal{V}} = (x'_1, \ldots, x'_n)$ for the *new* coordinates of x with respect to the basis $\mathcal{V}$, we have

$$x_{\mathcal{U}} = P_{\mathcal{V},\mathcal{U}}\, x_{\mathcal{V}}, \quad x_{\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}\, x_{\mathcal{U}}.$$

The above may look backward, but remember that the matrix $M_{\mathcal{U},\mathcal{V}}(f)$ takes input expressed over the basis $\mathcal{U}$ to output expressed over the basis $\mathcal{V}$. Consequently, $P_{\mathcal{V},\mathcal{U}}$ takes input expressed over the basis $\mathcal{V}$ to output expressed over the basis $\mathcal{U}$, and $x_{\mathcal{U}} = P_{\mathcal{V},\mathcal{U}}\, x_{\mathcal{V}}$ matches this point of view!

Beware that some authors (such as Artin [7]) define the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ as $P_{\mathcal{U},\mathcal{V}} = P_{\mathcal{V},\mathcal{U}}^{-1}$. Under this point of view, the old basis $\mathcal{U}$ is expressed in terms of the new basis $\mathcal{V}$. We find this a bit unnatural. Also, in practice, it seems that the new basis is often expressed in terms of the old basis, rather than the other way around.

Since the matrix $P = P_{\mathcal{V},\mathcal{U}}$ expresses the *new* basis $(v_1, \ldots, v_n)$ in terms of the *old* basis $(u_1, \ldots, u_n)$, we observe that the coordinates (x_i) of a vector x vary in the *opposite direction* of the change of basis. For this reason, vectors are sometimes said to be *contravariant*. However, this expression does not make sense! Indeed, a vector in an intrinsic quantity that does not depend on a specific basis. What makes sense is that the *coordinates* of a vector vary in a contravariant fashion.

Let us consider some concrete examples of change of bases.

Example 4.1. Let $E = F = \mathbb{R}^2$, with $u_1 = (1, 0)$, $u_2 = (0, 1)$, $v_1 = (1, 1)$ and $v_2 = (-1, 1)$. The change of basis matrix P from the basis $\mathcal{U} = (u_1, u_2)$ to the basis $\mathcal{V} = (v_1, v_2)$ is

$$P = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

and its inverse is

$$P^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

The old coordinates (x_1, x_2) with respect to (u_1, u_2) are expressed in terms of the new coordinates (x_1', x_2') with respect to (v_1, v_2) by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \end{pmatrix},$$

and the new coordinates (x_1', x_2') with respect to (v_1, v_2) are expressed in terms of the old coordinates (x_1, x_2) with respect to (u_1, u_2) by

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Example 4.2. Let $E = F = \mathbb{R}[X]_3$ be the set of polynomials of degree at most 3, and consider the bases $\mathcal{U} = (1, x, x^2, x^3)$ and $\mathcal{V} = (B_0^3(x), B_1^3(x), B_2^3(x), B_3^3(x))$, where $B_0^3(x), B_1^3(x), B_2^3(x), B_3^3(x)$ are the *Bernstein polynomials* of degree 3, given by

$$B_0^3(x) = (1-x)^3 \qquad B_1^3(x) = 3(1-x)^2 x \qquad B_2^3(x) = 3(1-x)x^2 \qquad B_3^3(x) = x^3.$$

By expanding the Bernstein polynomials, we find that the change of basis matrix $P_{\mathcal{V},\mathcal{U}}$ is given by

$$P_{\mathcal{V},\mathcal{U}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{pmatrix}.$$

We also find that the inverse of $P_{\mathcal{V},\mathcal{U}}$ is

$$P_{\mathcal{V},\mathcal{U}}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1/3 & 0 & 0 \\ 1 & 2/3 & 1/3 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Therefore, the coordinates of the polynomial $2x^3 - x + 1$ over the basis $\mathcal{V}$ are

$$\begin{pmatrix} 1 \\ 2/3 \\ 1/3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1/3 & 0 & 0 \\ 1 & 2/3 & 1/3 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix},$$

and so

$$2x^3 - x + 1 = B_0^3(x) + \frac{2}{3}B_1^3(x) + \frac{1}{3}B_2^3(x) + 2B_3^3(x).$$

4.4 The Effect of a Change of Bases on Matrices

The effect of a change of bases on the representation of a linear map is described in the following proposition.

Proposition 4.5. *Let E and F be vector spaces, let $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{U}' = (u'_1, \ldots, u'_n)$ be two bases of E, and let $\mathcal{V} = (v_1, \ldots, v_m)$ and $\mathcal{V}' = (v'_1, \ldots, v'_m)$ be two bases of F. Let $P = P_{\mathcal{U}',\mathcal{U}}$ be the change of basis matrix from $\mathcal{U}$ to $\mathcal{U}'$, and let $Q = P_{\mathcal{V}',\mathcal{V}}$ be the change of basis matrix from $\mathcal{V}$ to $\mathcal{V}'$. For any linear map $f\colon E \to F$, let $M(f) = M_{\mathcal{U},\mathcal{V}}(f)$ be the matrix associated to f w.r.t. the bases $\mathcal{U}$ and $\mathcal{V}$, and let $M'(f) = M_{\mathcal{U}',\mathcal{V}'}(f)$ be the matrix associated to f w.r.t. the bases $\mathcal{U}'$ and $\mathcal{V}'$. We have*

$$M'(f) = Q^{-1}M(f)P,$$

or more explicitly

$$M_{\mathcal{U}',\mathcal{V}'}(f) = P^{-1}_{\mathcal{V}',\mathcal{V}} M_{\mathcal{U},\mathcal{V}}(f) P_{\mathcal{U}',\mathcal{U}} = P_{\mathcal{V},\mathcal{V}'} M_{\mathcal{U},\mathcal{V}}(f) P_{\mathcal{U}',\mathcal{U}}.$$

Proof. Since $f\colon E \to F$ can be written as $f = \mathrm{id}_F \circ f \circ \mathrm{id}_E$, since $P = P_{\mathcal{U}',\mathcal{U}}$ is the matrix of id_E w.r.t. the bases $(u'_1, \ldots, u'_n)$ and $(u_1, \ldots, u_n)$, and $Q^{-1} = P^{-1}_{\mathcal{V}',\mathcal{V}} = P_{\mathcal{V},\mathcal{V}'}$ is the matrix of id_F w.r.t. the bases $(v_1, \ldots, v_m)$ and $(v'_1, \ldots, v'_m)$ as illustrated by the following diagram

$$\begin{tikzcd}
\mathcal{U}, E \arrow[r, "f", "M_{\mathcal{U},\mathcal{V}}(f)"'] & \mathcal{V}, F \arrow[d, "P^{-1}_{\mathcal{V}',\mathcal{V}}"', "\mathrm{id}_F"] \\
\mathcal{U}', E \arrow[u, "P_{\mathcal{U}',\mathcal{U}}", "\mathrm{id}_E"'] \arrow[r, "M_{\mathcal{U}',\mathcal{V}'}(f)", "f"'] & \mathcal{V}', F,
\end{tikzcd}$$

by Proposition 4.2, we have $M'(f) = Q^{-1}M(f)P$. $\square$

As a corollary, we get the following result.

Corollary 4.6. *Let E be a vector space, and let $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{U}' = (u'_1, \ldots, u'_n)$ be two bases of E. Let $P = P_{\mathcal{U}',\mathcal{U}}$ be the change of basis matrix from $\mathcal{U}$ to $\mathcal{U}'$. For any linear map $f\colon E \to E$, let $M(f) = M_{\mathcal{U}}(f)$ be the matrix associated to f w.r.t. the basis $\mathcal{U}$, and let $M'(f) = M_{\mathcal{U}'}(f)$ be the matrix associated to f w.r.t. the basis $\mathcal{U}'$. We have*

$$M'(f) = P^{-1}M(f)P,$$

or more explicitly,

$$M_{\mathcal{U}'}(f) = P^{-1}_{\mathcal{U}',\mathcal{U}} M_{\mathcal{U}}(f) P_{\mathcal{U}',\mathcal{U}} = P_{\mathcal{U},\mathcal{U}'} M_{\mathcal{U}}(f) P_{\mathcal{U}',\mathcal{U}},$$

as illustrated by the following diagram

$$\begin{tikzcd}
\mathcal{U}, E \arrow[r, "f"', "M_{\mathcal{U}},(f)"] & \mathcal{U}, E \arrow[d, "P^{-1}_{\mathcal{U}',\mathcal{U}}"', "\mathrm{id}_E"] \\
\mathcal{U}', E \arrow[u, "P_{\mathcal{U}',\mathcal{U}}", "\mathrm{id}_E"'] \arrow[r, "M_{\mathcal{U}'}(f)", "f"'] & \mathcal{U}', E.
\end{tikzcd}$$

Example 4.3. Let $E = \mathbb{R}^2$, $\mathcal{U} = (e_1, e_2)$ where $e_1 = (1, 0)$ and $e_2 = (0, 1)$ are the canonical basis vectors, let $\mathcal{V} = (v_1, v_2) = (e_1, e_1 - e_2)$, and let

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

The change of basis matrix $P = P_{\mathcal{V},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{V}$ is

$$P = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix},$$

and we check that

$$P^{-1} = P.$$

Therefore, in the basis $\mathcal{V}$, the matrix representing the linear map f defined by A is

$$A' = P^{-1}AP = PAP = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} = D,$$

a diagonal matrix. In the basis $\mathcal{V}$, it is clear what the action of f is: it is a stretch by a factor of 2 in the v_1 direction and it is the identity in the v_2 direction. Observe that v_1 and v_2 are not orthogonal.

What happened is that we *diagonalized* the matrix A. The diagonal entries 2 and 1 are the *eigenvalues* of A (and f), and v_1 and v_2 are corresponding *eigenvectors*. We will come back to eigenvalues and eigenvectors later on.

The above example showed that the same linear map can be represented by different matrices. This suggests making the following definition:

Definition 4.5. Two $n \times n$ matrices A and B are said to be *similar* iff there is some invertible matrix P such that

$$B = P^{-1}AP.$$

It is easily checked that similarity is an equivalence relation. From our previous considerations, *two $n \times n$ matrices A and B are similar iff they represent the same linear map with respect to two different bases.* The following surprising fact can be shown: Every square matrix A is similar to its transpose $A^\top$. The proof requires advanced concepts (the Jordan form or similarity invariants).

If $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ are two bases of E, the change of basis matrix

$$P = P_{\mathcal{V},\mathcal{U}} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$ is the matrix whose jth column consists of the coordinates of v_j over the basis $(u_1, \ldots, u_n)$, which means that

$$v_j = \sum_{i=1}^n a_{ij} u_i.$$

It is natural to extend the matrix notation and to express the vector $\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ in E^n as the product of a matrix times the vector $\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ in E^n, namely as

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

but notice that the matrix involved is not P, but its *transpose* $P^\top$.

This observation has the following consequence: if $\mathcal{U} = (u_1, \ldots, u_n)$ and $\mathcal{V} = (v_1, \ldots, v_n)$ are two bases of E and if

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = A \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

that is,

$$v_i = \sum_{j=1}^n a_{ij} u_j,$$

for any vector $w \in E$, if

$$w = \sum_{i=1}^n x_i u_i = \sum_{k=1}^n y_k v_k = \sum_{k=1}^n y_k \left(\sum_{j=1}^n a_{kj} u_j \right) = \sum_{j=1}^n \left(\sum_{k=1}^n a_{kj} y_k \right) u_j,$$

so

$$x_i = \sum_{k=1}^{n} a_{kj} y_k,$$

which means (note the inevitable transposition) that

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = A^\top \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and so

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = (A^\top)^{-1} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

It is easy to see that $(A^\top)^{-1} = (A^{-1})^\top$. Also, if $\mathcal{U} = (u_1, \ldots, u_n)$, $\mathcal{V} = (v_1, \ldots, v_n)$, and $\mathcal{W} = (w_1, \ldots, w_n)$ are three bases of E, and if the change of basis matrix from $\mathcal{U}$ to $\mathcal{V}$ is $P = P_{\mathcal{V},\mathcal{U}}$ and the change of basis matrix from $\mathcal{V}$ to $\mathcal{W}$ is $Q = P_{\mathcal{W},\mathcal{V}}$, then

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = P^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \quad \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = Q^\top \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

so

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = Q^\top P^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = (PQ)^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

which means that the change of basis matrix $P_{\mathcal{W},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{W}$ is PQ. This proves that

$$P_{\mathcal{W},\mathcal{U}} = P_{\mathcal{V},\mathcal{U}} P_{\mathcal{W},\mathcal{V}}.$$

Remark: In order to avoid the transposition involved in writing

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = P^\top \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

as a more convenient notation we may write

$$\begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} P.$$

Here we are defining the product

$$\begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} p_{1j} \\ \vdots \\ p_{nj} \end{pmatrix} \tag{$*$}$$

of a row of vectors $(u_1 \ \cdots \ u_n)$ by the jth column of P as the linear combination

$$\sum_{i=1}^{n} p_{ij} u_i.$$

Such a definition is needed since scalar multiplication of a vector by a scalar is only defined if the scalar is on the left of the vector, but in the matrix expression $(*)$ above, the vectors are on the left of the scalars!

Even though matrices are indispensable since they are *the* major tool in applications of linear algebra, one should not lose track of the fact that

linear maps are more fundamental because they are intrinsic objects that do not depend on the choice of bases. Consequently, we advise the reader to try to think in terms of linear maps rather than reduce everything to matrices.

In our experience, this is particularly effective when it comes to proving results about linear maps and matrices, where proofs involving linear maps are often more "conceptual." These proofs are usually more general because they do not depend on the fact that the dimension is finite. Also, instead of thinking of a matrix decomposition as a purely algebraic operation, it is often illuminating to view it as a *geometric decomposition*. This is the case of the SVD, which in geometric terms says that every linear map can be factored as a rotation, followed by a rescaling along orthogonal axes and then another rotation.

After all,

a matrix is a representation of a linear map,

and most decompositions of a matrix reflect the fact that with a *suitable choice of a basis (or bases)*, the linear map is a represented by a matrix having a special shape. The problem is then to find such bases.

Still, for the beginner, matrices have a certain irresistible appeal, and we confess that it takes a certain amount of practice to reach the point where it becomes more natural to deal with linear maps. We still recommend it! For example, try to translate a result stated in terms of matrices into a result stated in terms of linear maps. Whenever we tried this exercise, we learned something.

Also, always try to keep in mind that

linear maps are geometric in nature; they act on space.

4.5 Summary

The main concepts and results of this chapter are listed below:

- The representation of linear maps by *matrices*.
- The *matrix representation mapping* $M\colon \mathrm{Hom}(E, F) \to \mathrm{M}_{n,p}$ and the representation isomorphism (Proposition 4.2).
- *Change of basis matrix* and Proposition 4.5.

4.6 Problems

Problem 4.1. Prove that the column vectors of the matrix A_1 given by

$$A_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 3 & 1 \end{pmatrix}$$

are linearly independent.

Prove that the coordinates of the column vectors of the matrix B_1 over the basis consisting of the column vectors of A_1 given by

$$B_1 = \begin{pmatrix} 3 & 5 & 1 \\ 1 & 2 & 1 \\ 4 & 3 & -6 \end{pmatrix}$$

are the columns of the matrix P_1 given by

$$P_1 = \begin{pmatrix} -27 & -61 & -41 \\ 9 & 18 & 9 \\ 4 & 10 & 8 \end{pmatrix}.$$

Give a nontrivial linear dependence of the columns of P_1. Check that $B_1 = A_1 P_1$. Is the matrix B_1 invertible?

Problem 4.2. Prove that the column vectors of the matrix A_2 given by

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 3 \end{pmatrix}$$

are linearly independent.

Prove that the column vectors of the matrix B_2 given by

$$B_2 = \begin{pmatrix} 1 & -2 & 2 & -2 \\ 0 & -3 & 2 & -3 \\ 3 & -5 & 5 & -4 \\ 3 & -4 & 4 & -4 \end{pmatrix}$$

are linearly independent.

Prove that the coordinates of the column vectors of the matrix B_2 over the basis consisting of the column vectors of A_2 are the columns of the matrix P_2 given by

$$P_2 = \begin{pmatrix} 2 & 0 & 1 & -1 \\ -3 & 1 & -2 & 1 \\ 1 & -2 & 2 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix}.$$

Check that $A_2P_2 = B_2$. Prove that

$$P_2^{-1} = \begin{pmatrix} -1 & -1 & -1 & 1 \\ 2 & 1 & 1 & -2 \\ 2 & 1 & 2 & -3 \\ -1 & -1 & 0 & -1 \end{pmatrix}.$$

What are the coordinates over the basis consisting of the column vectors of B_2 of the vector whose coordinates over the basis consisting of the column vectors of A_2 are $(2, -3, 0, 0)$?

Problem 4.3. Consider the polynomials

$$\begin{array}{llll} B_0^2(t) = (1-t)^2 & B_1^2(t) = 2(1-t)t & B_2^2(t) = t^2 & \\ B_0^3(t) = (1-t)^3 & B_1^3(t) = 3(1-t)^2t & B_2^3(t) = 3(1-t)t^2 & B_3^3(t) = t^3, \end{array}$$

known as the *Bernstein polynomials* of degree 2 and 3.

(1) Show that the Bernstein polynomials $B_0^2(t), B_1^2(t), B_2^2(t)$ are expressed as linear combinations of the basis $(1, t, t^2)$ of the vector space of polynomials of degree at most 2 as follows:

$$\begin{pmatrix} B_0^2(t) \\ B_1^2(t) \\ B_2^2(t) \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix}.$$

Prove that

$$B_0^2(t) + B_1^2(t) + B_2^2(t) = 1.$$

(2) Show that the Bernstein polynomials $B_0^3(t), B_1^3(t), B_2^3(t), B_3^3(t)$ are expressed as linear combinations of the basis $(1, t, t^2, t^3)$ of the vector space of polynomials of degree at most 3 as follows:

$$\begin{pmatrix} B_0^3(t) \\ B_1^3(t) \\ B_2^3(t) \\ B_3^3(t) \end{pmatrix} = \begin{pmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ t \\ t^2 \\ t^3 \end{pmatrix}.$$

Prove that

$$B_0^3(t) + B_1^3(t) + B_2^3(t) + B_3^3(t) = 1.$$

(3) Prove that the Bernstein polynomials of degree 2 are linearly independent, and that the Bernstein polynomials of degree 3 are linearly independent.

Problem 4.4. Recall that the *binomial coefficient* $\binom{m}{k}$ is given by

$$\binom{m}{k} = \frac{m!}{k!(m-k)!},$$

with $0 \le k \le m$.

For any $m \ge 1$, we have the $m+1$ *Bernstein polynomials* of degree m given by

$$B_k^m(t) = \binom{m}{k}(1-t)^{m-k}t^k, \quad 0 \le k \le m.$$

(1) Prove that

$$B_k^m(t) = \sum_{j=k}^{m} (-1)^{j-k}\binom{m}{j}\binom{j}{k}t^j. \tag{$*$}$$

Use the above to prove that $B_0^m(t), \ldots, B_m^m(t)$ are linearly independent.

(2) Prove that

$$B_0^m(t) + \cdots + B_m^m(t) = 1.$$

(3) What can you say about the symmetries of the $(m+1) \times (m+1)$ matrix expressing $B_0^m, \ldots, B_m^m$ in terms of the basis $1, t, \ldots, t^m$?

Prove your claim (beware that in equation $(*)$ the coefficient of t^j in B_k^m is the entry on the $(k+1)$th row of the $(j+1)$th column, since $0 \le k, j \le m$. Make appropriate modifications to the indices).

What can you say about the sum of the entries on each row of the above matrix? What about the sum of the entries on each column?

(4) The purpose of this question is to express the t^i in terms of the Bernstein polynomials $B_0^m(t), \ldots, B_m^m(t)$, with $0 \le i \le m$.

First, prove that

$$t^i = \sum_{j=0}^{m-i} t^i B_j^{m-i}(t), \quad 0 \le i \le m.$$

Then prove that

$$\binom{m}{i}\binom{m-i}{j} = \binom{m}{i+j}\binom{i+j}{i}.$$

Use the above facts to prove that

$$t^i = \sum_{j=0}^{m-i} \frac{\binom{i+j}{i}}{\binom{m}{i}} B_{i+j}^m(t).$$

Conclude that the Bernstein polynomials $B_0^m(t), \ldots, B_m^m(t)$ form a basis of the vector space of polynomials of degree $\leq m$.

Compute the matrix expressing $1, t, t^2$ in terms of $B_0^2(t), B_1^2(t), B_2^2(t)$, and the matrix expressing $1, t, t^2, t^3$ in terms of $B_0^3(t), B_1^3(t), B_2^3(t), B_3^3(t)$.

You should find

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1/3 & 2/3 & 1 \\ 0 & 0 & 1/3 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(5) A *polynomial curve $C(t)$ of degree m* in the plane is the set of points $C(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ given by two polynomials of degree $\leq m$,

$$\begin{aligned} x(t) &= \alpha_0 t^{m_1} + \alpha_1 t^{m_1 - 1} + \cdots + \alpha_{m_1} \\ y(t) &= \beta_0 t^{m_2} + \beta_1 t^{m_2 - 1} + \cdots + \beta_{m_2}, \end{aligned}$$

with $1 \leq m_1, m_2 \leq m$ and $\alpha_0, \beta_0 \neq 0$.

Prove that there exist $m+1$ points $b_0, \ldots, b_m \in \mathbb{R}^2$ so that

$$C(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = B_0^m(t) b_0 + B_1^m(t) b_1 + \cdots + B_m^m(t) b_m$$

for all $t \in \mathbb{R}$, with $C(0) = b_0$ and $C(1) = b_m$. Are the points $b_1, \ldots, b_{m-1}$ generally on the curve?

We say that the curve C is a *Bézier curve* and $(b_0, \ldots, b_m)$ is the list of *control points* of the curve (control points need not be distinct).

Remark: Because $B_0^m(t) + \cdots + B_m^m(t) = 1$ and $B_i^m(t) \geq 0$ when $t \in [0, 1]$, the curve segment $C[0, 1]$ corresponding to $t \in [0, 1]$ belongs to the convex hull of the control points. This is an important property of Bézier curves which is used in geometric modeling to find the intersection of curve segments. Bézier curves play an important role in computer graphics and geometric modeling, but also in robotics because they can be used to model the trajectories of moving objects.

Problem 4.5. Consider the $n \times n$ matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & 0 & \cdots & 0 & -a_{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & -a_2 \\ 0 & 0 & 0 & \cdots & 1 & -a_1 \end{pmatrix},$$

with $a_n \neq 0$.

(1) Find a matrix P such that

$$A^\top = P^{-1}AP.$$

What happens when $a_n = 0$?

Hint. First, try $n = 3, 4, 5$. Such a matrix must have zeros above the "antidiagonal," and identical entries p_{ij} for all $i, j \geq 0$ such that $i + j = n + k$, where $k = 1, \ldots, n$.

(2) Prove that if $a_n = 1$ and if $a_1, \ldots, a_{n-1}$ are integers, then P can be chosen so that the entries in P^{-1} are also integers.

Problem 4.6. For any matrix $A \in \mathrm{M}_n(\mathbb{C})$, let R_A and L_A be the maps from $\mathrm{M}_n(\mathbb{C})$ to itself defined so that

$$L_A(B) = AB, \quad R_A(B) = BA, \quad \text{for all } B \in \mathrm{M}_n(\mathbb{C}).$$

(1) Check that L_A and R_A are linear, and that L_A and R_B commute for all A, B.

Let $\mathrm{ad_A} \colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the linear map given by

$$\mathrm{ad}_A(B) = L_A(B) - R_A(B) = AB - BA = [A, B], \quad \text{for all } B \in \mathrm{M}_n(\mathbb{C}).$$

Note that $[A, B]$ is the Lie bracket.

(2) Prove that if A is invertible, then L_A and R_A are invertible; in fact, $(L_A)^{-1} = L_{A^{-1}}$ and $(R_A)^{-1} = R_{A^{-1}}$. Prove that if $A = PBP^{-1}$ for some invertible matrix P, then

$$L_A = L_P \circ L_B \circ L_P^{-1}, \quad R_A = R_P^{-1} \circ R_B \circ R_P.$$

(3) Recall that the n^2 matrices E_{ij} defined such that all entries in E_{ij} are zero except the (i, j)th entry, which is equal to 1, form a basis of the vector space $\mathrm{M}_n(\mathbb{C})$. Consider the partial ordering of the E_{ij} defined such that for $i = 1, \ldots, n$, if $n \geq j > k \geq 1$, then then E_{ij} precedes E_{ik}, and for $j = 1, \ldots, n$, if $1 \leq i < h \leq n$, then E_{ij} precedes E_{hj}.

Draw the Hasse diagram of the partial order defined above when $n = 3$.

There are total orderings extending this partial ordering. How would you find them algorithmically? Check that the following is such a total order:

$$(1, 3),\ (1, 2),\ (1, 1),\ (2, 3),\ (2, 2),\ (2, 1),\ (3, 3),\ (3, 2),\ (3, 1).$$

(4) Let the total order of the basis (E_{ij}) extending the partial ordering defined in (2) be given by

$$(i,j) < (h,k) \quad \text{iff} \quad \begin{cases} i = h \text{ and } j > k \\ \text{or } i < h. \end{cases}$$

Let R be the $n \times n$ permutation matrix given by

$$R = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Observe that $R^{-1} = R$. Prove that for any $n \geq 1$, the matrix of L_A is given by $A \otimes I_n$, and the matrix of R_A is given by $I_n \otimes RA^\top R$ (over the basis (E_{ij}) ordered as specified above), where $\otimes$ is the *Kronecker product* (also called *tensor product*) of matrices defined in Definition 5.4.

Hint. Figure out what are $R_B(E_{ij}) = E_{ij}B$ and $L_B(E_{ij}) = BE_{ij}$.

(5) Prove that if A is upper triangular, then the matrices representing L_A and R_A are also upper triangular.

Note that if instead of the ordering

$$E_{1n}, E_{1n-1}, \dots, E_{11}, E_{2n}, \dots, E_{21}, \dots, E_{nn}, \dots, E_{n1},$$

that I proposed you use the standard lexicographic ordering

$$E_{11}, E_{12}, \dots, E_{1n}, E_{21}, \dots, E_{2n}, \dots, E_{n1}, \dots, E_{nn},$$

then the matrix representing L_A is still $A \otimes I_n$, but the matrix representing R_A is $I_n \otimes A^\top$. In this case, if A is upper-triangular, then the matrix of R_A is *lower triangular*. This is the motivation for using the first basis (avoid upper becoming lower).

Chapter 5

Haar Bases, Haar Wavelets, Hadamard Matrices

In this chapter, we discuss two types of matrices that have applications in computer science and engineering:

(1) Haar matrices and the corresponding Haar wavelets, a fundamental tool in signal processing and computer graphics.

2) Hadamard matrices which have applications in error correcting codes, signal processing, and low rank approximation.

5.1 Introduction to Signal Compression Using Haar Wavelets

We begin by considering *Haar wavelets* in $\mathbb{R}^4$. Wavelets play an important role in audio and video signal processing, especially for *compressing* long signals into much smaller ones that still retain enough information so that when they are played, we can't see or hear any difference.

Consider the four vectors w_1, w_2, w_3, w_4 given by

$$w_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \qquad w_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \qquad w_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \qquad w_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

Note that these vectors are pairwise orthogonal, which means that their inner product is 0 (see Section 12.1, Example 12.1, and Section 12.2, Definition 12.2), so they are indeed linearly independent (see Proposition 12.4). Let $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$ be the *Haar basis*, and let

$\mathcal{U} = \{e_1, e_2, e_3, e_4\}$ be the canonical basis of $\mathbb{R}^4$. The change of basis matrix $W = P_{\mathcal{W},\mathcal{U}}$ from $\mathcal{U}$ to $\mathcal{W}$ is given by

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

and we easily find that the inverse of W is given by

$$W^{-1} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Observe that the second matrix in the above product is $W^\top$ and the first matrix in this product is $(W^\top W)^{-1}$. So the vector $v = (6, 4, 5, 1)$ over the basis $\mathcal{U}$ becomes $c = (c_1, c_2, c_3, c_4)$ over the Haar basis $\mathcal{W}$, with

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 6 \\ 4 \\ 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \\ 1 \\ 2 \end{pmatrix}.$$

Given a signal $v = (v_1, v_2, v_3, v_4)$, we first *transform* v into its coefficients $c = (c_1, c_2, c_3, c_4)$ over the Haar basis by computing $c = W^{-1}v$. Observe that

$$c_1 = \frac{v_1 + v_2 + v_3 + v_4}{4}$$

is the overall *average* value of the signal v. The coefficient c_1 corresponds to the background of the image (or of the sound). Then, c_2 gives the coarse details of v, whereas, c_3 gives the details in the first part of v, and c_4 gives the details in the second half of v.

Reconstruction of the signal consists in computing $v = Wc$. The trick for good *compression* is to throw away some of the coefficients of c (set them to zero), obtaining a *compressed signal* $\widehat{c}$, and still retain enough crucial information so that the reconstructed signal $\widehat{v} = W\widehat{c}$ looks almost as good as the original signal v. Thus, the steps are:

$$\text{input } v \longrightarrow \text{ coefficients } c = W^{-1}v \longrightarrow \text{ compressed } \widehat{c} \longrightarrow \text{ compressed } \widehat{v} = W\widehat{c}.$$

This kind of compression scheme makes modern video conferencing possible.

It turns out that there is a faster way to find $c = W^{-1}v$, without actually using W^{-1}. This has to do with the multiscale nature of Haar wavelets.

Given the original signal $v = (6, 4, 5, 1)$ shown in Figure 5.1, we compute averages and half differences obtaining Figure 5.2. We get the coefficients $c_3 = 1$ and $c_4 = 2$. Then

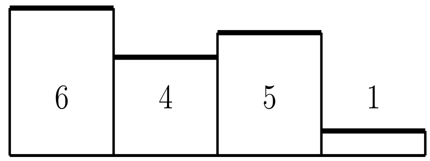


Figure 5.1: The original signal v.

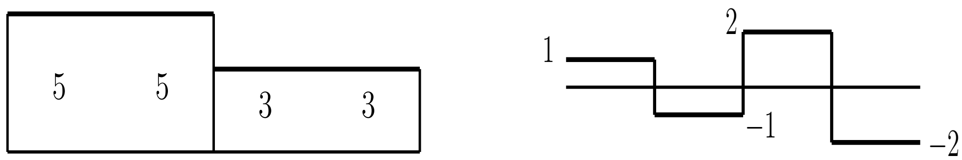


Figure 5.2: First averages and first half differences.

again we compute averages and half differences obtaining Figure 5.3. We get the coefficients $c_1 = 4$ and $c_2 = 1$. Note that the original signal v can be reconstructed from the two signals in Figure 5.2, and the signal on the left of Figure 5.2 can be reconstructed from the two signals in Figure 5.3. In particular, the data from Figure 5.2 gives us

$$\begin{aligned}
5 + 1 &= \frac{v_1 + v_2}{2} + \frac{v_1 - v_2}{2} = v_1 \\
5 - 1 &= \frac{v_1 + v_2}{2} - \frac{v_1 - v_2}{2} = v_2 \\
3 + 2 &= \frac{v_3 + v_4}{2} + \frac{v_3 - v_4}{2} = v_3 \\
3 - 2 &= \frac{v_3 + v_4}{2} - \frac{v_3 - v_4}{2} = v_4.
\end{aligned}$$

5.2 Haar Bases and Haar Matrices, Scaling Properties of Haar Wavelets

The method discussed in Section 5.1 can be generalized to signals of any length 2^n. The previous case corresponds to $n = 2$. Let us consider the case $n = 3$. The *Haar basis*

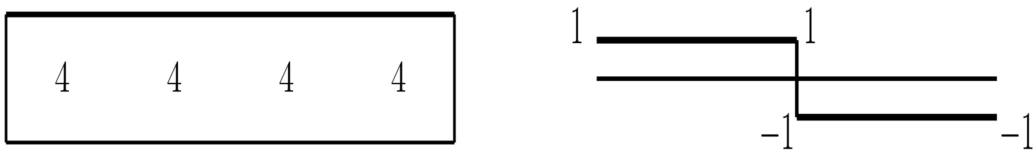


Figure 5.3: Second averages and second half differences.

$(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)$ is given by the matrix

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

The columns of this matrix are orthogonal, and it is easy to see that

$$W^{-1} = \mathrm{diag}(1/8, 1/8, 1/4, 1/4, 1/2, 1/2, 1/2, 1/2)W^{\top}.$$

A pattern is beginning to emerge. It looks like the second Haar basis vector w_2 is the "mother" of all the other basis vectors, except the first, whose purpose is to perform averaging. Indeed, in general, given

$$w_2 = \underbrace{(1, \ldots, 1, -1, \ldots, -1)}_{2^n},$$

the other Haar basis vectors are obtained by a "scaling and shifting process." Starting from w_2, the scaling process generates the vectors

$$w_3, w_5, w_9, \ldots, w_{2^j+1}, \ldots, w_{2^{n-1}+1},$$

such that $w_{2^{j+1}+1}$ is obtained from w_{2^j+1} by forming two consecutive blocks of 1 and -1 of half the size of the blocks in w_{2^j+1}, and setting all other entries to zero. Observe that w_{2^j+1} has 2^j blocks of 2^{n-j} elements. The shifting process consists in shifting the blocks of 1 and -1 in w_{2^j+1} to the right by inserting a block of $(k-1)2^{n-j}$ zeros from the left, with $0 \leq j \leq n-1$ and $1 \leq k \leq 2^j$. Note that our convention is to use j as the scaling index and k as the shifting index. Thus, we obtain the following formula for w_{2^j+k}:

$$w_{2^j+k}(i) = \begin{cases} 0 & 1 \leq i \leq (k-1)2^{n-j} \\ 1 & (k-1)2^{n-j} + 1 \leq i \leq (k-1)2^{n-j} + 2^{n-j-1} \\ -1 & (k-1)2^{n-j} + 2^{n-j-1} + 1 \leq i \leq k2^{n-j} \\ 0 & k2^{n-j} + 1 \leq i \leq 2^n, \end{cases}$$

with $0 \leq j \leq n-1$ and $1 \leq k \leq 2^j$. Of course

$$w_1 = \underbrace{(1, \ldots, 1)}_{2^n}.$$

The above formulae look a little better if we change our indexing slightly by letting k vary from 0 to $2^j - 1$, and using the index j instead of 2^j.

Definition 5.1. The vectors of the *Haar basis* of dimension 2^n are denoted by

$$w_1, h_0^0, h_0^1, h_1^1, h_0^2, h_1^2, h_2^2, h_3^2, \ldots, h_k^j, \ldots, h_{2^{n-1}-1}^{n-1},$$

where

$$h_k^j(i) = \begin{cases} 0 & 1 \leq i \leq k2^{n-j} \\ 1 & k2^{n-j} + 1 \leq i \leq k2^{n-j} + 2^{n-j-1} \\ -1 & k2^{n-j} + 2^{n-j-1} + 1 \leq i \leq (k+1)2^{n-j} \\ 0 & (k+1)2^{n-j} + 1 \leq i \leq 2^n, \end{cases}$$

with $0 \leq j \leq n-1$ and $0 \leq k \leq 2^j - 1$. The $2^n \times 2^n$ matrix whose columns are the vectors

$$w_1, h_0^0, h_0^1, h_1^1, h_0^2, h_1^2, h_2^2, h_3^2, \ldots, h_k^j, \ldots, h_{2^{n-1}-1}^{n-1},$$

(in that order), is called the *Haar matrix* of dimension 2^n, and is denoted by W_n.

It turns out that there is a way to understand these formulae better if we interpret a vector $u = (u_1, \ldots, u_m)$ as a piecewise linear function over the interval $[0, 1)$.

Definition 5.2. Given a vector $u = (u_1, \ldots, u_m)$, the *piecewise linear function*[1] $\mathrm{plf}(u)$ is defined such that

$$\mathrm{plf}(u)(x) = u_i, \qquad \frac{i-1}{m} \leq x < \frac{i}{m},\ 1 \leq i \leq m.$$

In words, the function $\mathrm{plf}(u)$ has the value u_1 on the interval $[0, 1/m)$, the value u_2 on $[1/m, 2/m)$, *etc.*, and the value u_m on the interval $[(m-1)/m, 1)$.

For example, the piecewise linear function associated with the vector

$$u = (2.4, 2.2, 2.15, 2.05, 6.8, 2.8, -1.1, -1.3)$$

is shown in Figure 5.4.

Then each basis vector h_k^j corresponds to the function

$$\psi_k^j = \mathrm{plf}(h_k^j).$$

[1]Piecewise constant function might be a more accurate name.

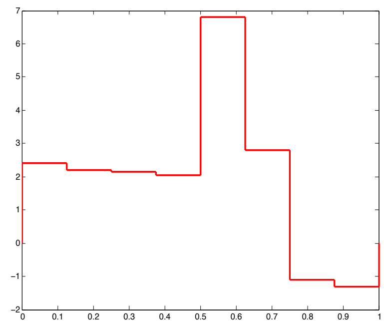


Figure 5.4: The piecewise linear function $\text{plf}(u)$.

In particular, for all n, the Haar basis vectors

$$h_0^0 = w_2 = \underbrace{(1, \ldots, 1, -1, \ldots, -1)}_{2^n}$$

yield the same piecewise linear function ψ given by

$$\psi(x) = \begin{cases} 1 & \text{if} \quad 0 \leq x < 1/2 \\ -1 & \text{if} \quad 1/2 \leq x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

whose graph is shown in Figure 5.5. It is easy to see that ψ_k^j is given by the simple expression

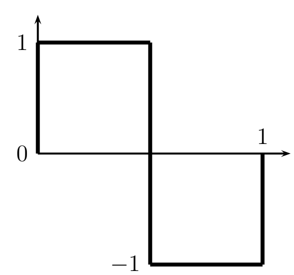


Figure 5.5: The Haar wavelet ψ.

$$\psi_k^j(x) = \psi(2^j x - k), \quad 0 \leq j \leq n-1,\ 0 \leq k \leq 2^j - 1.$$

The above formula makes it clear that ψ_k^j is obtained from ψ by scaling and shifting.

Definition 5.3. The function $\phi_0^0 = \text{plf}(w_1)$ is the piecewise linear function with the constant value 1 on $[0, 1)$, and the functions $\psi_k^j = \text{plf}(h_k^j)$ together with ϕ_0^0 are known as the *Haar wavelets*.

Rather than using W^{-1} to convert a vector u to a vector c of coefficients over the Haar basis, and the matrix W to reconstruct the vector u from its Haar coefficients c, we can use faster algorithms that use averaging and differencing.

If c is a vector of Haar coefficients of dimension 2^n, we compute the sequence of vectors $u^0, u^1, \ldots, u^n$ as follows:

$$\begin{aligned} u^0 &= c \\ u^{j+1} &= u^j \\ u^{j+1}(2i-1) &= u^j(i) + u^j(2^j+i) \\ u^{j+1}(2i) &= u^j(i) - u^j(2^j+i), \end{aligned}$$

for $j = 0, \ldots, n-1$ and $i = 1, \ldots, 2^j$. The reconstructed vector (signal) is $u = u^n$.

If u is a vector of dimension 2^n, we compute the sequence of vectors $c^n, c^{n-1}, \ldots, c^0$ as follows:

$$\begin{aligned} c^n &= u \\ c^j &= c^{j+1} \\ c^j(i) &= (c^{j+1}(2i-1) + c^{j+1}(2i))/2 \\ c^j(2^j+i) &= (c^{j+1}(2i-1) - c^{j+1}(2i))/2, \end{aligned}$$

for $j = n-1, \ldots, 0$ and $i = 1, \ldots, 2^j$. The vector over the Haar basis is $c = c^0$.

We leave it as an exercise to implement the above programs in `Matlab` using two variables u and c, and by building iteratively 2^j. Here is an example of the conversion of a vector to its Haar coefficients for $n = 3$.

Given the sequence $u = (31, 29, 23, 17, -6, -8, -2, -4)$, we get the sequence

$$\begin{aligned} c^3 &= (31, 29, 23, 17, -6, -8, -2, -4) \\ c^2 &= \left(\frac{31+29}{2}, \frac{23+17}{2}, \frac{-6-8}{2}, \frac{-2-4}{2}, \frac{31-29}{2}, \frac{23-17}{2}, \frac{-6-(-8)}{2}, \frac{-2-(-4)}{2}\right) \\ &= (30, 20, -7, -3, 1, 3, 1, 1) \\ c^1 &= \left(\frac{30+20}{2}, \frac{-7-3}{2}, \frac{30-20}{2}, \frac{-7-(-3)}{2}, 1, 3, 1, 1\right) \\ &= (25, -5, 5, -2, 1, 3, 1, 1) \\ c^0 &= \left(\frac{25-5}{2}, \frac{25-(-5)}{2}, 5, -2, 1, 3, 1, 1\right) = (10, 15, 5, -2, 1, 3, 1, 1) \end{aligned}$$

so $c = (10, 15, 5, -2, 1, 3, 1, 1)$. Conversely, given $c = (10, 15, 5, -2, 1, 3, 1, 1)$, we get the sequence

$$\begin{aligned}
u^0 &= (10, 15, 5, -2, 1, 3, 1, 1)\\
u^1 &= (10 + 15, 10 - 15, 5, -2, 1, 3, 1, 1) = (25, -5, 5, -2, 1, 3, 1, 1)\\
u^2 &= (25 + 5, 25 - 5, -5 + (-2), -5 - (-2), 1, 3, 1, 1) = (30, 20, -7, -3, 1, 3, 1, 1)\\
u^3 &= (30 + 1, 30 - 1, 20 + 3, 20 - 3, -7 + 1, -7 - 1, -3 + 1, -3 - 1)\\
&= (31, 29, 23, 17, -6, -8, -2, -4),
\end{aligned}$$

which gives back $u = (31, 29, 23, 17, -6, -8, -2, -4)$.

5.3 Kronecker Product Construction of Haar Matrices

There is another recursive method for constructing the Haar matrix W_n of dimension 2^n that makes it clearer why the columns of W_n are pairwise orthogonal, and why the above algorithms are indeed correct (which nobody seems to prove!). If we split W_n into two $2^n \times 2^{n-1}$ matrices, then the second matrix containing the last 2^{n-1} columns of W_n has a very simple structure: it consists of the vector

$$\underbrace{(1, -1, 0, \ldots, 0)}_{2^n}$$

and $2^{n-1} - 1$ shifted copies of it, as illustrated below for $n = 3$:

$$\begin{pmatrix}
1 & 0 & 0 & 0\\
-1 & 0 & 0 & 0\\
0 & 1 & 0 & 0\\
0 & -1 & 0 & 0\\
0 & 0 & 1 & 0\\
0 & 0 & -1 & 0\\
0 & 0 & 0 & 1\\
0 & 0 & 0 & -1
\end{pmatrix}.$$

Observe that this matrix can be obtained from the identity matrix $I_{2^{n-1}}$, in our example

$$I_4 = \begin{pmatrix}
1 & 0 & 0 & 0\\
0 & 1 & 0 & 0\\
0 & 0 & 1 & 0\\
0 & 0 & 0 & 1
\end{pmatrix},$$

by forming the $2^n \times 2^{n-1}$ matrix obtained by replacing each 1 by the column vector

$$\begin{pmatrix} 1\\ -1 \end{pmatrix}$$

and each zero by the column vector

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Now the first half of W_n, that is the matrix consisting of the first 2^{n-1} columns of W_n, can be obtained from W_{n-1} by forming the $2^n \times 2^{n-1}$ matrix obtained by replacing each 1 by the column vector

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

each -1 by the column vector

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix},$$

and each zero by the column vector

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

For $n = 3$, the first half of W_3 is the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

which is indeed obtained from

$$W_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

using the process that we just described.

These matrix manipulations can be described conveniently using a product operation on matrices known as the Kronecker product.

Definition 5.4. Given a $m \times n$ matrix $A = (a_{ij})$ and a $p \times q$ matrix $B = (b_{ij})$, the *Kronecker product* (or *tensor product*) $A \otimes B$ of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

It can be shown that $\otimes$ is associative and that

$$\begin{aligned}(A \otimes B)(C \otimes D) &= AC \otimes BD \\ (A \otimes B)^\top &= A^\top \otimes B^\top,\end{aligned}$$

whenever AC and BD are well defined. Then it is immediately verified that W_n is given by the following neat recursive equations:

$$W_n = \left(W_{n-1} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad I_{2^{n-1}} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right),$$

with $W_0 = (1)$. If we let

$$B_1 = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

and for $n \geq 1$,

$$B_{n+1} = 2 \begin{pmatrix} B_n & 0 \\ 0 & I_{2^n} \end{pmatrix},$$

then it is not hard to use the Kronecker product formulation of W_n to obtain a rigorous proof of the equation

$$W_n^\top W_n = B_n, \quad \text{for all } n \geq 1.$$

The above equation offers a clean justification of the fact that the columns of W_n are pairwise orthogonal.

Observe that the right block (of size $2^n \times 2^{n-1}$) shows clearly how the detail coefficients in the second half of the vector c are added and subtracted to the entries in the first half of the partially reconstructed vector after $n-1$ steps.

5.4 Multiresolution Signal Analysis with Haar Bases

An important and attractive feature of the Haar basis is that it provides a *multiresolution analysis* of a signal. Indeed, given a signal u, if $c = (c_1, \ldots, c_{2^n})$ is the vector of its Haar coefficients, the coefficients with low index give coarse information about u, and the coefficients with high index represent fine information. For example, if u is an audio signal corresponding to a Mozart concerto played by an orchestra, c_1 corresponds to the "background noise," c_2 to the bass, c_3 to the first cello, c_4 to the second cello, c_5, c_6, c_7, c_7 to the violas, then the violins, *etc.* This multiresolution feature of wavelets can be exploited to compress a signal, that is, to use fewer coefficients to represent it. Here is an example.

Consider the signal

$$u = (2.4, 2.2, 2.15, 2.05, 6.8, 2.8, -1.1, -1.3),$$

whose Haar transform is

$$c = (2, 0.2, 0.1, 3, 0.1, 0.05, 2, 0.1).$$

The piecewise-linear curves corresponding to u and c are shown in Figure 5.6. Since some of the coefficients in c are small (smaller than or equal to 0.2) we can compress c by replacing them by 0. We get

$$c_2 = (2, 0, 0, 3, 0, 0, 2, 0),$$

and the reconstructed signal is

$$u_2 = (2, 2, 2, 2, 7, 3, -1, -1).$$

The piecewise-linear curves corresponding to u_2 and c_2 are shown in Figure 5.7.

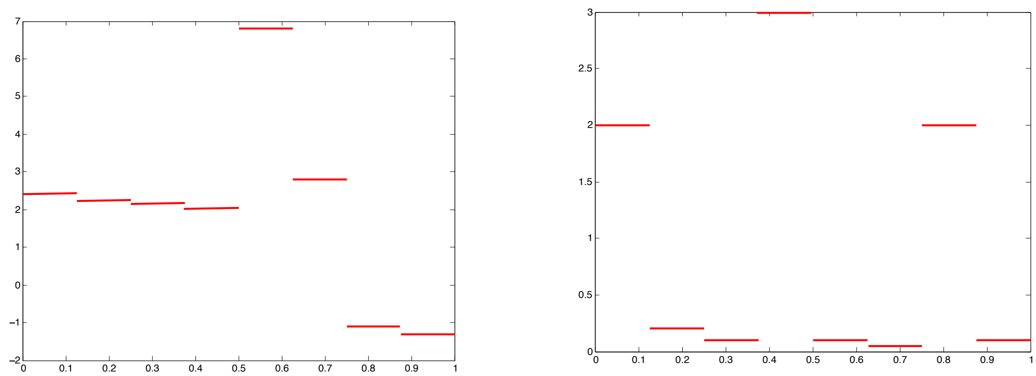


Figure 5.6: A signal and its Haar transform.

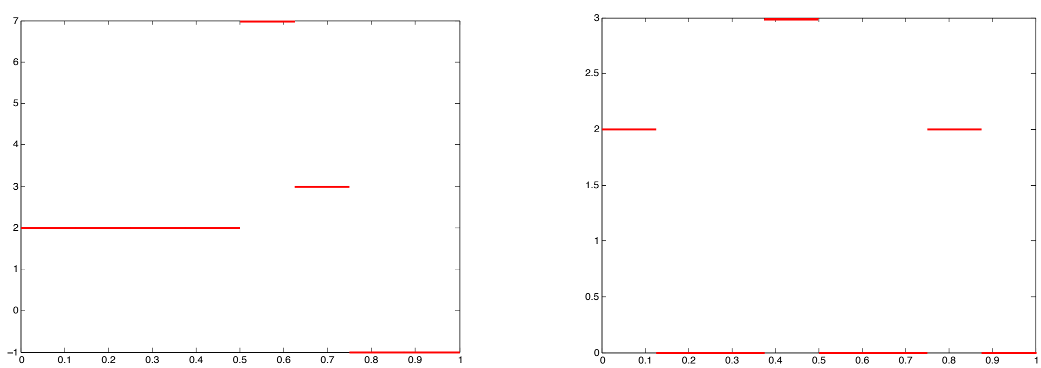


Figure 5.7: A compressed signal and its compressed Haar transform.

An interesting (and amusing) application of the Haar wavelets is to the compression of audio signals. It turns out that if your type `load handel` in `Matlab` an audio file will be loaded in a vector denoted by y, and if you type `sound(y)`, the computer will play this piece of music. You can convert y to its vector of Haar coefficients c. The length of y is 73113,

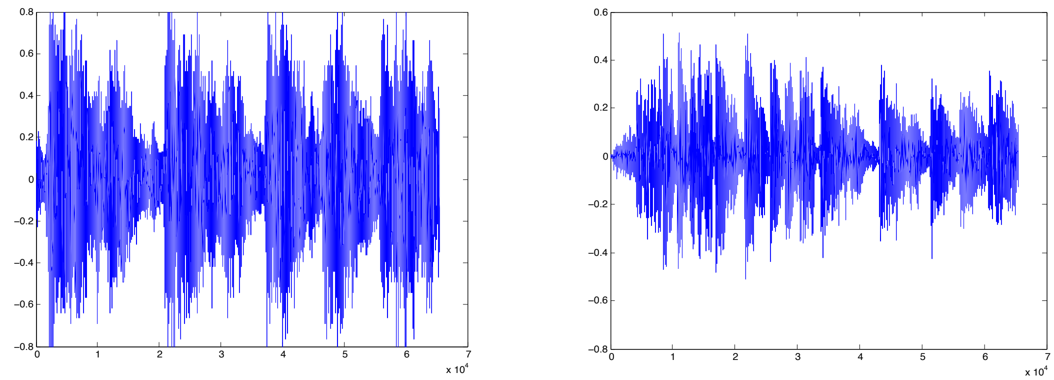


Figure 5.8: The signal "handel" and its Haar transform.

so first tuncate the tail of y to get a vector of length $65536 = 2^{16}$. A plot of the signals corresponding to y and c is shown in Figure 5.8. Then run a program that sets all coefficients of c whose absolute value is less that 0.05 to zero. This sets 37272 coefficients to 0. The resulting vector c_2 is converted to a signal y_2. A plot of the signals corresponding to y_2 and c_2 is shown in Figure 5.9. When you type `sound(y2)`, you find that the music doesn't differ

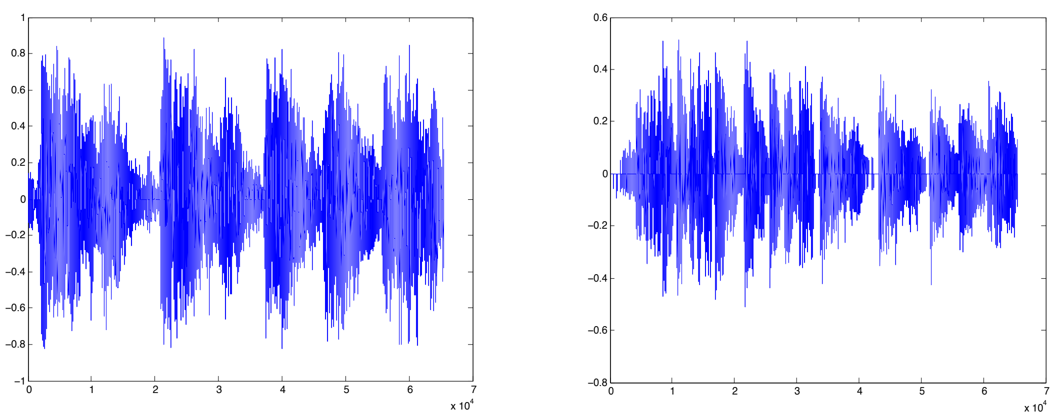


Figure 5.9: The compressed signal "handel" and its Haar transform.

much from the original, although it sounds less crisp. You should play with other numbers greater than or less than 0.05. You should hear what happens when you type `sound(c)`. It plays the music corresponding to the Haar transform c of y, and it is quite funny.

5.5 Haar Transform for Digital Images

Another neat property of the Haar transform is that it can be instantly generalized to matrices (even rectangular) without any extra effort! This allows for the compression of digital images. But first we address the issue of normalization of the Haar coefficients. As we observed earlier, the $2^n \times 2^n$ matrix W_n of Haar basis vectors has orthogonal columns, but its columns do not have unit length. As a consequence, $W_n^\top$ is not the inverse of W_n, but rather the matrix

$$W_n^{-1} = D_n W_n^\top$$

with $D_n = \mathrm{diag}\Big(2^{-n}, \underbrace{2^{-n}}_{2^0}, \underbrace{2^{-(n-1)}, 2^{-(n-1)}}_{2^1}, \underbrace{2^{-(n-2)}, \ldots, 2^{-(n-2)}}_{2^2}, \ldots, \underbrace{2^{-1}, \ldots, 2^{-1}}_{2^{n-1}}\Big)$.

Definition 5.5. The orthogonal matrix

$$H_n = W_n D_n^{\frac{1}{2}}$$

whose columns are the normalized Haar basis vectors, with

$$D_n^{\frac{1}{2}} = \mathrm{diag}\Big(2^{-\frac{n}{2}}, \underbrace{2^{-\frac{n}{2}}}_{2^0}, \underbrace{2^{-\frac{n-1}{2}}, 2^{-\frac{n-1}{2}}}_{2^1}, \underbrace{2^{-\frac{n-2}{2}}, \ldots, 2^{-\frac{n-2}{2}}}_{2^2}, \ldots, \underbrace{2^{-\frac{1}{2}}, \ldots, 2^{-\frac{1}{2}}}_{2^{n-1}}\Big)$$

is called the *normalized Haar transform matrix*. Given a vector (signal) u, we call $c = H_n^\top u$ the *normalized Haar coefficients* of u.

Because H_n is orthogonal, $H_n^{-1} = H_n^\top$.

Then a moment of reflection shows that we have to slightly modify the algorithms to compute $H_n^\top u$ and $H_n c$ as follows: When computing the sequence of u^js, use

$$\begin{aligned} u^{j+1}(2i-1) &= (u^j(i) + u^j(2^j + i))/\sqrt{2} \\ u^{j+1}(2i) &= (u^j(i) - u^j(2^j + i))/\sqrt{2}, \end{aligned}$$

and when computing the sequence of c^js, use

$$\begin{aligned} c^j(i) &= (c^{j+1}(2i-1) + c^{j+1}(2i))/\sqrt{2} \\ c^j(2^j + i) &= (c^{j+1}(2i-1) - c^{j+1}(2i))/\sqrt{2}. \end{aligned}$$

Note that things are now more symmetric, at the expense of a division by $\sqrt{2}$. However, for long vectors, it turns out that these algorithms are numerically more stable.

Remark: Some authors (for example, Stollnitz, Derose and Salesin [168]) rescale c by $1/\sqrt{2^n}$ and u by $\sqrt{2^n}$. This is because the norm of the basis functions ψ_k^j is not equal to 1 (under the inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. The normalized basis functions are the functions $\sqrt{2^j}\psi^j_k$.

Let us now explain the 2D version of the Haar transform. We describe the version using the matrix W_n, the method using H_n being identical (except that $H_n^{-1} = H_n^\top$, but this does not hold for W_n^{-1}). Given a $2^m \times 2^n$ matrix A, we can first convert the *rows* of A to their Haar coefficients using the Haar transform W_n^{-1}, obtaining a matrix B, and then convert the *columns* of B to their Haar coefficients, using the matrix W_m^{-1}. Because columns and rows are exchanged in the first step,

$$B = A(W_n^{-1})^\top,$$

and in the second step $C = W_m^{-1}B$, thus, we have

$$C = W_m^{-1}A(W_n^{-1})^\top = D_m W_m^\top A W_n\, D_n.$$

In the other direction, given a $2^m \times 2^n$ matrix C of Haar coefficients, we reconstruct the matrix A (the image) by first applying W_m to the columns of C, obtaining B, and then $W_n^\top$ to the rows of B. Therefore

$$A = W_m C W_n^\top.$$

Of course, we don't actually have to invert W_m and W_n and perform matrix multiplications. We just have to use our algorithms using averaging and differencing. Here is an example.

If the data matrix (the image) is the 8×8 matrix

$$A = \begin{pmatrix} 64 & 2 & 3 & 61 & 60 & 6 & 7 & 57 \\ 9 & 55 & 54 & 12 & 13 & 51 & 50 & 16 \\ 17 & 47 & 46 & 20 & 21 & 43 & 42 & 24 \\ 40 & 26 & 27 & 37 & 36 & 30 & 31 & 33 \\ 32 & 34 & 35 & 29 & 28 & 38 & 39 & 25 \\ 41 & 23 & 22 & 44 & 45 & 19 & 18 & 48 \\ 49 & 15 & 14 & 52 & 53 & 11 & 10 & 56 \\ 8 & 58 & 59 & 5 & 4 & 62 & 63 & 1 \end{pmatrix},$$

then applying our algorithms, we find that

$$C = \begin{pmatrix} 32.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0.5 & 0.5 & 27 & -25 & 23 & -21 \\ 0 & 0 & -0.5 & -0.5 & -11 & 9 & -7 & 5 \\ 0 & 0 & 0.5 & 0.5 & -5 & 7 & -9 & 11 \\ 0 & 0 & -0.5 & -0.5 & 21 & -23 & 25 & -27 \end{pmatrix}.$$

As we can see, C has more zero entries than A; it is a compressed version of A. We can further compress C by setting to 0 all entries of absolute value at most 0.5. Then we get

$$C_2 = \begin{pmatrix} 32.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 4 & -4 & 4 & -4 \\ 0 & 0 & 0 & 0 & 27 & -25 & 23 & -21 \\ 0 & 0 & 0 & 0 & -11 & 9 & -7 & 5 \\ 0 & 0 & 0 & 0 & -5 & 7 & -9 & 11 \\ 0 & 0 & 0 & 0 & 21 & -23 & 25 & -27 \end{pmatrix}.$$

We find that the reconstructed image is

$$A_2 = \begin{pmatrix} 63.5 & 1.5 & 3.5 & 61.5 & 59.5 & 5.5 & 7.5 & 57.5 \\ 9.5 & 55.5 & 53.5 & 11.5 & 13.5 & 51.5 & 49.5 & 15.5 \\ 17.5 & 47.5 & 45.5 & 19.5 & 21.5 & 43.5 & 41.5 & 23.5 \\ 39.5 & 25.5 & 27.5 & 37.5 & 35.5 & 29.5 & 31.5 & 33.5 \\ 31.5 & 33.5 & 35.5 & 29.5 & 27.5 & 37.5 & 39.5 & 25.5 \\ 41.5 & 23.5 & 21.5 & 43.5 & 45.5 & 19.5 & 17.5 & 47.5 \\ 49.5 & 15.5 & 13.5 & 51.5 & 53.5 & 11.5 & 9.5 & 55.5 \\ 7.5 & 57.5 & 59.5 & 5.5 & 3.5 & 61.5 & 63.5 & 1.5 \end{pmatrix},$$

which is pretty close to the original image matrix A.

It turns out that `Matlab` has a wonderful command, `image(X)` (also `imagesc(X)`, which often does a better job), which displays the matrix X has an image in which each entry is shown as a little square whose gray level is proportional to the numerical value of that entry (lighter if the value is higher, darker if the value is closer to zero; negative values are treated as zero). The images corresponding to A and C are shown in Figure 5.10. The compressed images corresponding to A_2 and C_2 are shown in Figure 5.11. The compressed versions appear to be indistinguishable from the originals!

If we use the normalized matrices H_m and H_n, then the equations relating the image matrix A and its normalized Haar transform C are

$$\begin{aligned} C &= H_m^{\top} A H_n \\ A &= H_m C H_n^{\top}. \end{aligned}$$

The Haar transform can also be used to send large images progressively over the internet. Indeed, we can start sending the Haar coefficients of the matrix C starting from the coarsest coefficients (the first column from top down, then the second column, *etc.*), and at the receiving end we can start reconstructing the image as soon as we have received enough data.

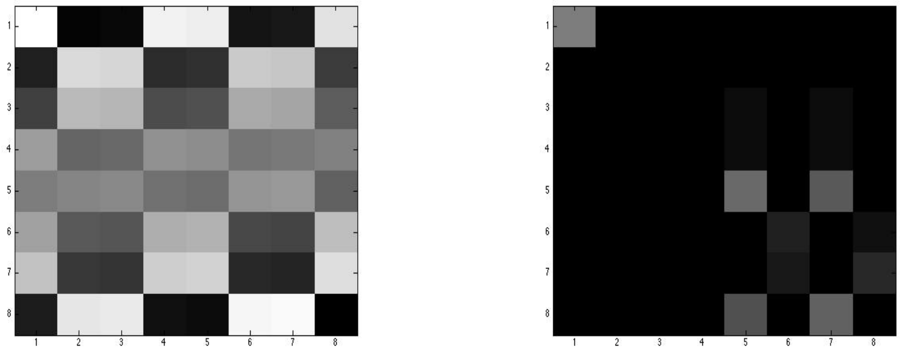


Figure 5.10: An image and its Haar transform.

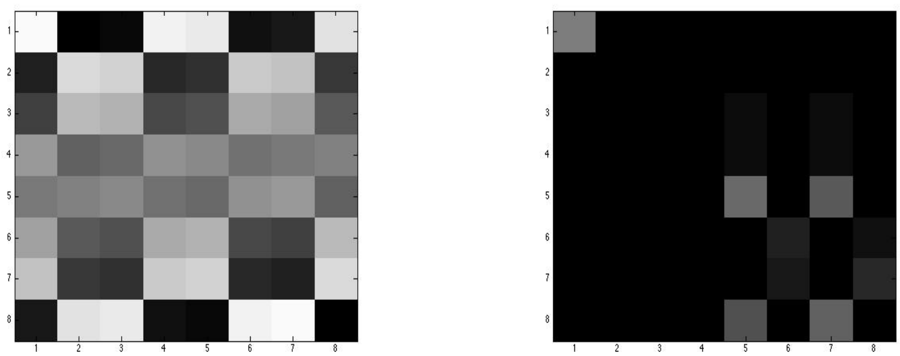


Figure 5.11: Compressed image and its Haar transform.

Observe that instead of performing all rounds of averaging and differencing on each row and each column, we can perform partial encoding (and decoding). For example, we can perform a single round of averaging and differencing for each row and each column. The result is an image consisting of four subimages, where the top left quarter is a coarser version of the original, and the rest (consisting of three pieces) contain the finest detail coefficients. We can also perform two rounds of averaging and differencing, or three rounds, *etc.* The second round of averaging and differencing is applied to the top left quarter of the image. Generally, the kth round is applied to the $2^{m+1-k} \times 2^{n+1-k}$ submatrix consisting of the first 2^{m+1-k} rows and the first 2^{n+1-k} columns $(1 \leq k \leq n)$ of the matrix obtained at the end of the previous round. This process is illustrated on the image shown in Figure 5.12. The result of performing one round, two rounds, three rounds, and nine rounds of averaging is shown in Figure 5.13. Since our images have size 512×512, nine rounds of averaging yields the Haar transform, displayed as the image on the bottom right. The original image has completely

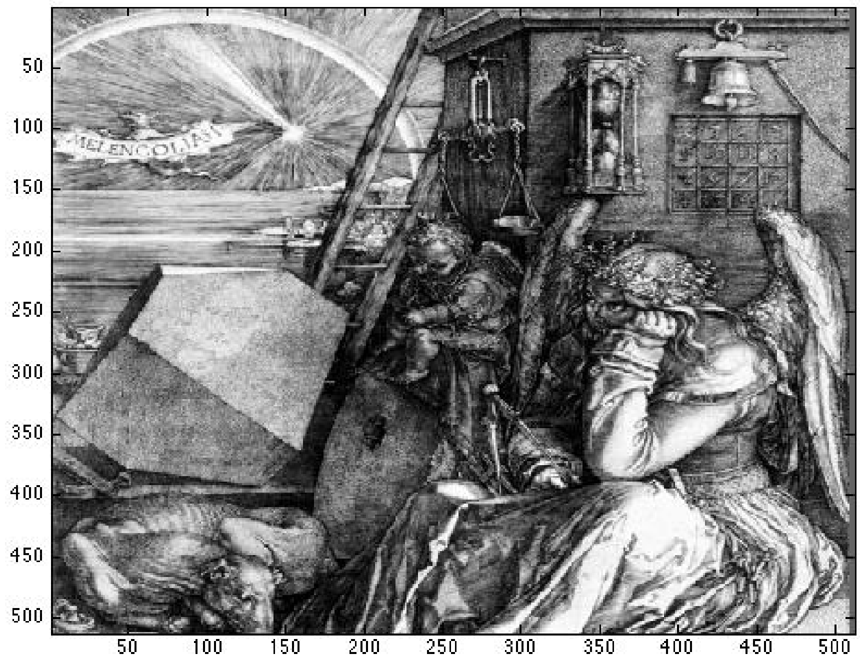


Figure 5.12: Original drawing by Durer.

disappeared! We leave it as a fun exercise to modify the algorithms involving averaging and differencing to perform k rounds of averaging/differencing. The reconstruction algorithm is a little tricky.

A nice and easily accessible account of wavelets and their uses in image processing and computer graphics can be found in Stollnitz, Derose and Salesin [168]. A very detailed account is given in Strang and and Nguyen [172], but this book assumes a fair amount of background in signal processing.

We can find easily a basis of $2^n \times 2^n = 2^{2n}$ vectors w_{ij} ($2^n \times 2^n$ matrices) for the linear map that reconstructs an image from its Haar coefficients, in the sense that for any $2^n \times 2^n$ matrix C of Haar coefficients, the image matrix A is given by

$$A = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} c_{ij} w_{ij}.$$

Indeed, the matrix w_{ij} is given by the so-called outer product

$$w_{ij} = w_i (w_j)^\top.$$

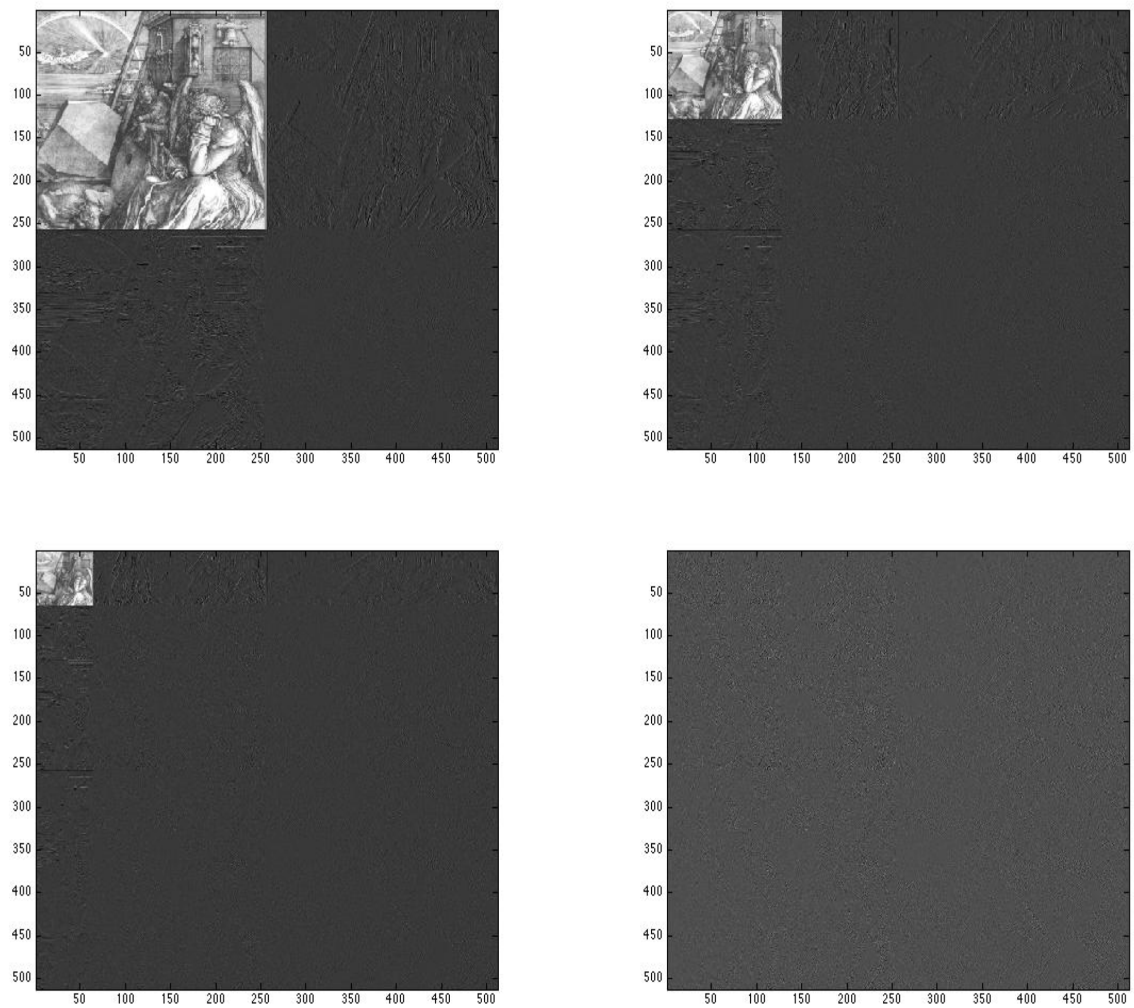


Figure 5.13: Haar tranforms after one, two, three, and nine rounds of averaging.

Similarly, there is a basis of $2^n \times 2^n = 2^{2n}$ vectors h_{ij} ($2^n \times 2^n$ matrices) for the 2D Haar transform, in the sense that for any $2^n \times 2^n$ matrix A, its matrix C of Haar coefficients is given by

$$C = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} a_{ij} h_{ij}.$$

If the columns of W^{-1} are $w'_1, \ldots, w'_{2^n}$, then

$$h_{ij} = w'_i (w'_j)^\top.$$

We leave it as exercise to compute the bases (w_{ij}) and (h_{ij}) for $n = 2$, and to display the corresponding images using the command `imagesc`.

5.6 Hadamard Matrices

There is another famous family of matrices somewhat similar to Haar matrices, but these matrices have entries $+1$ and -1 (no zero entries).

Definition 5.6. A real $n \times n$ matrix H is a *Hadamard matrix* if $h_{ij} = \pm 1$ for all i, j such that $1 \leq i, j \leq n$ and if

$$H^\top H = nI_n.$$

Thus the columns of a Hadamard matrix are pairwise orthogonal. Because H is a square matrix, the equation $H^\top H = nI_n$ shows that H is invertible, so we also have $HH^\top = nI_n$. The following matrices are example of Hadamard matrices:

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

and

$$H_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}.$$

A natural question is to determine the positive integers n for which a Hadamard matrix of dimension n exists, but surprisingly this is an *open problem*. The *Hadamard conjecture* is that for every positive integer of the form $n = 4k$, there is a Hadamard matrix of dimension n.

What is known is a necessary condition and various sufficient conditions.

Theorem 5.1. *If H is an $n \times n$ Hadamard matrix, then either $n = 1, 2$, or $n = 4k$ for some positive integer k.*

Sylvester introduced a family of Hadamard matrices and proved that there are Hadamard matrices of dimension $n = 2^m$ for all $m \geq 1$ using the following construction.

Proposition 5.2. *(Sylvester, 1867) If H is a Hadamard matrix of dimension n, then the block matrix of dimension $2n$,*

$$\begin{pmatrix} H & H \\ H & -H \end{pmatrix},$$

is a Hadamard matrix.

If we start with

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

we obtain an infinite family of symmetric Hadamard matrices usually called *Sylvester–Hadamard* matrices and denoted by H_{2^m}. The Sylvester–Hadamard matrices H_2, H_4 and H_8 are shown on the previous page.

In 1893, Hadamard gave examples of Hadamard matrices for $n = 12$ and $n = 20$. At the present, Hadamard matrices are known for all $n = 4k \leq 1000$, *except for* $n = 668, 716$, *and* 892.

Hadamard matrices have various applications to error correcting codes, signal processing, and numerical linear algebra; see Seberry, Wysocki and Wysocki [154] and Tropp [177]. For example, there is a code based on H_{32} that can correct 7 errors in any 32-bit encoded block, and can detect an eighth. This code was used on a Mariner spacecraft in 1969 to transmit pictures back to the earth.

For every $m \geq 0$, the piecewise affine functions $\mathrm{plf}((H_{2^m})_i)$ associated with the 2^m rows of the Sylvester–Hadamard matrix H_{2^m} are functions on $[0, 1]$ known as the *Walsh functions*. It is customary to index these 2^m functions by the integers $0, 1, \ldots, 2^m - 1$ in such a way that the Walsh function $\mathrm{Wal}(k, t)$ is equal to the function $\mathrm{plf}((H_{2^m})_i)$ associated with the Row i of H_{2^m} that contains k changes of signs between consecutive groups of $+1$ and consecutive groups of -1. For example, the fifth row of H_8, namely

$$\begin{pmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \end{pmatrix},$$

has five consecutive blocks of $+1$s and -1s, four sign changes between these blocks, and thus is associated with $\mathrm{Wal}(4, t)$. In particular, Walsh functions corresponding to the rows of H_8 (from top down) are:

$$\mathrm{Wal}(0, t),\ \mathrm{Wal}(7, t),\ \mathrm{Wal}(3, t),\ \mathrm{Wal}(4, t),\ \mathrm{Wal}(1, t),\ \mathrm{Wal}(6, t),\ \mathrm{Wal}(2, t),\ \mathrm{Wal}(5, t).$$

Because of the connection between Sylvester–Hadamard matrices and Walsh functions, Sylvester–Hadamard matrices are called *Walsh–Hadamard matrices* by some authors. For every m, the 2^m Walsh functions are pairwise orthogonal. The countable set of Walsh functions $\mathrm{Wal}(k,t)$ for all $m \geq 0$ and all k such that $0 \leq k \leq 2^m - 1$ can be ordered in such a way that it is an orthogonal Hilbert basis of the Hilbert space $\mathrm{L}^2([0,1)]$; see Seberry, Wysocki and Wysocki [154].

The Sylvester–Hadamard matrix H_{2^m} plays a role in various algorithms for dimension reduction and low-rank matrix approximation. There is a type of structured dimension-reduction map known as the *subsampled randomized Hadamard transform*, for short SRHT; see Tropp [177] and Halko, Martinsson and Tropp [86]. For $\ell \ll n = 2^m$, an *SRHT matrix* is an $\ell \times n$ matrix of the form

$$\Phi = \sqrt{\frac{n}{\ell}} RHD,$$

where

1. D is a random $n \times n$ diagonal matrix whose entries are independent random signs.

2. $H = n^{-1/2} H_n$, a normalized Sylvester–Hadamard matrix of dimension n.

3. R is a random $\ell \times n$ matrix that restricts an n-dimensional vector to ℓ coordinates, chosen uniformly at random.

It is explained in Tropp [177] that for any input x such that $\|x\|_2 = 1$, the probability that $|(HDx)_i| \geq \sqrt{n^{-1}\log(n)}$ for any i is quite small. Thus HD has the effect of "flattening" the input x. The main result about the SRHT is that it preserves the geometry of an entire subspace of vectors; see Tropp [177] (Theorem 1.3).

5.7 Summary

The main concepts and results of this chapter are listed below:

- Haar basis vectors and a glimpse at *Haar wavelets*.

- *Kronecker product* (or *tensor product*) of matrices.

- Hadamard and Sylvester–Hadamard matrices.

- Walsh functions.

5.8 Problems

Problem 5.1. (Haar extravaganza) Consider the matrix

$$W_{3,3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

(1) Show that given any vector $c = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8)$, the result $W_{3,3}c$ of applying $W_{3,3}$ to c is

$$W_{3,3}c = (c_1 + c_5, c_1 - c_5, c_2 + c_6, c_2 - c_6, c_3 + c_7, c_3 - c_7, c_4 + c_8, c_4 - c_8),$$

the last step in reconstructing a vector from its Haar coefficients.

(2) Prove that the inverse of $W_{3,3}$ is $(1/2)W_{3,3}^\top$. Prove that the columns and the rows of $W_{3,3}$ are orthogonal.

(3) Let $W_{3,2}$ and $W_{3,1}$ be the following matrices:

$$W_{3,2} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad W_{3,1} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Show that given any vector $c = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8)$, the result $W_{3,2}c$ of applying $W_{3,2}$ to c is

$$W_{3,2}c = (c_1 + c_3, c_1 - c_3, c_2 + c_4, c_2 - c_4, c_5, c_6, c_7, c_8),$$

the second step in reconstructing a vector from its Haar coefficients, and the result $W_{3,1}c$ of applying $W_{3,1}$ to c is

$$W_{3,1}c = (c_1 + c_2, c_1 - c_2, c_3, c_4, c_5, c_6, c_7, c_8),$$

the first step in reconstructing a vector from its Haar coefficients.

Conclude that

$$W_{3,3}W_{3,2}W_{3,1} = W_3,$$

the Haar matrix

$$W_3 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

Hint. First check that

$$W_{3,2}W_{3,1} = \begin{pmatrix} W_2 & 0_{4,4} \\ 0_{4,4} & I_4 \end{pmatrix},$$

where

$$W_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

(4) Prove that the columns and the rows of $W_{3,2}$ and $W_{3,1}$ are orthogonal. Deduce from this that the columns of W_3 are orthogonal, and the rows of W_3^{-1} are orthogonal. Are the rows of W_3 orthogonal? Are the columns of W_3^{-1} orthogonal? Find the inverse of $W_{3,2}$ and the inverse of $W_{3,1}$.

Problem 5.2. This is a continuation of Problem 5.1.

(1) For any $n \geq 2$, the $2^n \times 2^n$ matrix $W_{n,n}$ is obtained form the two rows

$$\underbrace{1, 0, \ldots, 0}_{2^{n-1}}, \underbrace{1, 0, \ldots, 0}_{2^{n-1}}$$
$$\underbrace{1, 0, \ldots, 0}_{2^{n-1}}, \underbrace{-1, 0, \ldots, 0}_{2^{n-1}}$$

by shifting them $2^{n-1} - 1$ times over to the right by inserting a zero on the left each time.

Given any vector $c = (c_1, c_2, \ldots, c_{2^n})$, show that $W_{n,n}c$ is the result of the last step in the process of reconstructing a vector from its Haar coefficients c. Prove that $W_{n,n}^{-1} = (1/2)W_{n,n}^{\top}$, and that the columns and the rows of $W_{n,n}$ are orthogonal.

(2) Given a $m \times n$ matrix $A = (a_{ij})$ and a $p \times q$ matrix $B = (b_{ij})$, the *Kronecker product* (or *tensor product*) $A \otimes B$ of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

It can be shown (and you may use these facts without proof) that $\otimes$ is associative and that

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$
$$(A \otimes B)^\top = A^\top \otimes B^\top,$$

whenever AC and BD are well defined.

Check that

$$W_{n,n} = \begin{pmatrix} I_{2^{n-1}} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} & I_{2^{n-1}} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \end{pmatrix},$$

and that

$$W_n = \begin{pmatrix} W_{n-1} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} & I_{2^{n-1}} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \end{pmatrix}.$$

Use the above to reprove that

$$W_{n,n} W_{n,n}^\top = 2I_{2^n}.$$

Let

$$B_1 = 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

and for $n \geq 1$,

$$B_{n+1} = 2\begin{pmatrix} B_n & 0 \\ 0 & I_{2^n} \end{pmatrix}.$$

Prove that

$$W_n^\top W_n = B_n, \quad \text{for all } n \geq 1.$$

(3) The matrix $W_{n,i}$ is obtained from the matrix $W_{i,i}$ $(1 \leq i \leq n-1)$ as follows:

$$W_{n,i} = \begin{pmatrix} W_{i,i} & 0_{2^i, 2^n - 2^i} \\ 0_{2^n - 2^i, 2^i} & I_{2^n - 2^i} \end{pmatrix}.$$

It consists of four blocks, where $0_{2^i, 2^n - 2^i}$ and $0_{2^n - 2^i, 2^i}$ are matrices of zeros and $I_{2^n - 2^i}$ is the identity matrix of dimension $2^n - 2^i$.

Explain what $W_{n,i}$ does to c and prove that

$$W_{n,n} W_{n,n-1} \cdots W_{n,1} = W_n,$$

where W_n is the Haar matrix of dimension 2^n.

Hint. Use induction on k, with the induction hypothesis

$$W_{n,k} W_{n,k-1} \cdots W_{n,1} = \begin{pmatrix} W_k & 0_{2^k, 2^n - 2^k} \\ 0_{2^n - 2^k, 2^k} & I_{2^n - 2^k} \end{pmatrix}.$$

Prove that the columns and rows of $W_{n,k}$ are orthogonal, and use this to prove that the columns of W_n and the rows of W_n^{-1} are orthogonal. Are the rows of W_n orthogonal? Are the columns of W_n^{-1} orthogonal? Prove that

$$W_{n,k}^{-1} = \begin{pmatrix} \frac{1}{2}W_{k,k}^{\top} & 0_{2^k,2^n-2^k} \\ 0_{2^n-2^k,2^k} & I_{2^n-2^k} \end{pmatrix}.$$

Problem 5.3. Prove that if H is a Hadamard matrix of dimension n, then the block matrix of dimension $2n$,

$$\begin{pmatrix} H & H \\ H & -H \end{pmatrix},$$

is a Hadamard matrix.

Problem 5.4. Plot the graphs of the eight Walsh functions $\mathrm{Wal}(k,t)$ for $k = 0, 1, \ldots, 7$.

Problem 5.5. Describe a recursive algorithm to compute the product $H_{2^m}\, x$ of the Sylvester–Hadamard matrix H_{2^m} by a vector $x \in \mathbb{R}^{2^m}$ that uses m recursive calls.

Chapter 6

Direct Sums

In this chapter all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$.

6.1 Sums, Direct Sums, Direct Products

There are some useful ways of forming new vector spaces from older ones, in particular, direct products and direct sums. Regarding direct sums, there is a subtle point, which is that if we attempt to define the direct sum $E \coprod F$ of two vector spaces using the cartesian product $E \times F$, we don't quite get the right notion because elements of $E \times F$ are ordered pairs, but we want $E \coprod F = F \coprod E$. Thus, we want to think of the elements of $E \coprod F$ as unordrered pairs of elements. It is possible to do so by considering the direct sum of a *family* $(E_i)_{i\in\{1,2\}}$, and more generally of a family $(E_i)_{i\in I}$. For simplicity, we begin by considering the case where $I = \{1, 2\}$.

Definition 6.1. Given a family $(E_i)_{i\in\{1,2\}}$ of two vector spaces, we define the *(external) direct sum* $E_1 \coprod E_2$ (or *coproduct*) of the family $(E_i)_{i\in\{1,2\}}$ as the set

$$E_1 \coprod E_2 = \{\{\langle 1, u\rangle, \langle 2, v\rangle\} \mid u \in E_1,\ v \in E_2\},$$

with addition

$$\{\langle 1, u_1\rangle, \langle 2, v_1\rangle\} + \{\langle 1, u_2\rangle, \langle 2, v_2\rangle\} = \{\langle 1, u_1 + u_2\rangle, \langle 2, v_1 + v_2\rangle\},$$

and scalar multiplication

$$\lambda\{\langle 1, u\rangle, \langle 2, v\rangle\} = \{\langle 1, \lambda u\rangle, \langle 2, \lambda v\rangle\}.$$

We define the *injections* $in_1 \colon E_1 \to E_1 \coprod E_2$ and $in_2 \colon E_2 \to E_1 \coprod E_2$ as the linear maps defined such that,

$$in_1(u) = \{\langle 1, u\rangle, \langle 2, 0\rangle\},$$

and

$$in_2(v) = \{\langle 1, 0\rangle, \langle 2, v\rangle\}.$$

Note that

$$E_2 \coprod E_1 = \{\{\langle 2, v\rangle, \langle 1, u\rangle\} \mid v \in E_2,\ u \in E_1\} = E_1 \coprod E_2.$$

Thus, every member $\{\langle 1, u\rangle, \langle 2, v\rangle\}$ of $E_1 \coprod E_2$ can be viewed as an *unordered pair* consisting of the two vectors u and v, tagged with the index 1 and 2, respectively.

Remark: In fact, $E_1 \coprod E_2$ is just the product $\prod_{i\in\{1,2\}} E_i$ of the family $(E_i)_{i\in\{1,2\}}$.

This is not to be confused with the cartesian product $E_1 \times E_2$. The vector space $E_1 \times E_2$ is the set of all ordered pairs $\langle u, v\rangle$, where $u \in E_1$, and $v \in E_2$, with addition and multiplication by a scalar defined such that

$$\begin{aligned} \langle u_1, v_1\rangle + \langle u_2, v_2\rangle &= \langle u_1 + u_2, v_1 + v_2\rangle, \\ \lambda\langle u, v\rangle &= \langle \lambda u, \lambda v\rangle. \end{aligned}$$

There is a bijection between $\prod_{i\in\{1,2\}} E_i$ and $E_1 \times E_2$, but as we just saw, elements of $\prod_{i\in\{1,2\}} E_i$ are certain sets. The product $E_1 \times \cdots \times E_n$ of any number of vector spaces can also be defined. We will do this shortly.

The following property holds.

Proposition 6.1. *Given any two vector spaces, E_1 and E_2, the set $E_1 \coprod E_2$ is a vector space. For every pair of linear maps, $f\colon E_1 \to G$ and $g\colon E_2 \to G$, there is a unique linear map, $f + g\colon E_1 \coprod E_2 \to G$, such that $(f + g) \circ in_1 = f$ and $(f + g) \circ in_2 = g$, as in the following diagram:*

$$\begin{tikzcd} E_1 \arrow[d, "in_1"'] \arrow[dr, "f"] & \\ E_1 \coprod E_2 \arrow[r, "f+g"] & G \\ E_2 \arrow[u, "in_2"] \arrow[ur, "g"'] & \end{tikzcd}$$

Proof. Define

$$(f + g)(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = f(u) + g(v),$$

for every $u \in E_1$ and $v \in E_2$. It is immediately verified that $f + g$ is the unique linear map with the required properties. $\square$

We already noted that $E_1 \coprod E_2$ is in bijection with $E_1 \times E_2$. If we define the *projections* $\pi_1\colon E_1 \coprod E_2 \to E_1$ and $\pi_2\colon E_1 \coprod E_2 \to E_2$, such that

$$\pi_1(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = u,$$

and

$$\pi_2(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = v,$$

we have the following proposition.

Proposition 6.2. *Given any two vector spaces, E_1 and E_2, for every pair of linear maps, $f\colon D \to E_1$ and $g\colon D \to E_2$, there is a unique linear map, $f \times g\colon D \to E_1 \coprod E_2$, such that $\pi_1 \circ (f \times g) = f$ and $\pi_2 \circ (f \times g) = g$, as in the following diagram:*

$$\begin{tikzcd}
 & E_1 \\
D \arrow[ur, "f"] \arrow[r, "f\times g"] \arrow[dr, "g"'] & E_1 \coprod E_2 \arrow[u, "\pi_1"'] \arrow[d, "\pi_2"] \\
 & E_2
\end{tikzcd}$$

Proof. Define

$$(f \times g)(w) = \{\langle 1, f(w)\rangle, \langle 2, g(w)\rangle\},$$

for every $w \in D$. It is immediately verified that $f \times g$ is the unique linear map with the required properties. $\square$

Remark: It is a peculiarity of linear algebra that direct sums and products of finite families are isomorphic. However, this is no longer true for products and sums of infinite families.

When U, V are subspaces of a vector space E, letting $i_1\colon U \to E$ and $i_2\colon V \to E$ be the inclusion maps, if $U \coprod V$ is isomomorphic to E under the map $i_1 + i_2$ given by Proposition 6.1, we say that E *is a direct sum* of U and V, and we write $E = U \coprod V$ (with a slight abuse of notation, since E and $U \coprod V$ are only isomorphic). It is also convenient to define the sum $U_1 + \cdots + U_p$ and the internal direct sum $U_1 \oplus \cdots \oplus U_p$ of any number of subspaces of E.

Definition 6.2. Given $p \geq 2$ vector spaces $E_1, \ldots, E_p$, the product $F = E_1 \times \cdots \times E_p$ can be made into a vector space by defining addition and scalar multiplication as follows:

$$\begin{aligned}
(u_1, \ldots, u_p) + (v_1, \ldots, v_p) &= (u_1 + v_1, \ldots, u_p + v_p) \\
\lambda(u_1, \ldots, u_p) &= (\lambda u_1, \ldots, \lambda u_p),
\end{aligned}$$

for all $u_i, v_i \in E_i$ and all $\lambda \in \mathbb{R}$. The zero vector of $E_1 \times \cdots \times E_p$ is the p-tuple

$$(\underbrace{0, \ldots, 0}_{p}),$$

where the ith zero is the zero vector of E_i.

With the above addition and multiplication, the vector space $F = E_1 \times \cdots \times E_p$ is called the *direct product* of the vector spaces $E_1, \ldots, E_p$.

As a special case, when $E_1 = \cdots = E_p = \mathbb{R}$, we find again the vector space $F = \mathbb{R}^p$. The *projection maps* $pr_i\colon E_1 \times \cdots \times E_p \to E_i$ given by

$$pr_i(u_1, \ldots, u_p) = u_i$$

are clearly linear. Similarly, the maps $\mathrm{in}_i\colon E_i \to E_1 \times \cdots \times E_p$ given by

$$\mathrm{in}_i(u_i) = (0, \ldots, 0, u_i, 0, \ldots, 0)$$

are injective and linear. If $\dim(E_i) = n_i$ and if $(e_1^i, \ldots, e_{n_i}^i)$ is a basis of E_i for $i = 1, \ldots, p$, then it is easy to see that the $n_1 + \cdots + n_p$ vectors

$$\begin{array}{ccc}
(e_1^1, 0, \ldots, 0), & \ldots, & (e_{n_1}^1, 0, \ldots, 0), \\
\vdots & \vdots & \vdots \\
(0, \ldots, 0, e_1^i, 0, \ldots, 0), & \ldots, & (0, \ldots, 0, e_{n_i}^i, 0, \ldots, 0), \\
\vdots & \vdots & \vdots \\
(0, \ldots, 0, e_1^p), & \ldots, & (0, \ldots, 0, e_{n_p}^p)
\end{array}$$

form a basis of $E_1 \times \cdots \times E_p$, and so

$$\dim(E_1 \times \cdots \times E_p) = \dim(E_1) + \cdots + \dim(E_p).$$

Let us now consider a vector space E and p subspaces $U_1, \ldots, U_p$ of E. We have a map

$$a\colon U_1 \times \cdots \times U_p \to E$$

given by

$$a(u_1, \ldots, u_p) = u_1 + \cdots + u_p,$$

with $u_i \in U_i$ for $i = 1, \ldots, p$. It is clear that this map is linear, and so its image is a subspace of E denoted by

$$U_1 + \cdots + U_p$$

and called the *sum* of the subspaces $U_1, \ldots, U_p$. By definition,

$$U_1 + \cdots + U_p = \{u_1 + \cdots + u_p \mid u_i \in U_i,\ 1 \leq i \leq p\},$$

and it is immediately verified that $U_1 + \cdots + U_p$ is the smallest subspace of E containing $U_1, \ldots, U_p$. This also implies that $U_1 + \cdots + U_p$ does not depend on the order of the factors U_i; in particular,

$$U_1 + U_2 = U_2 + U_1.$$

Definition 6.3. For any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$ of E, if the map a defined above is injective, then the sum $U_1 + \cdots + U_p$ is called a *direct sum* and it is denoted by

$$U_1 \oplus \cdots \oplus U_p.$$

The space E is the *direct sum* of the subspaces U_i if

$$E = U_1 \oplus \cdots \oplus U_p.$$

As in the case of a sum, $U_1 \oplus U_2 = U_2 \oplus U_1$.

If the map a is injective, then by Proposition 3.17 we have $\operatorname{Ker} a = \{(\underbrace{0, \ldots, 0}_{p})\}$ where each 0 is the zero vector of E, which means that if $u_i \in U_i$ for $i = 1, \ldots, p$ and if

$$u_1 + \cdots + u_p = 0,$$

then $(u_1, \ldots, u_p) = (0, \ldots, 0)$, that is, $u_1 = 0, \ldots, u_p = 0$.

Proposition 6.3. *If the map $a \colon U_1 \times \cdots \times U_p \to E$ is injective, then every $u \in U_1 + \cdots + U_p$ has a unique expression as a sum*

$$u = u_1 + \cdots + u_p,$$

with $u_i \in U_i$, for $i = 1, \ldots, p$.

Proof. If

$$u = v_1 + \cdots + v_p = w_1 + \cdots + w_p,$$

with $v_i, w_i \in U_i$, for $i = 1, \ldots, p$, then we have

$$w_1 - v_1 + \cdots + w_p - v_p = 0,$$

and since $v_i, w_i \in U_i$ and each U_i is a subspace, $w_i - v_i \in U_i$. The injectivity of a implies that $w_i - v_i = 0$, that is, $w_i = v_i$ for $i = 1, \ldots, p$, which shows the uniqueness of the decomposition of u. ☐

Proposition 6.4. *If the map $a \colon U_1 \times \cdots \times U_p \to E$ is injective, then any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent.*

Proof. To see this, assume that

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0$$

for some $\lambda_i \in \mathbb{R}$. Since $u_i \in U_i$ and U_i is a subspace, $\lambda_i u_i \in U_i$, and the injectivity of a implies that $\lambda_i u_i = 0$, for $i = 1, \ldots, p$. Since $u_i \neq 0$, we must have $\lambda_i = 0$ for $i = 1, \ldots, p$; that is, $u_1, \ldots, u_p$ with $u_i \in U_i$ and $u_i \neq 0$ are linearly independent. ☐

Observe that if a is injective, then we must have $U_i \cap U_j = (0)$ whenever $i \neq j$. However, this condition is generally not sufficient if $p \geq 3$. For example, if $E = \mathbb{R}^2$ and U_1 the line spanned by $e_1 = (1, 0)$, U_2 is the line spanned by $d = (1, 1)$, and U_3 is the line spanned by $e_2 = (0, 1)$, then $U_1 \cap U_2 = U_1 \cap U_3 = U_2 \cap U_3 = \{(0, 0)\}$, but $U_1 + U_2 = U_1 + U_3 = U_2 + U_3 = \mathbb{R}^2$, so $U_1 + U_2 + U_3$ is not a direct sum. For example, d is expressed in two different ways as

$$d = (1, 1) = (1, 0) + (0, 1) = e_1 + e_2.$$

See Figure 6.1.

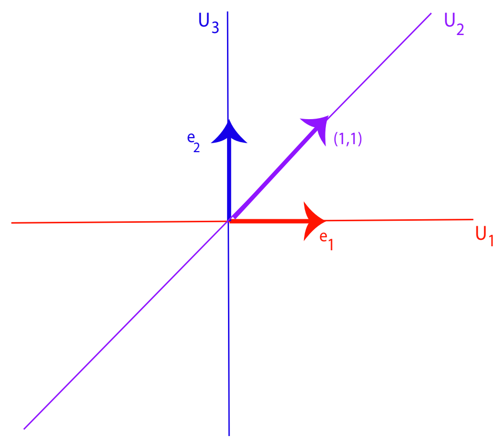


Figure 6.1: The linear subspaces U_1, U_2, and U_3 illustrated as lines in $\mathbb{R}^2$.

As in the case of a sum, $U_1 \oplus U_2 = U_2 \oplus U_1$. *Observe that when the map a is injective, then it is a linear isomorphism between $U_1 \times \cdots \times U_p$ and $U_1 \oplus \cdots \oplus U_p$.* The difference is that $U_1 \times \cdots \times U_p$ is defined even if the spaces U_i are not assumed to be subspaces of some common space.

If E is a direct sum $E = U_1 \oplus \cdots \oplus U_p$, since any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent, if we pick a basis $(u_k)_{k \in I_j}$ in U_j for $j = 1, \ldots, p$, then $(u_i)_{i \in I}$ with $I = I_1 \cup \cdots \cup I_p$ is a basis of E. Intuitively, E is split into p independent subspaces.

Conversely, given a basis $(u_i)_{i \in I}$ of E, if we partition the index set I as $I = I_1 \cup \cdots \cup I_p$, then each subfamily $(u_k)_{k \in I_j}$ spans some subspace U_j of E, and it is immediately verified that we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

Definition 6.4. Let $f\colon E \to E$ be a linear map. For any subspace U of E, if $f(U) \subseteq U$ we say that U is *invariant under f*.

Assume that E is finite-dimensional, a direct sum $E = U_1 \oplus \cdots \oplus U_p$, and that each U_j is invariant under f. If we pick a basis $(u_i)_{i \in I}$ as above with $I = I_1 \cup \cdots \cup I_p$ and with each $(u_k)_{k \in I_j}$ a basis of U_j, since each U_j is invariant under f, the image $f(u_k)$ of every basis vector u_k with $k \in I_j$ belongs to U_j, so the matrix A representing f over the basis $(u_i)_{i \in I}$ is a *block diagonal* matrix of the form

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_p \end{pmatrix},$$

with each block A_j a $d_j \times d_j$-matrix with $d_j = \dim(U_j)$ and all other entries equal to 0. If $d_j = 1$ for $j = 1, \ldots, p$, the matrix A is a diagonal matrix.

There are natural injections from each U_i to E denoted by $\mathrm{in}_i \colon U_i \to E$.

Now, if $p = 2$, it is easy to determine the kernel of the map $a \colon U_1 \times U_2 \to E$. We have

$$a(u_1, u_2) = u_1 + u_2 = 0 \quad \text{iff} \quad u_1 = -u_2,\ u_1 \in U_1, u_2 \in U_2,$$

which implies that

$$\mathrm{Ker}\, a = \{(u, -u) \mid u \in U_1 \cap U_2\}.$$

Now, $U_1 \cap U_2$ is a subspace of E and the linear map $u \mapsto (u, -u)$ is clearly an isomorphism between $U_1 \cap U_2$ and $\mathrm{Ker}\, a$, so $\mathrm{Ker}\, a$ is isomorphic to $U_1 \cap U_2$. As a consequence, we get the following result:

Proposition 6.5. *Given any vector space E and any two subspaces U_1 and U_2, the sum $U_1 + U_2$ is a direct sum iff $U_1 \cap U_2 = (0)$.*

An interesting illustration of the notion of direct sum is the decomposition of a square matrix into its symmetric part and its skew-symmetric part. Recall that an $n \times n$ matrix $A \in \mathrm{M}_n$ is *symmetric* if $A^\top = A$, *skew -symmetric* if $A^\top = -A$. It is clear that s

$$\mathbf{S}(n) = \{A \in \mathrm{M}_n \mid A^\top = A\} \quad \text{and} \quad \mathbf{Skew}(n) = \{A \in \mathrm{M}_n \mid A^\top = -A\}$$

are subspaces of M_n, and that $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$. Observe that for any matrix $A \in \mathrm{M}_n$, the matrix $H(A) = (A + A^\top)/2$ is symmetric and the matrix $S(A) = (A - A^\top)/2$ is skew-symmetric. Since

$$A = H(A) + S(A) = \frac{A + A^\top}{2} + \frac{A - A^\top}{2},$$

we see that $\mathrm{M}_n = \mathbf{S}(n) + \mathbf{Skew}(n)$, and since $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$, we have the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n).$$

Remark: The vector space $\mathbf{Skew}(n)$ of skew-symmetric matrices is also denoted by $\mathfrak{so}(n)$. It is the *Lie algebra* of the group $\mathbf{SO}(n)$.

Proposition 6.5 can be generalized to any $p \geq 2$ subspaces at the expense of notation. The proof of the following proposition is left as an exercise.

Proposition 6.6. *Given any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$, the following properties are equivalent:*

(1) The sum $U_1 + \cdots + U_p$ is a direct sum.

(2) We have

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0), \quad i = 1, \ldots, p.$$

(3) We have

$$U_i \cap \left(\sum_{j=1}^{i-1} U_j \right) = (0), \quad i = 2, \ldots, p.$$

Because of the isomorphism

$$U_1 \times \cdots \times U_p \approx U_1 \oplus \cdots \oplus U_p,$$

we have

Proposition 6.7. *If E is any vector space, for any (finite-dimensional) subspaces $U_1, \ldots,$ U_p of E, we have*

$$\dim(U_1 \oplus \cdots \oplus U_p) = \dim(U_1) + \cdots + \dim(U_p).$$

If E is a direct sum

$$E = U_1 \oplus \cdots \oplus U_p,$$

since every $u \in E$ can be written in a unique way as

$$u = u_1 + \cdots + u_p$$

with $u_i \in U_i$ for $i = 1 \ldots, p$, we can define the maps $\pi_i \colon E \to U_i$, called *projections*, by

$$\pi_i(u) = \pi_i(u_1 + \cdots + u_p) = u_i.$$

It is easy to check that these maps are linear and satisfy the following properties:

$$\pi_j \circ \pi_i = \begin{cases} \pi_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

$$\pi_1 + \cdots + \pi_p = \mathrm{id}_E.$$

For example, in the case of the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n),$$

the projection onto $\mathbf{S}(n)$ is given by

$$\pi_1(A) = H(A) = \frac{A + A^\top}{2},$$

and the projection onto $\mathbf{Skew}(n)$ is given by

$$\pi_2(A) = S(A) = \frac{A - A^\top}{2}.$$

Clearly, $H(A)+S(A) = A$, $H(H(A)) = H(A)$, $S(S(A)) = S(A)$, and $H(S(A)) = S(H(A)) = 0$.

A function f such that $f \circ f = f$ is said to be *idempotent*. Thus, the projections π_i are idempotent. Conversely, the following proposition can be shown:

Proposition 6.8. *Let E be a vector space. For any $p \geq 2$ linear maps $f_i \colon E \to E$, if*

$$f_j \circ f_i = \begin{cases} f_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$
$$f_1 + \cdots + f_p = \mathrm{id}_E,$$

then if we let $U_i = f_i(E)$, we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

We also have the following proposition characterizing idempotent linear maps whose proof is also left as an exercise.

Proposition 6.9. *For every vector space E, if $f \colon E \to E$ is an idempotent linear map, i.e., $f \circ f = f$, then we have a direct sum*

$$E = \mathrm{Ker}\, f \oplus \mathrm{Im}\, f,$$

so that f is the projection onto its image $\mathrm{Im}\, f$.

We now give the definition of a direct sum for any arbitrary nonempty index set I. First, let us recall the notion of the product of a family $(E_i)_{i\in I}$. Given a family of sets $(E_i)_{i\in I}$, its product $\prod_{i\in I} E_i$, is the set of all functions $f \colon I \to \bigcup_{i\in I} E_i$, such that, $f(i) \in E_i$, for every $i \in I$. It is one of the many versions of the axiom of choice, that, if $E_i \neq \emptyset$ for every $i \in I$, then $\prod_{i\in I} E_i \neq \emptyset$. A member $f \in \prod_{i\in I} E_i$, is often denoted as $(f_i)_{i\in I}$. For every $i \in I$, we have the *projection* $\pi_i \colon \prod_{i\in I} E_i \to E_i$, defined such that, $\pi_i((f_i)_{i\in I}) = f_i$. We now define direct sums.

Definition 6.5. Let I be any nonempty set, and let $(E_i)_{i\in I}$ be a family of vector spaces. The *(external) direct sum* $\coprod_{i\in I} E_i$ (or *coproduct*) of the family $(E_i)_{i\in I}$ is defined as follows:

$\coprod_{i\in I} E_i$ consists of all $f \in \prod_{i\in I} E_i$, which have finite support, and addition and multiplication by a scalar are defined as follows:

$$(f_i)_{i\in I} + (g_i)_{i\in I} = (f_i + g_i)_{i\in I},$$
$$\lambda(f_i)_{i\in I} = (\lambda f_i)_{i\in I}.$$

We also have *injection maps* $in_i \colon E_i \to \coprod_{i\in I} E_i$, defined such that, $in_i(x) = (f_i)_{i\in I}$, where $f_i = x$, and $f_j = 0$, for all $j \in (I - \{i\})$.

The following proposition is an obvious generalization of Proposition 6.1.

Proposition 6.10. *Let I be any nonempty set, let $(E_i)_{i\in I}$ be a family of vector spaces, and let G be any vector space. The direct sum $\coprod_{i\in I} E_i$ is a vector space, and for every family $(h_i)_{i\in I}$ of linear maps $h_i\colon E_i \to G$, there is a unique linear map*

$$\Big(\sum_{i\in I} h_i\Big)\colon \coprod_{i\in I} E_i \to G,$$

such that, $(\sum_{i\in I} h_i) \circ in_i = h_i$, for every $i \in I$.

Remarks:

(1) One might wonder why the direct sum $\coprod_{i\in I} E_i$ consists of familes of finite support instead of arbitrary families; in other words, why didn't we define the direct sum of the family $(E_i)_{i\in I}$ as $\prod_{i\in I} E_i$? The product space $\prod_{i\in I} E_i$ with addition and scalar multiplication defined as above is also a vector space but the problem is that any linear map $\widehat{h}\colon \prod_{i\in I} E_i \to G$ such that $\widehat{h} \circ in_i = h_i$ for all $\in I$ must be given by

$$\widehat{h}((u_i)_{\in I}) = \sum_{i\in I} h_i(u_i),$$

and if I is infinite, the sum on the right-hand side is infinite, and thus undefined! If I is finite then $\prod_{i\in I} E_i$ and $\coprod_{i\in I} E_i$ are isomorphic.

(2) When $E_i = E$, for all $i \in I$, we denote $\coprod_{i\in I} E_i$ by $E^{(I)}$. In particular, when $E_i = K$, for all $i \in I$, we find the vector space $K^{(I)}$ of Definition 3.11.

We also have the following basic proposition about injective or surjective linear maps.

Proposition 6.11. *Let E and F be vector spaces, and let $f\colon E \to F$ be a linear map. If $f\colon E \to F$ is injective, then there is a surjective linear map $r\colon F \to E$ called a retraction, such that $r \circ f = \mathrm{id}_E$. See Figure 6.2. If $f\colon E \to F$ is surjective, then there is an injective linear map $s\colon F \to E$ called a section, such that $f \circ s = \mathrm{id}_F$. See Figure 6.3.*

Proof. Let $(u_i)_{i\in I}$ be a basis of E. Since $f\colon E \to F$ is an injective linear map, by Proposition 3.18, $(f(u_i))_{i\in I}$ is linearly independent in F. By Theorem 3.7, there is a basis $(v_j)_{j\in J}$ of F, where $I \subseteq J$, and where $v_i = f(u_i)$, for all $i \in I$. By Proposition 3.18, a linear map $r\colon F \to E$ can be defined such that $r(v_i) = u_i$, for all $i \in I$, and $r(v_j) = w$ for all $j \in (J - I)$, where w is any given vector in E, say $w = 0$. Since $r(f(u_i)) = u_i$ for all $i \in I$, by Proposition 3.18, we have $r \circ f = \mathrm{id}_E$.

Now, assume that $f\colon E \to F$ is surjective. Let $(v_j)_{j\in J}$ be a basis of F. Since $f\colon E \to F$ is surjective, for every $v_j \in F$, there is some $u_j \in E$ such that $f(u_j) = v_j$. Since $(v_j)_{j\in J}$ is a basis of F, by Proposition 3.18, there is a unique linear map $s\colon F \to E$ such that $s(v_j) = u_j$. Also, since $f(s(v_j)) = v_j$, by Proposition 3.18 (again), we must have $f \circ s = \mathrm{id}_F$. □

The converse of Proposition 6.11 is obvious.

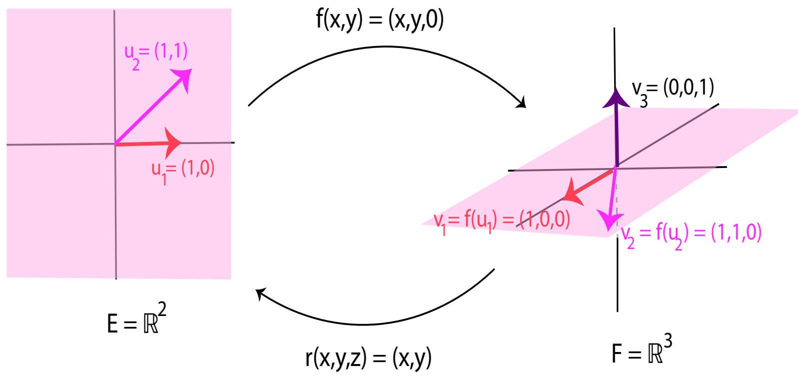


Figure 6.2: Let $f\colon E \to F$ be the injective linear map from $\mathbb{R}^2$ to $\mathbb{R}^3$ given by $f(x, y) = (x, y, 0)$. Then a surjective retraction is given by $r\colon \mathbb{R}^3 \to \mathbb{R}^2$ is given by $r(x, y, z) = (x, y)$. Observe that $r(v_1) = u_1$, $r(v_2) = u_2$, and $r(v_3) = 0$.

6.2 Matrices of Linear Maps and Multiplication by Blocks

Direct sums yield a fairly easy justification of matrix block multiplication. The key idea is that the representation of a linear map $f\colon E \to F$ over a basis $(u_1, \ldots, u_n)$ of E and a basis $(v_1, \ldots, v_m)$ of F by a matrix $A = (a_{ij})$ of scalars (in K) can be generalized to the representation of f over a direct sum decomposition $E = E_1 \oplus \cdots \oplus E_n$ of E and a direct sum decomposition $F = F_1 \oplus \cdots \oplus F_m$ of F in terms of a matrix (f_{ij}) of linear maps $f_{ij}\colon E_j \to F_i$. Futhermore, matrix multiplication of scalar matrices extends naturally to matrix multiplication of matrices of linear maps. We simply have to replace multiplication of scalars in K by the composition of linear maps.

Let E and F be two vector spaces and assume that they are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i.$$

Definition 6.6. Given any linear map $f\colon E \to F$, we define the linear maps $f_{ij}\colon E_j \to F_i$ as follows. Let $pr_i^F\colon F \to F_i$ be the projection of $F = F_1 \oplus \cdots \oplus F_m$ onto F_i. If $f_j\colon E_j \to F$ is the restriction of f to E_j, which means that for every vector $x_j \in E_j$,

$$f_j(x_j) = f(x_j),$$

then we define the map $f_{ij}\colon E_j \to F_i$ by

$$f_{ij} = pr_i^F \circ f_j,$$

so that if $x_j \in E_j$, then

$$f(x_j) = f_j(x_j) = \sum_{i=1}^{m} f_{ij}(x_j), \quad \text{with } f_{ij}(x_j) \in F_i. \tag{$\dagger_1$}$$

Observe that we are summing over the index i, which eventually corresponds to summing the entries in the jth column of the matrix representing f; see Definition 6.7.

We see that for every vector $x = x_1 + \cdots + x_n \in E$, with $x_j \in E_j$, we have

$$f(x) = \sum_{j=1}^{n} f_j(x_j) = \sum_{j=1}^{n}\sum_{i=1}^{m} f_{ij}(x_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_j).$$

Observe that conversely, given a family $(f_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ of linear maps $f_{ij}\colon E_j \to F_i$, we obtain the linear map $f\colon E \to F$ defined such that for every $x = x_1 + \cdots + x_n \in E$, with $x_j \in E_j$, we have

$$f(x) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_j).$$

As a consequence, it is easy to check that there is an isomorphism between the vector spaces

$$\mathrm{Hom}(E, F) \quad \text{and} \quad \prod_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} \mathrm{Hom}(E_j, F_i).$$

Example 6.1. Suppose that $E = E_1 \oplus E_2$ and $F = F_1 \oplus F_2 \oplus F_3$, and that we have six maps $f_{ij}\colon E_j \to F_i$ for $i = 1, 2, 3$ and $j = 1, 2$. For any $x = x_1 + x_2$, with $x_1 \in E_1$ and $x_2 \in E_2$, we have

$$\begin{aligned} y_1 &= f_{11}(x_1) + f_{12}(x_2) \in F_1 \\ y_2 &= f_{21}(x_1) + f_{22}(x_2) \in F_2 \\ y_3 &= f_{31}(x_1) + f_{32}(x_2) \in F_3, \end{aligned}$$

$$\begin{aligned} f_1(x_1) &= f_{11}(x_1) + f_{21}(x_1) + f_{31}(x_1) \\ f_2(x_2) &= f_{12}(x_2) + f_{22}(x_2) + f_{32}(x_2), \end{aligned}$$

and

$$\begin{aligned} f(x) &= f_1(x_1) + f_2(x_2) = y_1 + y_2 + y_3 \\ &= f_{11}(x_1) + f_{12}(x_2) + f_{21}(x_1) + f_{22}(x_2) + f_{31}(x_1) + f_{32}(x_2). \end{aligned}$$

These equations suggest the matrix notation

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

In general we proceed as follows. For any $x = x_1 + \cdots + x_n$ with $x_j \in E_j$, if $y = f(x)$, since $F = F_1 \oplus \cdots \oplus F_m$, the vector $y \in F$ has a unique decomposition $y = y_1 + \cdots + y_m$ with $y_i \in F_i$, and since $f_{ij} \colon E_j \to F_i$, we have $\sum_{j=1}^n f_{ij}(x_j) \in F_i$, so $\sum_{j=1}^n f_{ij}(x_j) \in F_i$ is the ith component of $f(x)$ over the direct sum $F = F_1 \oplus \cdots \oplus F_m$; equivalently

$$pr_i^F(f(x)) = \sum_{j=1}^n f_{ij}(x_j), \quad 1 \leq i \leq m.$$

Consequently, we have

$$y_i = \sum_{j=1}^n f_{ij}(x_j), \quad 1 \leq i \leq m. \tag{$\dagger_2$}$$

This time we are summing over the index j, which eventually corresponds to multiplying the ith row of the matrix representing f by the n-tuple $(x_1, \ldots, x_n)$; see Definition 6.7.

All this suggests a generalization of the matrix notation Ax, where A is a matrix of scalars and x is a column vector of scalars, namely to write

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} f_{1\,1} & \ldots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \ldots & f_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \tag{$\dagger_3$}$$

which is an abbreviation for the m equations

$$y_i = \sum_{j=1}^n f_{ij}(x_j), \quad i = 1, \ldots, m.$$

The interpretation of the multiplication of an $m \times n$ matrix of linear maps f_{ij} by a column vector of n vectors $x_j \in E_j$ is to apply each f_{ij} to x_j to obtain $f_{ij}(x_j)$ and to sum over the index j to obtain the ith output vector. This is the generalization of multiplying the scalar a_{ij} by the scalar x_j. Also note that the jth column of the matrix (f_{ij}) consists of the maps $(f_{1j}, \ldots, f_{mj})$ such that $(f_{1j}(x_j), \ldots, f_{mj}(x_j))$ are the components of $f(x_j) = f_j(x_j)$ over the direct sum $F = F_1 \oplus \cdots \oplus F_m$.

In the special case in which each E_j and each F_i is one-dimensional, this is equivalent to choosing a basis $(u_1, \ldots, u_n)$ in E so that E_j is the one-dimensional subspace $E_j = Ku_j$, and a basis $(v_1, \ldots, v_m)$ in F_j so that F_i is the one-dimensional subspace $F_i = Kv_i$. In this case every vector $x \in E$ is expressed as $x = x_1u_1 + \cdots + x_nu_n$, where the $x_i \in K$ are scalars and similarly every vector $y \in F$ is expressed as $y = y_1v_1 + \cdots + y_mv_m$, where the $y_i \in K$ are scalars. Each linear map $f_{ij} \colon E_j \to F_i$ is a map between the one-dimensional spaces Ku_j and Kv_i, so it is of the form $f_{ij}(x_ju_j) = a_{ij}x_jv_i$, with $x_j \in K$, and so the matrix (f_{ij}) of linear maps f_{ij} is in one-to-one correspondence with the matrix (a_{ij}) of scalars in K, and Equation ($\dagger_3$) where the x_j and y_i are vectors is equivalent to the same familar equation where the x_j and y_i are the scalar coordinates of x and y over the respective bases.

Definition 6.7. Let E and F be two vector spaces and assume that they are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i.$$

Given any linear map $f\colon E \to F$, if $(f_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ is the familiy of linear maps $f_{ij}\colon E_j \to F_i$ defined in Definition 6.6, the $m \times n$ matrix of linear maps

$$M(f) = \begin{pmatrix} f_{1\,1} & \dots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \dots & f_{m\,n} \end{pmatrix}$$

is called the *matrix of f with respect to the decompositions $\bigoplus_{j=1}^{n} E_j$, and $\bigoplus_{i=1}^{m} F_i$ of E and F as direct sums.*

For any $x = x_1 + \cdots + x_n \in E$ with $x_j \in E_j$ and any $y = y_1 + \cdots + y_m \in F$ with $y_i \in F_i$, we have $y = f(x)$ iff

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} f_{1\,1} & \dots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \dots & f_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

where the matrix equation above means that the system of m equations

$$y_i = \sum_{j=1}^{n} f_{ij}(x_j), \quad i = 1\dots, m, \tag{$\dagger$}$$

holds.

But now we can also promote matrix multiplication. Suppose we have a third space G written as a direct sum. It is more convenient to write

$$E = \bigoplus_{k=1}^{p} E_k, \quad F = \bigoplus_{j=1}^{n} F_j, \quad G = \bigoplus_{i=1}^{m} G_i.$$

Assume we also have two linear maps $f\colon E \to F$ and $g\colon F \to G$. Now we have the $n \times p$ matrix of linear maps $B = (f_{jk})$ and the $m \times n$ matrix of linear maps $A = (g_{ij})$. We would like to find the $m \times p$ matrix associated with $g \circ f$.

By definition of $f_k\colon E_k \to F$ and $f_{jk}\colon E_k \to F_j$, if $x_k \in E_k$, then

$$f_k(x_k) = f(x_k) = \sum_{j=1}^{n} f_{jk}(x_k), \quad \text{with } f_{jk}(x_k) \in F_j, \tag{$*_1$}$$

and similarly, by definition of $g_j\colon F_j \to G$ and $g_{ij}\colon F_j \to G_i$, if $y_j \in F_j$, then

$$g_j(y_j) = g(y_j) = \sum_{i=1}^{m} g_{ij}(y_j), \quad \text{with } g_{ij}(y_j) \in G_i. \tag{$*_2$}$$

If we write $h = g \circ f$, we need to find the maps $h_k \colon E_k \to G$, with

$$h_k(x_k) = h(x_k) = (g \circ f)(x_k),$$

and $h_{ik} \colon E_k \to G_i$ given by

$$h_{ik} = pr_i^G \circ h_k.$$

We have

$$\begin{aligned}
h_k(x_k) &= (g \circ f)(x_k) = g(f(x_k)) = g(f_k(x_k)) \\
&= g\Bigg(\sum_{j=1}^{n} f_{jk}(x_k)\Bigg) && \text{by } (*_1) \\
&= \sum_{j=1}^{n} g(f_{jk}(x_k)) = \sum_{j=1}^{n} g_j(f_{jk}(x_k)) && \text{since } g \text{ is linear} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{m} g_{ij}(f_{jk}(x_k)) = \sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij}(f_{jk}(x_k)), && \text{by } (*_2)
\end{aligned}$$

and since $\sum_{j=1}^{n} g_{ij}(f_{jk}(x_k)) \in G_i$, we conclude that

$$h_{ik}(x_k) = \sum_{j=1}^{n} g_{ij}(f_{jk}(x_k)) = \sum_{j=1}^{n} (g_{ij} \circ f_{jk})(x_k), \qquad (*_3)$$

which can also be expressed as

$$h_{ik} = \sum_{j=1}^{n} g_{ij} \circ f_{jk}. \qquad (*_4)$$

Equation $(*_4)$ is exactly the analog of the formula for the multiplication of matrices of scalars! We just have to replace multiplication by composition. The $m \times p$ matrix of linear maps (h_{ik}) is the "product" AB of the matrices of linear maps $A = (g_{ij})$ and $B = (f_{jk})$, except that multiplication is replaced by composition.

In summary we just proved the following result.

Proposition 6.12. *Let E, F, G be three vector spaces expressed as direct sums*

$$E = \bigoplus_{k=1}^{p} E_k, \quad F = \bigoplus_{j=1}^{n} F_j, \quad G = \bigoplus_{i=1}^{m} G_i.$$

For any two linear maps $f \colon E \to F$ and $g \colon F \to G$, let $B = (f_{jk})$ be the $n \times p$ matrix of linear maps associated with f (with respect to the decomposition of E and F as direct sums) and let $A = (g_{ij})$ be the $m \times n$ matrix of linear maps associated with g (with respect to the decomposition of F and G as direct sums). Then the $m \times p$ matrix $C = (h_{ik})$ of linear maps *associated with $h = g \circ f$ (with respect to the decomposition of E and G as direct sums) is given by*

$$C = AB,$$

with

$$h_{ik} = \sum_{j=1}^{n} g_{ij} \circ f_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p.$$

We will use Proposition 6.12 to justify the rule for the block multiplication of matrices. The difficulty is mostly notational. Again suppose that E and F are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i,$$

and let $f\colon E \to F$ be a linear map. Furthermore, suppose that E has a finite basis $(u_t)_{t\in T}$, where T is the disjoint union $T = T_1 \cup \cdots \cup T_n$ of nonempty subsets T_j so that $(u_t)_{t\in T_j}$ is a basis of E_j, and similarly F has a finite basis $(v_s)_{s\in S}$, where S is the disjoint union $S = S_1 \cup \cdots \cup S_m$ of nonempty subsets S_i so that $(v_s)_{s\in S_i}$ is a basis of F_i. Let $M = |S|$, $N = |T|$, $s_i = |S_i|$, and let $t_j = |T_j|$. Since s_i is the number of elements in the basis $(v_s)_{s\in S_i}$ of F_i and $F = F_1 \oplus \cdots \oplus F_m$, we have $M = \dim(F) = s_1 + \cdots + s_m$. Similarly, since t_j is the number of elements in the basis $(u_t)_{t\in T_j}$ of E_j and $E = E_1 \oplus \cdots \oplus E_n$, we have $N = \dim(E) = t_1 + \cdots + t_n$.

Let $A = (a_{st})_{(s,t)\in S\times T}$ be the (ordinary) $M \times N$ matrix of scalars (in K) representing f with respect to the basis $(u_t)_{t\in T}$ of E and the basis $(v_s)_{s\in S}$ of F with $M = r_1 + \cdots + r_m$ and $N = s_1 + \cdots + s_n$, which means that for any $t \in T$, the tth column of A consists of the components a_{st} of $f(u_t)$ over the basis $(v_s)_{s\in S}$, as in the beginning of Section 4.1.

For any i and any j such that $1 \leq i \leq m$ and $1 \leq j \leq n$, we can form the $s_i \times t_j$ matrix A_{S_i,T_j} obtained by deleting all rows in A of index $s \notin S_i$ and all columns in A of index $t \notin T_j$. The matrix A_{S_i,T_j} is the indexed family $(a_{st})_{(s,t)\in S_i\times T_j}$, as explained at the beginning of Section 4.1.

Observe that the matrix A_{S_i,T_j} is actually the matrix representing the linear map $f_{ij}\colon E_j \to F_i$ of Definition 6.7 with respect to the basis $(u_t)_{t\in T_j}$ of E_j and the basis $(v_s)_{s\in S_i}$ of F_i, in the sense that for any $t \in T_j$, the tth column of A_{S_i,T_j} consists of the components a_{st} of $f_{ij}(u_t)$ over the basis $(v_s)_{s\in S_i}$.

Definition 6.8. Given an $M \times N$ matrix A (with entries in K), we define the $m \times n$ matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ whose entry A_{ij} is the matrix $A_{ij} = A_{S_i,T_j}$, $1 \leq i \leq m$, $1 \leq j \leq n$, and we call it the *block matrix of A associated with the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$*. The matrix A_{S_i,T_j} is an $s_i \times t_j$ matrix called the (i,j)th *block* of this block matrix.

Here we run into a notational dilemma which does not seem to be addressed in the literature. Horn and Johnson [95] (Section 0.7) define partitioned matrices as we do, but they do not propose a notation for block matrices.

The problem is that the block matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ is *not* equal to the original matrix A. First of all, the block matrix is $m\times n$ and its entries are matrices, but the matrix A is $M\times N$ and its entries are scalars. But even if we think of the block matrix as an $M\times N$ matrix of scalars, some rows and some columns of the original matrix A may have been *permuted* due to the choice of the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$; see Example 6.3.

We propose to denote the block matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ by $[A]$. This is not entirely satisfactory since all information about the partitions of S and T are lost, but at least this allows us to distinguish between A and a block matrix arising from A.

To be completely rigorous we may proceed as follows. Let $[m] = \{1, \ldots, m\}$ and $[n] = \{1, \ldots, n\}$.

Definition 6.9. For any two finite sets of indices S and T, an $S\times T$ *matrix* A is an $S\times T$-indexed family with values in K, that is, a function

$$A\colon S\times T\to K.$$

Denote the space of $S\times T$ matrices with entries in K by $\mathrm{M}_{S,T}(K)$.

An $S\times T$ matrix A is an $S\times T$-indexed family $(a_{st})_{(s,t)\in S\times T}$, but the standard representation of a matrix by a rectangular array of scalars is not quite correct because it assumes that the rows are indexed by indices in the "canonical index set" $[m]$ and that the columns are indexed by indices in the "canonical index set" $[n]$. Also the index sets need not be ordered, but in practice they are, so if $S = \{s_1, \ldots, s_m\}$ and $T = \{t_1, \ldots, t_n\}$, we denote an $S\times T$ matrix A by the rectangular array

$$A = \begin{pmatrix} a_{s_1t_1} & \cdots & a_{s_1t_n} \\ \vdots & \ddots & \vdots \\ a_{s_mt_1} & \cdots & a_{s_mt_n} \end{pmatrix}.$$

Even if the index sets are not ordered, the product of an $R\times S$ matrix A and of an $S\times T$ matrix B is well defined and $C = AB$ is an $R\times T$ matrix (where R, S, T are finite index sets); see Proposition 6.13.

Then an $m\times n$ block matrix X induced by two partitions $S = S_1\cup\cdots\cup S_m$ and $T = T_1\cup\cdots\cup T_n$ is an $[m]\times[n]$-indexed family

$$X\colon [m]\times[n]\to \prod_{(i,j)\in[m]\times[n]} \mathrm{M}_{S_i,T_j}(K),$$

such that $X(i,j)\in \mathrm{M}_{S_i,T_j}(K)$, which means that $X(i,j)$ is an $S_i\times T_j$ matrix with entries in K. The map X also defines the $M\times N$ matrix $A = (a_{st})_{s\in S, t\in T}$, with

$$a_{st} = X(i,j)_{st},$$

for any $s \in S_i$ and any $j \in T_j$, so in fact $X = [A]$ and $X(i,j) = A_{S_i,T_j}$. But remember that we abbreviate $X(i,j)$ as X_{ij}, so the (i,j)th entry in the block matrix $[A]$ of A associated with the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ should be denoted by $[A]_{ij}$. To minimize notation we will use the simpler notation A_{ij}. Schematically we represent the block matrix $[A]$ as

$$[A] = \begin{pmatrix} A_{S_1,T_1} & \cdots & A_{S_1,T_n} \\ \vdots & \ddots & \vdots \\ A_{S_m,T_1} & \cdots & A_{S_m,T_n} \end{pmatrix} \quad \text{or simply as} \quad [A] = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}.$$

In the simplified notation we lose the information about the index sets of the blocks.

Remark: It is easy to check that the set of $m \times n$ block matrices induced by two partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ is a vector space. In fact, it is isomorphic to the direct sum

$$\bigoplus_{(i,j)\in[m]\times[n]} \mathrm{M}_{S_i,T_j}(K).$$

Addition and rescaling are performed blockwise.

Example 6.2. Let $S = \{1,2,3,4,5,6\}$, with $S_1 = \{1,2\}$, $S_2 = \{3\}$, $S_3 = \{4,5,6\}$, and $T = \{1,2,3,4,5\}$, with $T_1 = \{1,2\}$, $T_2 = \{3,4\}$, $T_3 = \{5\}$, and Then $s_1 = 2$, $s_2 = 1$, $s_3 = 3$ and $t_1 = 2, t_2 = 2, t_3 = 1$. The original matrix is a 6×5 matrix $A = (a_{ij})$. Schematically we obtain a 3×3 matrix of nine blocks. where A_{11}, A_{12}, A_{13} are respectively 2×2, 2×2 and 2×1, A_{21}, A_{22}, A_{23} are respectively 1×2, 1×2 and 1×1, and A_{31}, A_{32}, A_{33} are respectively 3×2, 3×2 and 3×1, as illustrated below.

$$[A] = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} & \begin{bmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{bmatrix} & \begin{bmatrix} a_{15} \\ a_{25} \end{bmatrix} \\ \begin{bmatrix} a_{31} & a_{32} \end{bmatrix} & \begin{bmatrix} a_{33} & a_{34} \end{bmatrix} & \begin{bmatrix} a_{35} \end{bmatrix} \\ \begin{bmatrix} a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \end{bmatrix} & \begin{bmatrix} a_{43} & a_{44} \\ a_{53} & a_{54} \\ a_{63} & a_{64} \end{bmatrix} & \begin{bmatrix} a_{45} \\ a_{55} \\ a_{65} \end{bmatrix} \end{pmatrix}.$$

Technically, the blocks are obtained from A in terms of the subsets S_i, T_j. For example,

$$A_{12} = A_{\{1,2\},\{3,4\}} = \begin{bmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{bmatrix}.$$

Example 6.3. Let $S = \{1,2,3\}$, with $S_1 = \{1,3\}$, $S_2 = \{2\}$, and $T = \{1,2,3\}$, with $T_1 = \{1,3\}$, $T_2 = \{2\}$. Then $s_1 = 2$, $s_2 = 1$, and $t_1 = 2, t_2 = 1$. The block 2×2 matrix $[A]$ associated with above partitions is

$$[A] = \begin{pmatrix} A_{\{1,3\},\{1,3\}} & A_{\{1,3\},\{2\}} \\ A_{\{2\},\{1,3\}} & A_{\{2\},\{2\}} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} & \begin{bmatrix} a_{12} \\ a_{32} \end{bmatrix} \\ \begin{bmatrix} a_{21} & a_{23} \end{bmatrix} & \begin{bmatrix} a_{22} \end{bmatrix} \end{pmatrix}.$$

Observe that $[A]$ viewed as a 3×3 scalar matrix is definitely different from

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

In practice, $S = \{1, \ldots, M\}$ and $T = \{1, \ldots, N\}$, so there are bijections $\alpha_i \colon \{1, \ldots, s_i\} \to S_i$ and $\beta_j \colon \{1, \ldots, t_j\} \to T_j$, $1 \leq i \leq m$, $1 \leq j \leq n$. Each $s_i \times t_j$ matrix A_{S_i,T_j} is considered as a submatrix of A, this is why the rows are indexed by S_i and the columns are indexed by T_j, but this matrix can also be viewed as an $s_i \times t_j$ matrix $A'_{ij} = \big((a'_{ij})_{st}\big)$ by itself, with the rows indexed by $\{1, \ldots, s_i\}$ and the columns indexed by $\{1, \ldots, t_j\}$, with

$$(a'_{ij})_{st} = a_{\alpha(s)\beta(t)}, \quad 1 \leq s \leq s_i,\ 1 \leq t \leq t_j.$$

Symbolic systems like `Matlab` have commands to construct such matrices. But be careful that to put a matrix such as A'_{ij} back into A at the correct row and column locations requires viewing this matrix as A_{S_i,T_j}. Symbolic systems like `Matlab` also have commands to assign row vectors and column vectors to specific rows or columns of a matrix. Technically, to be completely rigorous, the matrices A_{S_i,T_j} and A'_{ij} are different, even though they contain the same entries. The reason they are different is that in A_{S_i,T_j} the entries are indexed by the index sets S_i and T_j, but in A'_{ij} they are indexed by the index sets $\{1, \ldots, s_i\}$ and $\{1, \ldots, t_j\}$. This depends whether we view A_{S_i,T_j} as a submatrix of A or as a matrix on its own.

In most cases, the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ are chosen so that

$$\begin{aligned} S_i &= \{s \mid s_1 + \cdots + s_{i-1} + 1 \leq s \leq s_1 + \cdots + s_i\} \\ T_j &= \{t \mid t_1 + \cdots + t_{i-1} + 1 \leq t \leq t_1 + \cdots + t_j\}, \end{aligned}$$

with $s_i = |S_i| \geq 1$, $t_j = |T_j| \geq 1$, $1 \leq i \leq m, 1 \leq j \leq n$. For $i = 1$, we have $S_1 = \{1, \ldots, s_1\}$ and $T_1 = \{1, \ldots, t_1\}$. *This means that we partition into consecutive subsets of consecutive integers and we preserve the order of the bases.* In this case, $[A]$ can be viewed as A. But the results about block multiplication hold in the general case.

Finally we tackle block multiplication. But first we observe that the computation made in Section 4.2 can be immediately adapted to matrices indexed by arbitrary finite index sets I, J, K, not necessary of the form $\{1, \ldots, p\}$, $\{1, \ldots, n\}$, $\{1, \ldots, m\}$. We need this to deal with products of matrices occurring as blocks in other matrices, since such matrices are of the form A_{S_i,T_j}, *etc.*

We can prove immediately the following result generalizing Equation (4) proven in Section 4.2 (also see the fourth equation of Proposition 4.2).

Proposition 6.13. *Let I, J, K be any nonempty finite index sets. If the $I \times J$ matrix $A = (a_{ij})_{(i,j) \in I \times J}$ represents the linear map $g \colon F \to G$ with respect to the basis $(v_j)_{j \in J}$ of F and the basis $(w_i)_{i \in I}$ of G and if the $J \times K$ matrix $B = (b_{jk})_{(j,k) \in J \times K}$ represents the linear map $f \colon E \to F$ with respect to the basis $(u_k)_{k \in K}$ of E and the basis $(v_j)_{j \in J}$ of F, then the*

$I \times K$ matrix $C = (c_{ik})_{(i,k)\in I\times K}$ representing the linear map $g \circ f \colon E \to G$ with respect to the basis $(u_k)_{k\in K}$ of E and the basis $(w_i)_{i\in I}$ of G is given by

$$C = AB,$$

where for all $i \in I$ and all $k \in K$,

$$c_{ik} = \sum_{j\in J} a_{ij} b_{jk}.$$

Let E, F, G be three vector spaces expressed as direct sums

$$E = \bigoplus_{k=1}^{p} E_k, \quad F = \bigoplus_{j=1}^{n} F_j, \quad G = \bigoplus_{i=1}^{m} G_i,$$

and let $f \colon E \to F$ and $g \colon F \to G$ be two linear maps. Furthermore, assume that E has a finite basis $(u_t)_{t\in T}$, where T is the disjoint union $T = T_1 \cup \cdots \cup T_p$ of nonempty subsets T_k so that $(u_t)_{t\in T_k}$ is a basis of E_k, F has a finite basis $(v_s)_{s\in S}$, where S is the disjoint union $S = S_1 \cup \cdots \cup S_n$ of nonempty subsets S_j so that $(v_s)_{s\in S_j}$ is a basis of F_j, and G has a finite basis $(w_r)_{r\in R}$, where R is the disjoint union $R = R_1 \cup \cdots \cup R_m$ of nonempty subsets R_i so that $(w_r)_{r\in R_i}$ is a basis of G_i. Also let $M = |R|$, $N = |S|$, $P = |T|$, $r_i = |R_i|$, $s_j = |S_j|$, $t_k = |T_k|$, so that $M = \dim(G) = r_1 + \cdots + r_m$, $N = \dim(F) = s_1 + \cdots + s_n$, and $P = \dim(E) = t_1 + \cdots + t_p$.

Let B be the $N \times P$ matrix representing f with respect to the basis $(u_t)_{t\in T}$ of E and the basis $(v_s)_{s\in S}$ of F, let A be the $M \times N$ matrix representing g with respect to the basis $(v_s)_{s\in S}$ of F and the basis $(w_r)_{r\in R}$ of G, and let C be the $M \times P$ matrix representing $h = g \circ f$ with respect to the basis basis $(u_t)_{t\in T}$ of E and the basis $(w_r)_{r\in R}$ of G.

The matrix $[A]$ is an $m \times n$ block matrix of $r_i \times s_j$ matrices A_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$, the matrix $[B]$ is an $n \times p$ block matrix of $s_j \times t_k$ matrices B_{jk} $(1 \leq j \leq n, 1 \leq k \leq p)$, and the matrix $[C]$ is an $m \times p$ block matrix of $r_i \times t_k$ matrices C_{ik} $(1 \leq i \leq m, 1 \leq k \leq p)$. Furthermore, to be precise, $A_{ij} = A_{R_i,S_j}$, $B_{jk} = B_{S_j,T_k}$, and $C_{ik} = C_{R_i,T_k}$.

Now recall that the matrix A_{R_i,S_j} represents the linear map $g_{ij} \colon F_j \to G_i$ with respect to the basis $(v_s)_{s\in S_j}$ of F_j and the basis $(w_r)_{r\in R_i}$ of G_i, the matrix B_{S_j,T_k} represents the linear map $f_{jk} \colon E_k \to F_j$ with respect to the basis $(u_t)_{t\in T_k}$ of E_k and the basis $(v_s)_{s\in S_j}$ of F_j, and the matrix C_{R_i,T_k} represents the linear map $h_{ik} \colon E_k \to G_i$ with respect to the basis $(u_t)_{t\in T_k}$ of E_k and the basis $(w_r)_{r\in R_i}$ of G_i.

By Proposition 6.12, h_{ik} is given by the formula

$$h_{ik} = \sum_{j=1}^{n} g_{ij} \circ f_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p, \tag{$*_5$}$$

and since the matrix A_{R_i,S_j} represents $g_{ij}\colon F_j \to G_i$, the matrix B_{S_j,T_k} represents $f_{jk}\colon E_k \to F_j$, and the matrix C_{R_i,T_k} represents $h_{ik}\colon E_k \to G_i$, so $(*_5)$ implies the matrix equation

$$C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p, \tag{$*_6$}$$

establishing (when combined with Proposition 6.13) the fact that $[C] = [A][B]$, namely the product $C = AB$ of the matrices A and B can be performed by blocks, using the same product formula on matrices that is used on scalars.

We record the above fact in the following proposition.

Proposition 6.14. *Let M, N, P be any positive integers, and let $\{1, \ldots, M\} = R_1 \cup \cdots \cup R_m$, $\{1, \ldots, N\} = S_1 \cup \cdots \cup S_n$, and $\{1, \ldots, P\} = T_1 \cup \cdots \cup T_p$ be any partitions into nonempty subsets R_i, S_j, T_k, and write $r_i = |R_i|$, $s_j = |S_j|$ and $t_k = |T_k|$ ($1 \leq i \leq m, 1 \leq j \leq n, 1 \leq k \leq p$). Let A be an $M \times N$ matrix, let $[A]$ be the corresponding $m \times n$ block matrix of $r_i \times s_j$ matrices A_{ij} ($1 \leq i \leq m, 1 \leq j \leq n$), and let B be an $N \times P$ matrix and $[B]$ be the corresponding $n \times p$ block matrix of $s_j \times t_k$ matrices B_{jk} ($1 \leq j \leq n, 1 \leq k \leq p$). Then the $M \times P$ matrix $C = AB$ corresponds to an $m \times p$ block matrix $[C]$ of $r_i \times t_k$ matrices C_{ik} ($1 \leq i \leq m, 1 \leq k \leq p$), and we have*

$$[C] = [A][B],$$

which means that

$$C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p.$$

Remark: The product $A_{ij}B_{jk}$ of the blocks A_{ij} and B_{jk}, which are really the matrices A_{R_i,S_j} and B_{S_j,T_k}, can be computed using the matrices A'_{ij} and B'_{jk} (discussed after Example 6.3) that are indexed by the "canonical" index sets $\{1, \ldots, r_i\}$, $\{1, \ldots, s_j\}$ and $\{1, \ldots, t_k\}$. But after computing $A'_{ij}B'_{jk}$, we have to remember to insert it as a block in $[C]$ using the correct index sets R_i and T_k. This is easily achieved in `Matlab`.

Example 6.4. Consider the partition of the index set $R = \{1, 2, 3, 4, 5, 6\}$ given by $R_1 = \{1, 2\}$, $R_2 = \{3\}$, $R_3 = \{4, 5, 6\}$; of the index set $S = \{1, 2, 3\}$ given by $S_1 = \{1, 2\}$, $S_2 = \{3\}$; and of the index set $T = \{1, 2, 3, 4, 5, 6\}$ given by $T_1 = \{1\}$, $T_2 = \{2, 3\}$, $T_3 = \{4, 5, 6\}$. Let $[A]$ be the 3×2 block matrix

$$[A] = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \\ \quad \end{bmatrix} \\ \begin{bmatrix} \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \end{bmatrix} \\ \begin{bmatrix} \quad & \quad \\ \quad & \quad \\ \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} \end{pmatrix}$$

where A_{11}, A_{12} are 2×2, 2×1; A_{21}, A_{22} are 1×2, 1×1; and A_{31}, A_{32} are 3×2, 3×1, and $[B]$ be the 2×3 block matrix

$$[B] = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \\ \end{bmatrix} & \begin{bmatrix} & \\ & \end{bmatrix} & \begin{bmatrix} & & \\ & & \end{bmatrix} \\ \begin{bmatrix} \end{bmatrix} & \begin{bmatrix} & \end{bmatrix} & \begin{bmatrix} & & \end{bmatrix} \end{pmatrix},$$

where B_{11}, B_{12}, B_{13} are 2×1, 2×2, 2×3; and B_{21}, B_{22}, B_{23} are 1×1, 1×2, 1×3. Then $[C] = [A][B]$ is the 3×3 block matrix

$$[C] = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \\ \end{bmatrix} & \begin{bmatrix} & \\ & \end{bmatrix} & \begin{bmatrix} & & \\ & & \end{bmatrix} \\ \begin{bmatrix} \end{bmatrix} & \begin{bmatrix} & \end{bmatrix} & \begin{bmatrix} & & \end{bmatrix} \\ \begin{bmatrix} \\ \\ \end{bmatrix} & \begin{bmatrix} & \\ & \\ & \end{bmatrix} & \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \end{pmatrix},$$

where C_{11}, C_{12}, C_{13} are 2×1, 2×2, 2×3; C_{21}, C_{22}, C_{23} are 1×1, 1×2, 1×3; and C_{31}, C_{32}, C_{33} are 3×1, 3×2, 3×3. For example,

$$C_{32} = A_{31}B_{12} + A_{32}B_{22}.$$

Example 6.5. This example illustrates some of the subtleties having to do with the partitioning of the index sets. Consider the 1×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix}$$

and the 3×2 matrix

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

Consider the partition of the index set $R = \{1\}$ given by $R_1 = \{1\}$; of the index set $S = \{1, 2, 3\}$ given by $S_1 = \{1, 3\}$, $S_2 = \{2\}$; and of the index set $T = \{1, 2\}$ given by $T_1 = \{2\}$, $T_2 = \{1\}$. The corresponding block matrices are the 1×2 block matrix

$$[A] = \begin{pmatrix} A_{\{1\},\{1,3\}} & A_{\{1\},\{2\}} \end{pmatrix} = \left(\begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \quad \begin{bmatrix} a_{12} \end{bmatrix} \right),$$

and the 2×2 block matrix

$$[B] = \begin{pmatrix} B_{\{1,3\},\{2\}} & B_{\{1,3\},\{1\}} \\ B_{\{2\},\{2\}} & B_{\{2\},\{1\}} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} & \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} \\ \begin{bmatrix} b_{22} \end{bmatrix} & \begin{bmatrix} b_{21} \end{bmatrix} \end{pmatrix}.$$

The product of the 1×2 block matrix $[A]$ and the 2×2 block matrix $[B]$ is the 1×2 block matrix $[C]$ given by

$$\begin{aligned}
[C] = [A][B] &= \begin{pmatrix} \begin{bmatrix} a_{11} & a_{13} \end{bmatrix} & \begin{bmatrix} a_{12} \end{bmatrix} \end{pmatrix} \begin{pmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} & \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} \\ \begin{bmatrix} b_{22} \end{bmatrix} & \begin{bmatrix} b_{21} \end{bmatrix} \end{pmatrix} \\
&= \left(\begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} + \begin{bmatrix} a_{12} \end{bmatrix} \begin{bmatrix} b_{22} \end{bmatrix} \quad \begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} + \begin{bmatrix} a_{12} \end{bmatrix} \begin{bmatrix} b_{21} \end{bmatrix} \right) \\
&= \begin{pmatrix} \begin{bmatrix} a_{11}b_{12} + a_{13}b_{32} + a_{12}b_{22} \end{bmatrix} & \begin{bmatrix} a_{11}b_{11} + a_{13}b_{31} + a_{12}b_{21} \end{bmatrix} \end{pmatrix} \\
&= \begin{pmatrix} \begin{bmatrix} a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \end{bmatrix} & \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \end{bmatrix} \end{pmatrix}.
\end{aligned}$$

The block matrix $[C]$ is obtained from the 1×2 matrix $C = AB$ using the partitions of $R = \{1\}$ given by $R_1 = \{1\}$ and of $T = \{1, 2\}$ given by $T_1 = \{2\}$, $T_2 = \{1\}$, so

$$[C] = \begin{pmatrix} C_{\{1\},\{2\}} & C_{\{1\},\{1\}} \end{pmatrix},$$

which means that $[C]$ is obtained from C by permuting its two columns. Since

$$\begin{aligned}
C = AB &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \\
&= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \end{pmatrix},
\end{aligned}$$

we have confirmed that $[C]$ is correct.

Example 6.6. Matrix block multiplication is a very effective method to prove that if an upper-triangular matrix A is invertible, then its inverse is also upper-triangular. We proceed by induction on the dimensiopn n of A. If $n = 1$, then $A = (a)$, where a is a scalar, so A is invertible iff $a \neq 0$, and $A^{-1} = (a^{-1})$, which is trivially upper-triangular. For the induction step we can write an $(n+1) \times (n+1)$ upper triangular matrix A in block form as

$$A = \begin{pmatrix} T & U \\ 0_{1,n} & \alpha \end{pmatrix},$$

where T is an $n \times n$ upper triangular matrix, U is an $n \times 1$ matrix and $\alpha \in \mathbb{R}$. Assume that A is invertible and let B be its inverse, written in block form as

$$B = \begin{pmatrix} C & V \\ W & \beta \end{pmatrix},$$

where C is an $n \times n$ matrix, V is an $n \times 1$ matrix, W is a $1 \times n$ matrix, and $\beta \in \mathbb{R}$. Since B is the inverse of A, we have $AB = I_{n+1}$, which yields

$$\begin{pmatrix} T & U \\ 0_{1,n-1} & \alpha \end{pmatrix} \begin{pmatrix} C & V \\ W & \beta \end{pmatrix} = \begin{pmatrix} I_n & 0_{n,1} \\ 0_{1,n} & 1 \end{pmatrix}.$$

By block multiplication we get

$$\begin{aligned} TC + UW &= I_n \\ TV + \beta U &= 0_{n,1} \\ \alpha W &= 0_{1,n} \\ \alpha\beta &= 1. \end{aligned}$$

From the above equations we deduce that $\alpha, \beta \neq 0$ and $\beta = \alpha^{-1}$. Since $\alpha \neq 0$, the equation $\alpha W = 0_{1,n}$ yields $W = 0_{1,n}$, and so

$$TC = I_n, \quad TV + \beta U = 0_{n,1}.$$

It follows that T is invertible and C is its inverse, and since T is upper triangular, by the induction hypothesis, C is also upper triangular.

The above argument can be easily modified to prove that if A is invertible, then its diagonal entries are nonzero.

We are now ready to prove a very crucial result relating the rank and the dimension of the kernel of a linear map.

6.3 The Rank-Nullity Theorem; Grassmann's Relation

We begin with the following fundamental proposition.

Proposition 6.15. *Let E, F and G, be three vector spaces, $f\colon E \to F$ an injective linear map, $g\colon F \to G$ a surjective linear map, and assume that $\mathrm{Im}\, f = \mathrm{Ker}\, g$. Then, the following properties hold. (a) For any section $s\colon G \to F$ of g, we have $F = \mathrm{Ker}\, g \oplus \mathrm{Im}\, s$, and the linear map $f + s\colon E \oplus G \to F$ is an isomorphism.*[1]

(b) For any retraction $r\colon F \to E$ of f, we have $F = \mathrm{Im}\, f \oplus \mathrm{Ker}\, r$.[2]

$$\begin{tikzcd} E \arrow[r, shift left, "f"] & F \arrow[l, shift left, "r"] \arrow[r, shift left, "g"] & G \arrow[l, shift left, "s"] \end{tikzcd}$$

Proof. (a) Since $s\colon G \to F$ is a section of g, we have $g \circ s = \mathrm{id}_G$, and for every $u \in F$,

$$g(u - s(g(u))) = g(u) - g(s(g(u))) = g(u) - g(u) = 0.$$

Thus, $u - s(g(u)) \in \mathrm{Ker}\, g$, and we have $F = \mathrm{Ker}\, g + \mathrm{Im}\, s$. On the other hand, if $u \in \mathrm{Ker}\, g \cap \mathrm{Im}\, s$, then $u = s(v)$ for some $v \in G$ because $u \in \mathrm{Im}\, s$, $g(u) = 0$ because $u \in \mathrm{Ker}\, g$, and so,

$$g(u) = g(s(v)) = v = 0,$$

[1] The existence of a section $s\colon G \to F$ of g follows from Proposition 6.11.

[2] The existence of a retraction $r\colon F \to E$ of f follows from Proposition 6.11.

because $g \circ s = \mathrm{id}_G$, which shows that $u = s(v) = 0$. Thus, $F = \operatorname{Ker} g \oplus \operatorname{Im} s$, and since by assumption, $\operatorname{Im} f = \operatorname{Ker} g$, we have $F = \operatorname{Im} f \oplus \operatorname{Im} s$. But then, since f and s are injective, $f + s\colon E \oplus G \to F$ is an isomorphism. The proof of (b) is very similar. $\square$

Note that we can choose a retraction $r\colon F \to E$ so that $\operatorname{Ker} r = \operatorname{Im} s$, since $F = \operatorname{Ker} g \oplus \operatorname{Im} s = \operatorname{Im} f \oplus \operatorname{Im} s$ and f is injective so we can set $r \equiv 0$ on $\operatorname{Im} s$.

Given a sequence of linear maps $E \xrightarrow{f} F \xrightarrow{g} G$, when $\operatorname{Im} f = \operatorname{Ker} g$, we say that the sequence $E \xrightarrow{f} F \xrightarrow{g} G$ is *exact at F*. If in addition to being exact at F, f is injective and g is surjective, we say that we have a *short exact sequence*, and this is denoted as

$$0 \longrightarrow E \xrightarrow{f} F \xrightarrow{g} G \longrightarrow 0.$$

The property of a short exact sequence given by Proposition 6.15 is often described by saying that $0 \longrightarrow E \xrightarrow{f} F \xrightarrow{g} G \longrightarrow 0$ is a (short) *split exact sequence*.

As a corollary of Proposition 6.15, we have the following result which shows that given a linear map $f\colon E \to F$, its domain E is the direct sum of its kernel $\operatorname{Ker} f$ with some isomorphic copy of its image $\operatorname{Im} f$.

Theorem 6.16. *(Rank-nullity theorem) Let E and F be vector spaces, and let $f\colon E \to F$ be a linear map. Then, E is isomorphic to $\operatorname{Ker} f \oplus \operatorname{Im} f$, and thus,*

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f).$$

See Figure 6.3.

Proof. Consider

$$\operatorname{Ker} f \xrightarrow{i} E \xrightarrow{f'} \operatorname{Im} f,$$

where $\operatorname{Ker} f \xrightarrow{i} E$ is the inclusion map, and $E \xrightarrow{f'} \operatorname{Im} f$ is the surjection associated with $E \xrightarrow{f} F$. Then, we apply Proposition 6.15 to any section $\operatorname{Im} f \xrightarrow{s} E$ of f' to get an isomorphism between E and $\operatorname{Ker} f \oplus \operatorname{Im} f$, and Proposition 6.7, to get $\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f)$. $\square$

Definition 6.10. The dimension $\dim(\operatorname{Ker} f)$ of the kernel of a linear map f is called the *nullity* of f.

We now derive some important results using Theorem 6.16.

Proposition 6.17. *Given a vector space E, if U and V are any two subspaces of E, then*

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

an equation known as Grassmann's relation.

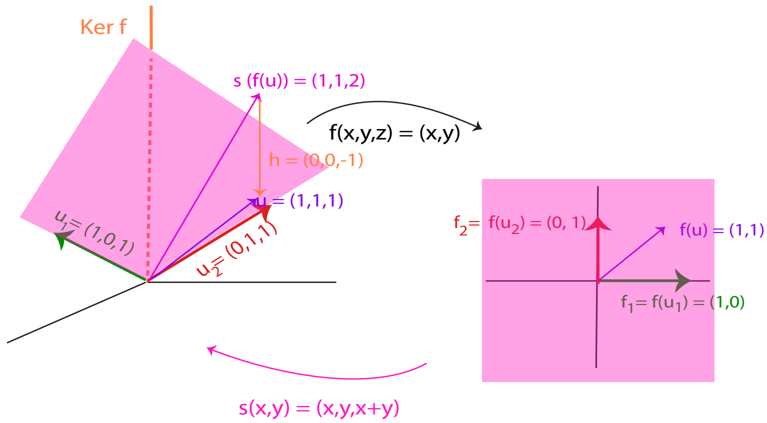


Figure 6.3: Let $f\colon E \to F$ be the linear map from $\mathbb{R}^3$ to $\mathbb{R}^2$ given by $f(x, y, z) = (x, y)$. Then $s\colon \mathbb{R}^2 \to \mathbb{R}^3$ is given by $s(x, y) = (x, y, x + y)$ and maps the pink $\mathbb{R}^2$ isomorphically onto the slanted pink plane of $\mathbb{R}^3$ whose equation is $-x - y + z = 0$. Theorem 6.16 shows that $\mathbb{R}^3$ is the direct sum of the plane $-x - y + z = 0$ and the kernel of f which the orange z-axis.

Proof. Recall that $U + V$ is the image of the linear map

$$a\colon U \times V \to E$$

given by

$$a(u, v) = u + v,$$

and that we proved earlier that the kernel $\mathrm{Ker}\, a$ of a is isomorphic to $U \cap V$. By Theorem 6.16,

$$\dim(U \times V) = \dim(\mathrm{Ker}\, a) + \dim(\mathrm{Im}\, a),$$

but $\dim(U \times V) = \dim(U) + \dim(V)$, $\dim(\mathrm{Ker}\, a) = \dim(U \cap V)$, and $\mathrm{Im}\, a = U + V$, so the Grassmann relation holds. $\square$

The Grassmann relation can be very useful to figure out whether two subspace have a nontrivial intersection in spaces of dimension > 3. For example, it is easy to see that in $\mathbb{R}^5$, there are subspaces U and V with $\dim(U) = 3$ and $\dim(V) = 2$ such that $U \cap V = (0)$; for example, let U be generated by the vectors $(1, 0, 0, 0, 0)$, $(0, 1, 0, 0, 0)$, $(0, 0, 1, 0, 0)$, and V be generated by the vectors $(0, 0, 0, 1, 0)$ and $(0, 0, 0, 0, 1)$. However, we claim that if $\dim(U) = 3$ and $\dim(V) = 3$, then $\dim(U \cap V) \geq 1$. Indeed, by the Grassmann relation, we have

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

namely

$$3 + 3 = 6 = \dim(U + V) + \dim(U \cap V),$$

and since $U + V$ is a subspace of $\mathbb{R}^5$, $\dim(U + V) \leq 5$, which implies

$$6 \leq 5 + \dim(U \cap V),$$

that is $1 \leq \dim(U \cap V)$.

As another consequence of Proposition 6.17, if U and V are two hyperplanes in a vector space of dimension n, so that $\dim(U) = n - 1$ and $\dim(V) = n - 1$, the reader should show that

$$\dim(U \cap V) \geq n - 2,$$

and so, if $U \neq V$, then

$$\dim(U \cap V) = n - 2.$$

Here is a characterization of direct sums that follows directly from Theorem 6.16.

Proposition 6.18. *If $U_1, \ldots, U_p$ are any subspaces of a finite dimensional vector space E, then*

$$\dim(U_1 + \cdots + U_p) \leq \dim(U_1) + \cdots + \dim(U_p),$$

and

$$\dim(U_1 + \cdots + U_p) = \dim(U_1) + \cdots + \dim(U_p)$$

iff the U_is form a direct sum $U_1 \oplus \cdots \oplus U_p$.

Proof. If we apply Theorem 6.16 to the linear map

$$a \colon U_1 \times \cdots \times U_p \to U_1 + \cdots + U_p$$

given by $a(u_1, \ldots, u_p) = u_1 + \cdots + u_p$, we get

$$\begin{aligned} \dim(U_1 + \cdots + U_p) &= \dim(U_1 \times \cdots \times U_p) - \dim(\operatorname{Ker} a) \\ &= \dim(U_1) + \cdots + \dim(U_p) - \dim(\operatorname{Ker} a), \end{aligned}$$

so the inequality follows. Since a is injective iff $\operatorname{Ker} a = (0)$, the U_is form a direct sum iff the second equation holds. $\square$

Another important corollary of Theorem 6.16 is the following result:

Proposition 6.19. *Let E and F be two vector spaces with the same finite dimension $\dim(E) = \dim(F) = n$. For every linear map $f \colon E \to F$, the following properties are equivalent:*

(a) f is bijective.

(b) f is surjective.

(c) f is injective.

(d) $\mathrm{Ker}\, f = (0)$.

Proof. Obviously, (a) implies (b).

If f is surjective, then $\mathrm{Im}\, f = F$, and so $\dim(\mathrm{Im}\, f) = n$. By Theorem 6.16,

$$\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f),$$

and since $\dim(E) = n$ and $\dim(\mathrm{Im}\, f) = n$, we get $\dim(\mathrm{Ker}\, f) = 0$, which means that $\mathrm{Ker}\, f = (0)$, and so f is injective (see Proposition 3.17). This proves that (b) implies (c).

If f is injective, then by Proposition 3.17, $\mathrm{Ker}\, f = (0)$, so (c) implies (d).

Finally, assume that $\mathrm{Ker}\, f = (0)$, so that $\dim(\mathrm{Ker}\, f) = 0$ and f is injective (by Proposition 3.17). By Theorem 6.16,

$$\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f),$$

and since $\dim(\mathrm{Ker}\, f) = 0$, we get

$$\dim(\mathrm{Im}\, f) = \dim(E) = \dim(F),$$

which proves that f is also surjective, and thus bijective. This proves that (d) implies (a) and concludes the proof. ☐

One should be warned that Proposition 6.19 fails in infinite dimension.

Here are a few applications of Proposition 6.19. Let A be an $n \times n$ matrix and assume that A some right inverse B, which means that B is an $n \times n$ matrix such that

$$AB = I.$$

The linear map associated with A is surjective, since for every $u \in \mathbb{R}^n$, we have $A(Bu) = u$. By Proposition 6.19, this map is bijective so B is actually the inverse of A; in particular $BA = I$.

Similarly, assume that A has a left inverse B, so that

$$BA = I.$$

This time the linear map associated with A is injective, because if $Au = 0$, then $BAu = B0 = 0$, and since $BA = I$ we get $u = 0$. Again, by Proposition 6.19, this map is bijective so B is actually the inverse of A; in particular $AB = I$.

Now assume that the linear system $Ax = b$ has some solution for every b. Then the linear map associated with A is surjective and by Proposition 6.19, A is invertible.

Finally assume that the linear system $Ax = b$ has at most one solution for every b. Then the linear map associated with A is injective and by Proposition 6.19, A is invertible.

The following Proposition will also be useful.

Proposition 6.20. *Let E be a vector space. If $E = U \oplus V$ and $E = U \oplus W$, then there is an isomorphism $f\colon V \to W$ between V and W.*

Proof. Let R be the relation between V and W, defined such that

$$\langle v, w\rangle \in R \quad \text{iff} \quad w - v \in U.$$

We claim that R is a functional relation that defines a linear isomorphism $f\colon V \to W$ between V and W, where $f(v) = w$ iff $\langle v, w\rangle \in R$ (R is the graph of f). If $w - v \in U$ and $w' - v \in U$, then $w' - w \in U$, and since $U \oplus W$ is a direct sum, $U \cap W = (0)$, and thus $w' - w = 0$, that is $w' = w$. Thus, R is functional. Similarly, if $w - v \in U$ and $w - v' \in U$, then $v' - v \in U$, and since $U \oplus V$ is a direct sum, $U \cap V = (0)$, and $v' = v$. Thus, f is injective. Since $E = U \oplus V$, for every $w \in W$, there exists a unique pair $\langle u, v\rangle \in U \times V$, such that $w = u + v$. Then, $w - v \in U$, and f is surjective. We also need to verify that f is linear. If

$$w - v = u$$

and

$$w' - v' = u',$$

where $u, u' \in U$, then, we have

$$(w + w') - (v + v') = (u + u'),$$

where $u + u' \in U$. Similarly, if

$$w - v = u$$

where $u \in U$, then we have

$$\lambda w - \lambda v = \lambda u,$$

where $\lambda u \in U$. Thus, f is linear. ☐

Given a vector space E and any subspace U of E, Proposition 6.20 shows that the dimension of any subspace V such that $E = U \oplus V$ depends only on U. We call $\dim(V)$ the *codimension* of U, and we denote it by $\mathrm{codim}(U)$. A subspace U of codimension 1 is called a *hyperplane*.

The notion of rank of a linear map or of a matrix is an important one, both theoretically and practically, since it is the key to the solvability of linear equations. Recall from Definition 3.19 that the *rank* $\mathrm{rk}(f)$ of a linear map $f\colon E \to F$ is the dimension $\dim(\mathrm{Im}\, f)$ of the image subspace $\mathrm{Im}\, f$ of F.

We have the following simple proposition.

Proposition 6.21. *Given a linear map $f\colon E \to F$, the following properties hold:*

(i) $\mathrm{rk}(f) = \mathrm{codim}(\mathrm{Ker}\, f)$.

(ii) $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$.

(iii) $\mathrm{rk}(f) \leq \min(\dim(E), \dim(F))$.

Proof. Since by Proposition 6.16, $\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f)$, and by definition, $\mathrm{rk}(f) = \dim(\mathrm{Im}\, f)$, we have $\mathrm{rk}(f) = \mathrm{codim}(\mathrm{Ker}\, f)$. Since $\mathrm{rk}(f) = \dim(\mathrm{Im}\, f)$, (ii) follows from $\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f)$. As for (iii), since $\mathrm{Im}\, f$ is a subspace of F, we have $\mathrm{rk}(f) \leq \dim(F)$, and since $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$, we have $\mathrm{rk}(f) \leq \dim(E)$. $\square$

The rank of a matrix is defined as follows.

Definition 6.11. Given a $m \times n$-matrix $A = (a_{i\,j})$ over the field K, the *rank* $\mathrm{rk}(A)$ of the matrix A is the maximum number of linearly independent columns of A (viewed as vectors in K^m).

In view of Proposition 3.8, the rank of a matrix A is the dimension of the subspace of K^m generated by the columns of A. Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis of E, and $(v_1, \ldots, v_m)$ a basis of F. Let $f \colon E \to F$ be a linear map, and let $M(f)$ be its matrix w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$. Since the rank $\mathrm{rk}(f)$ of f is the dimension of $\mathrm{Im}\, f$, which is generated by $(f(u_1), \ldots, f(u_n))$, the rank of f is the maximum number of linearly independent vectors in $(f(u_1), \ldots, f(u_n))$, which is equal to the number of linearly independent columns of $M(f)$, since F and K^m are isomorphic. Thus, we have $\mathrm{rk}(f) = \mathrm{rk}(M(f))$, for every matrix representing f.

We will see later, using duality, that the rank of a matrix A is also equal to the maximal number of linearly independent rows of A.

If U is a hyperplane, then $E = U \oplus V$ for some subspace V of dimension 1. However, a subspace V of dimension 1 is generated by any nonzero vector $v \in V$, and thus we denote V by Kv, and we write $E = U \oplus Kv$. Clearly, $v \notin U$. Conversely, let $x \in E$ be a vector such that $x \notin U$ (and thus, $x \neq 0$). We claim that $E = U \oplus Kx$. Indeed, since U is a hyperplane, we have $E = U \oplus Kv$ for some $v \notin U$ (with $v \neq 0$). Then, $x \in E$ can be written in a unique way as $x = u + \lambda v$, where $u \in U$, and since $x \notin U$, we must have $\lambda \neq 0$, and thus, $v = -\lambda^{-1}u + \lambda^{-1}x$. Since $E = U \oplus Kv$, this shows that $E = U + Kx$. Since $x \notin U$, we have $U \cap Kx = 0$, and thus $E = U \oplus Kx$. This argument shows that a hyperplane is a maximal proper subspace H of E.

Theorem 6.16 also yields a characterization of hyperplanes in terms of linear forms. Recall that given a vector space E, a *hyperplane* H in E is subspace of codimension 1, which means that there is a one-dimensional subspace L such that

$$E = H \oplus L.$$

Proposition 6.22. *Given a nontrivial vector space E over a field K, a subspace H of E is a hyperplane iff there is a nonzero linear form $\varphi \colon E \to K$ such that*

$$H = \mathrm{Ker}\, \varphi.$$

Furthermore, if φ_1 and φ_2 are any two nonzero linear forms defining the same hyperplane H, in the sense that $H = \operatorname{Ker}\varphi_1 = \operatorname{Ker}\varphi_2$, then there is some nonzero $\alpha \in K$ such that $\varphi_2 = \alpha\varphi_1$.

Proof. First assume that $\varphi\colon E \to K$ is a nonzero linear form and that $H = \operatorname{Ker}\varphi$. Then there is a nonzero vector $u_0 \in E$ such that $\varphi(u_0) = \lambda_0 \neq 0$ for some $\lambda_0 \in K$, and so for every $\lambda \in K$, we have

$$\varphi(\lambda\lambda_0^{-1}u_0) = \lambda\lambda_0^{-1}\varphi(u_0) = \lambda\lambda_0^{-1}\lambda_0 = \lambda,$$

which means that φ is surjective onto K. It follows that in Theorem 6.16 we can define s by $s(\lambda_0) = u_0$, so the subspace $L = \operatorname{Im} s = Ku_0$ is a one-dimensional space and we have

$$E = \operatorname{Ker}\varphi \oplus L = H \oplus L,$$

so H is a hyperplane.

Conversely assume that H is a hyperplane, so that $E = H \oplus L$ where L is a subspace of dimension 1. If we pick a nonzero vector $u_0 \in L$, since L has dimension 1 and $E = H \oplus L$, every $u \in E$ can be written uniquely as $u = h + \lambda u_0$ for some $h \in H$ and some $\lambda \in K$. If we define the map $\varphi\colon E \to K$ by

$$\varphi(u + \lambda u_0) = \lambda,$$

we check immediately that φ is linear and that its kernel is H.

Assume that $H = \operatorname{Ker}\varphi_1 = \operatorname{Ker}\varphi_2$ for some nonzero linear forms φ_1 and φ_2. We just proved that $E = H \oplus Ku_0$ for some $u_0 \in E$ such that $\varphi_1(u_0) \neq 0$, and we must also have $\varphi_2(u_0) \neq 0$, since otherwise, as $H = \operatorname{Ker}\varphi_2$, the linear form φ_2 would be zero on E. Then observe that

$$\varphi_2 - \frac{\varphi_2(u_0)}{\varphi_1(u_0)}\varphi_1$$

is a linear form that vanishes on H since both φ_1 and φ_2 vanish on H, but also vanishes on Ku_0 since

$$\begin{aligned}\left(\varphi_2 - \frac{\varphi_2(u_0)}{\varphi_1(u_0)}\varphi_1\right)(\lambda u_0) &= \varphi_2(\lambda u_0) - \frac{\varphi_2(u_0)}{\varphi_1(u_0)}\varphi_1(\lambda u_0)\\ &= \lambda\varphi_2(u_0) - \lambda\frac{\varphi_2(u_0)}{\varphi_1(u_0)}\varphi_1(u_0) = \lambda\varphi_2(u_0) - \lambda\varphi_2(u_0) = 0\end{aligned}$$

for all $\lambda \in K$. Since $E = H \oplus Ku_0$, we deduce that $\varphi_2 - \alpha\varphi_2$ vanishes on E, with

$$\alpha = \frac{\varphi_2(u_0)}{\varphi_1(u_0)} \neq 0,$$

and so $\varphi_2 = \alpha\varphi_1$, as claimed. $\square$

It is immediately verified that if H is a hyperplane in E defined by a nonzero linear form φ so that $H = \operatorname{Ker} \varphi$, then for any nonzero $\alpha \in K$, the linear form $\alpha\varphi$ is a nonzero linear form that also defines H, that is, $H = \operatorname{Ker} \alpha\varphi$. This fact with the second part of Proposition 6.22 shows that a hyperplane H in E is defined by the one-dimensional subspace of the dual E^* of E consisting of all the linear forms that vanish on H (including the zero linear form). This is an instance of duality.

6.4 Summary

The main concepts and results of this chapter are listed below:

- *Direct products, sums, direct sums.*
- *Projections.*
- The fundamental equation

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f)$$

  (Proposition 6.16).
- *Grassmann's relation*

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V).$$

- Characterizations of a bijective linear map $f \colon E \to F$.
- *Rank* of a matrix.

6.5 Problems

Problem 6.1. Let V and W be two subspaces of a vector space E. Prove that if $V \cup W$ is a subspace of E, then either $V \subseteq W$ or $W \subseteq V$.

Problem 6.2. Prove that for every vector space E, if $f \colon E \to E$ is an idempotent linear map, i.e., $f \circ f = f$, then we have a direct sum

$$E = \operatorname{Ker} f \oplus \operatorname{Im} f,$$

so that f is the projection onto its image $\operatorname{Im} f$.

Problem 6.3. Let $U_1, \ldots, U_p$ be any $p \geq 2$ subspaces of some vector space E and recall that the linear map

$$a \colon U_1 \times \cdots \times U_p \to E$$

is given by

$$a(u_1, \dots, u_p) = u_1 + \cdots + u_p,$$

with $u_i \in U_i$ for $i = 1, \dots, p$.

(1) If we let $Z_i \subseteq U_1 \times \cdots \times U_p$ be given by

$$Z_i = \left\{ \left(u_1, \dots, u_{i-1}, -\sum_{j=1, j\neq i}^{p} u_j, u_{i+1}, \dots, u_p \right) \;\middle|\; \sum_{j=1, j\neq i}^{p} u_j \in U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) \right\},$$

for $i = 1, \dots, p$, then prove that

$$\operatorname{Ker} a = Z_1 = \cdots = Z_p.$$

In general, for any given i, the condition $U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0)$ does not necessarily imply that $Z_i = (0)$. Thus, let

$$Z = \left\{ \left(u_1, \dots, u_{i-1}, u_i, u_{i+1}, \dots, u_p \right) \;\middle|\; u_i = -\sum_{j=1, j\neq i}^{p} u_j,\ u_i \in U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right),\ 1 \leq i \leq p \right\}.$$

Since $\operatorname{Ker} a = Z_1 = \cdots = Z_p$, we have $Z = \operatorname{Ker} a$. Prove that if

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0) \quad 1 \leq i \leq p,$$

then $Z = \operatorname{Ker} a = (0)$.

(2) Prove that $U_1 + \cdots + U_p$ is a direct sum iff

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0) \quad 1 \leq i \leq p.$$

Problem 6.4. Assume that E is finite-dimensional, and let $f_i \colon E \to E$ be any $p \geq 2$ linear maps such that

$$f_1 + \cdots + f_p = \mathrm{id}_E.$$

Prove that the following properties are equivalent:

(1) $f_i^2 = f_i$, $1 \leq i \leq p$.

(2) $f_j \circ f_i = 0$, for all $i \neq j$, $1 \leq i, j \leq p$.

Hint. Use Problem 6.2.

Let $U_1, \ldots, U_p$ be any $p \geq 2$ subspaces of some vector space E. Prove that $U_1 + \cdots + U_p$ is a direct sum iff

$$U_i \cap \left(\sum_{j=1}^{i-1} U_j \right) = (0), \quad i = 2, \ldots, p.$$

Problem 6.5. Given any vector space E, a linear map $f\colon E \to E$ is an *involution* if $f \circ f = \mathrm{id}$.

(1) Prove that an involution f is invertible. What is its inverse?

(2) Let E_1 and E_{-1} be the subspaces of E defined as follows:

$$\begin{aligned} E_1 &= \{u \in E \mid f(u) = u\} \\ E_{-1} &= \{u \in E \mid f(u) = -u\}. \end{aligned}$$

Prove that we have a direct sum

$$E = E_1 \oplus E_{-1}.$$

Hint. For every $u \in E$, write

$$u = \frac{u + f(u)}{2} + \frac{u - f(u)}{2}.$$

(3) If E is finite-dimensional and f is an involution, prove that there is some basis of E with respect to which the matrix of f is of the form

$$I_{k,n-k} = \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix},$$

where I_k is the $k \times k$ identity matrix (similarly for I_{n-k}) and $k = \dim(E_1)$. Can you give a geometric interpretation of the action of f (especially when $k = n - 1$)?

Problem 6.6. An $n \times n$ matrix H is *upper Hessenberg* if $h_{jk} = 0$ for all (j, k) such that $j - k \geq 0$. An upper Hessenberg matrix is *unreduced* if $h_{i+1i} \neq 0$ for $i = 1, \ldots, n - 1$.

Prove that if H is a singular unreduced upper Hessenberg matrix, then $\dim(\mathrm{Ker}\,(H)) = 1$.

Problem 6.7. Let A be any $n \times k$ matrix.

(1) Prove that the $k \times k$ matrix $A^\top A$ and the matrix A have the same nullspace. Use this to prove that $\mathrm{rank}(A^\top A) = \mathrm{rank}(A)$. Similarly, prove that the $n \times n$ matrix $AA^\top$ and the matrix $A^\top$ have the same nullspace, and conclude that $\mathrm{rank}(AA^\top) = \mathrm{rank}(A^\top)$.

We will prove later that $\mathrm{rank}(A^\top) = \mathrm{rank}(A)$.

(2) Let $a_1, \ldots, a_k$ be k linearly independent vectors in $\mathbb{R}^n$ $(1 \leq k \leq n)$, and let A be the $n \times k$ matrix whose ith column is a_i. Prove that $A^\top A$ has rank k, and that it is invertible. Let $P = A(A^\top A)^{-1} A^\top$ (an $n \times n$ matrix). Prove that

$$\begin{aligned} P^2 &= P \\ P^\top &= P. \end{aligned}$$

What is the matrix P when $k = 1$?

(3) Prove that the image of P is the subspace V spanned by $a_1, \ldots, a_k$, or equivalently the set of all vectors in $\mathbb{R}^n$ of the form Ax, with $x \in \mathbb{R}^k$. Prove that the nullspace U of P is the set of vectors $u \in \mathbb{R}^n$ such that $A^\top u = 0$. Can you give a geometric interpretation of U?

Conclude that P is a projection of $\mathbb{R}^n$ onto the subspace V spanned by $a_1, \ldots, a_k$, and that

$$\mathbb{R}^n = U \oplus V.$$

Problem 6.8. A *rotation* R_θ in the plane $\mathbb{R}^2$ is given by the matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

(1) Use `Matlab` to show the action of a rotation R_θ on a simple figure such as a triangle or a rectangle, for various values of θ, including $\theta = \pi/6, \pi/4, \pi/3, \pi/2$.

(2) Prove that R_θ is invertible and that its inverse is $R_{-\theta}$.

(3) For any two rotations R_α and R_β, prove that

$$R_\beta \circ R_\alpha = R_\alpha \circ R_\beta = R_{\alpha+\beta}.$$

Use (2)-(3) to prove that the rotations in the plane form a commutative group denoted $\mathbf{SO}(2)$.

Problem 6.9. Consider the affine map $R_{\theta,(a_1,a_2)}$ in $\mathbb{R}^2$ given by

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

(1) Prove that if $\theta \neq k2\pi$, with $k \in \mathbb{Z}$, then $R_{\theta,(a_1,a_2)}$ has a unique fixed point (c_1, c_2), that is, there is a unique point (c_1, c_2) such that

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = R_{\theta,(a_1,a_2)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

and this fixed point is given by

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{2\sin(\theta/2)} \begin{pmatrix} \cos(\pi/2 - \theta/2) & -\sin(\pi/2 - \theta/2) \\ \sin(\pi/2 - \theta/2) & \cos(\pi/2 - \theta/2) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

(2) In this question we still assume that $\theta \neq k2\pi$, with $k \in \mathbb{Z}$. By translating the coordinate system with origin $(0, 0)$ to the new coordinate system with origin (c_1, c_2), which

means that if (x_1, x_2) are the coordinates with respect to the standard origin $(0, 0)$ and if (x'_1, x'_2) are the coordinates with respect to the new origin (c_1, c_2), we have

$$\begin{aligned} x_1 &= x'_1 + c_1 \\ x_2 &= x'_2 + c_2 \end{aligned}$$

and similarly for (y_1, y_2) and (y'_1, y'_2), then show that

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = R_{\theta,(a_1,a_2)} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

becomes

$$\begin{pmatrix} y'_1 \\ y'_2 \end{pmatrix} = R_\theta \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix}.$$

Conclude that with respect to the new origin (c_1, c_2), the affine map $R_{\theta,(a_1,a_2)}$ becomes the rotation R_θ. We say that $R_{\theta,(a_1,a_2)}$ is a *rotation of center* (c_1, c_2).

(3) Use `Matlab` to show the action of the affine map $R_{\theta,(a_1,a_2)}$ on a simple figure such as a triangle or a rectangle, for $\theta = \pi/3$ and various values of (a_1, a_2). Display the center (c_1, c_2) of the rotation.

What kind of transformations correspond to $\theta = k2\pi$, with $k \in \mathbb{Z}$?

(4) Prove that the inverse of $R_{\theta,(a_1,a_2)}$ is of the form $R_{-\theta,(b_1,b_2)}$, and find (b_1, b_2) in terms of θ and (a_1, a_2).

(5) Given two affine maps $R_{\alpha,(a_1,a_2)}$ and $R_{\beta,(b_1,b_2)}$, prove that

$$R_{\beta,(b_1,b_2)} \circ R_{\alpha,(a_1,a_2)} = R_{\alpha+\beta,(t_1,t_2)}$$

for some (t_1, t_2), and find (t_1, t_2) in terms of β, (a_1, a_2) and (b_1, b_2).

Even in the case where $(a_1, a_2) = (0, 0)$, prove that in general

$$R_{\beta,(b_1,b_2)} \circ R_\alpha \neq R_\alpha \circ R_{\beta,(b_1,b_2)}.$$

Use (4)-(5) to show that the affine maps of the plane defined in this problem form a nonabelian group denoted $\mathbf{SE}(2)$.

Prove that $R_{\beta,(b_1,b_2)} \circ R_{\alpha,(a_1,a_2)}$ is not a translation (possibly the identity) iff $\alpha + \beta \neq k2\pi$, for all $k \in \mathbb{Z}$. Find its center of rotation when $(a_1, a_2) = (0, 0)$.

If $\alpha + \beta = k2\pi$, then $R_{\beta,(b_1,b_2)} \circ R_{\alpha,(a_1,a_2)}$ is a pure translation. Find the translation vector of $R_{\beta,(b_1,b_2)} \circ R_{\alpha,(a_1,a_2)}$.

Problem 6.10. (Affine subspaces) A subset $\mathcal{A}$ of $\mathbb{R}^n$ is called an *affine subspace* if either $\mathcal{A} = \emptyset$, or there is some vector $a \in \mathbb{R}^n$ and some subspace U of $\mathbb{R}^n$ such that

$$\mathcal{A} = a + U = \{a + u \mid u \in U\}.$$

We define the dimension $\dim(\mathcal{A})$ of $\mathcal{A}$ as the dimension $\dim(U)$ of U.

(1) If $\mathcal{A} = a + U$, why is $a \in \mathcal{A}$?

What are affine subspaces of dimension 0? What are affine subspaces of dimension 1 (begin with $\mathbb{R}^2$)? What are affine subspaces of dimension 2 (begin with $\mathbb{R}^3$)?

Prove that any nonempty affine subspace is closed under affine combinations.

(2) Prove that if $\mathcal{A} = a + U$ is any nonempty affine subspace, then $\mathcal{A} = b + U$ for any $b \in \mathcal{A}$.

(3) Let $\mathcal{A}$ be any nonempty subset of $\mathbb{R}^n$ closed under affine combinations. For any $a \in \mathcal{A}$, prove that

$$U_a = \{x - a \in \mathbb{R}^n \mid x \in \mathcal{A}\}$$

is a (linear) subspace of $\mathbb{R}^n$ such that

$$\mathcal{A} = a + U_a.$$

Prove that U_a does not depend on the choice of $a \in \mathcal{A}$; that is, $U_a = U_b$ for all $a, b \in \mathcal{A}$. In fact, prove that

$$U_a = U = \{y - x \in \mathbb{R}^n \mid x, y \in \mathcal{A}\}, \quad \text{for all } a \in \mathcal{A},$$

and so

$$\mathcal{A} = a + U, \quad \text{for any } a \in \mathcal{A}.$$

Remark: The subspace U is called the *direction* of $\mathcal{A}$.

(4) Two nonempty affine subspaces $\mathcal{A}$ and $\mathcal{B}$ are said to be *parallel* iff they have the same direction. Prove that that if $\mathcal{A} \neq \mathcal{B}$ and $\mathcal{A}$ and $\mathcal{B}$ are parallel, then $\mathcal{A} \cap \mathcal{B} = \emptyset$.

Remark: The above shows that affine subspaces behave quite differently from linear subspaces.

Problem 6.11. (Affine frames and affine maps) For any vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$, let $\widehat{v} \in \mathbb{R}^{n+1}$ be the vector $\widehat{v} = (v_1, \ldots, v_n, 1)$. Equivalently, $\widehat{v} = (\widehat{v}_1, \ldots, \widehat{v}_{n+1}) \in \mathbb{R}^{n+1}$ is the vector defined by

$$\widehat{v}_i = \begin{cases} v_i & \text{if } 1 \leq i \leq n, \\ 1 & \text{if } i = n + 1. \end{cases}$$

(1) For any $m + 1$ vectors $(u_0, u_1, \ldots, u_m)$ with $u_i \in \mathbb{R}^n$ and $m \leq n$, prove that if the m vectors $(u_1 - u_0, \ldots, u_m - u_0)$ are linearly independent, then the $m + 1$ vectors $(\widehat{u}_0, \ldots, \widehat{u}_m)$ are linearly independent.

(2) Prove that if the $m + 1$ vectors $(\widehat{u}_0, \ldots, \widehat{u}_m)$ are linearly independent, then for any choice of i, with $0 \leq i \leq m$, the m vectors $u_j - u_i$ for $j \in \{0, \ldots, m\}$ with $j - i \neq 0$ are linearly independent.

Any $m+1$ vectors $(u_0, u_1, \ldots, u_m)$ such that the $m+1$ vectors $(\widehat{u}_0, \ldots, \widehat{u}_m)$ are linearly independent are said to be *affinely independent*.

From (1) and (2), the vector $(u_0, u_1, \ldots, u_m)$ are affinely independent iff for any any choice of i, with $0 \leq i \leq m$, the m vectors $u_j - u_i$ for $j \in \{0, \ldots, m\}$ with $j - i \neq 0$ are linearly independent. If $m = n$, we say that $n+1$ affinely independent vectors $(u_0, u_1, \ldots, u_n)$ form an *affine frame* of $\mathbb{R}^n$.

(3) if $(u_0, u_1, \ldots, u_n)$ is an affine frame of $\mathbb{R}^n$, then prove that for every vector $v \in \mathbb{R}^n$, there is a unique $(n+1)$-tuple $(\lambda_0, \lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n+1}$, with $\lambda_0 + \lambda_1 + \cdots + \lambda_n = 1$, such that

$$v = \lambda_0 u_0 + \lambda_1 u_1 + \cdots + \lambda_n u_n.$$

The scalars $(\lambda_0, \lambda_1, \ldots, \lambda_n)$ are called the *barycentric* (or *affine*) *coordinates* of v w.r.t. the affine frame $(u_0, u_1, \ldots, u_n)$.

If we write $e_i = u_i - u_0$, for $i = 1, \ldots, n$, then prove that we have

$$v = u_0 + \lambda_1 e_1 + \cdots + \lambda_n e_n,$$

and since $(e_1, \ldots, e_n)$ is a basis of $\mathbb{R}^n$ (by (1) & (2)), the n-tuple $(\lambda_1, \ldots, \lambda_n)$ consists of the standard coordinates of $v - u_0$ over the basis $(e_1, \ldots, e_n)$.

Conversely, for any vector $u_0 \in \mathbb{R}^n$ and for any basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, let $u_i = u_0 + e_i$ for $i = 1, \ldots, n$. Prove that $(u_0, u_1, \ldots, u_n)$ is an affine frame of $\mathbb{R}^n$, and for any $v \in \mathbb{R}^n$, if

$$v = u_0 + x_1 e_1 + \cdots + x_n e_n,$$

with $(x_1, \ldots, x_n) \in \mathbb{R}^n$ (unique), then

$$v = (1 - (x_1 + \cdots + x_x))u_0 + x_1 u_1 + \cdots + x_n u_n,$$

so that $(1 - (x_1 + \cdots + x_x)), x_1, \cdots, x_n)$, are the barycentric coordinates of v w.r.t. the affine frame $(u_0, u_1, \ldots, u_n)$.

The above shows that there is a one-to-one correspondence between affine frames $(u_0, \ldots, u_n)$ and pairs $(u_0, (e_1, \ldots, e_n))$, with $(e_1, \ldots, e_n)$ a basis. Given an affine frame $(u_0, \ldots, u_n)$, we obtain the basis $(e_1, \ldots, e_n)$ with $e_i = u_i - u_0$, for $i = 1, \ldots, n$; given the pair $(u_0, (e_1, \ldots, e_n))$ where $(e_1, \ldots, e_n)$ is a basis, we obtain the affine frame $(u_0, \ldots, u_n)$, with $u_i = u_0 + e_i$, for $i = 1, \ldots, n$. There is also a one-to-one correspondence between barycentric coordinates w.r.t. the affine frame $(u_0, \ldots, u_n)$ and standard coordinates w.r.t. the basis $(e_1, \ldots, e_n)$. The barycentric cordinates $(\lambda_0, \lambda_1, \ldots, \lambda_n)$ of v (with $\lambda_0 + \lambda_1 + \cdots + \lambda_n = 1$) yield the standard coordinates $(\lambda_1, \ldots, \lambda_n)$ of $v - u_0$; the standard coordinates $(x_1, \ldots, x_n)$ of $v - u_0$ yield the barycentric coordinates $(1 - (x_1 + \cdots + x_n), x_1, \ldots, x_n)$ of v.

(4) Recall that an affine map is a map $f \colon E \to F$ between vector spaces that preserves *affine combinations*; that is,

$$f\left(\sum_{i=1}^{m} \lambda_i u_i\right) = \sum_{i=1}^{m} \lambda_i f(u_i),$$

for all $u_1 \dots, u_m \in E$ and all $\lambda_i \in K$ such that $\sum_{i=1}^m \lambda_i = 1$.

Let $(u_0, \dots, u_n)$ be any affine frame in $\mathbb{R}^n$ and let $(v_0, \dots, v_n)$ be any vectors in $\mathbb{R}^m$. Prove that there is a *unique* affine map $f\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(u_i) = v_i, \quad i = 0, \dots, n.$$

(5) Let $(a_0, \dots, a_n)$ be any affine frame in $\mathbb{R}^n$ and let $(b_0, \dots, b_n)$ be any $n+1$ points in $\mathbb{R}^n$. Prove that there is a unique $(n+1) \times (n+1)$ matrix

$$A = \begin{pmatrix} B & w \\ 0 & 1 \end{pmatrix}$$

corresponding to the unique affine map f such that

$$f(a_i) = b_i, \quad i = 0, \dots, n,$$

in the sense that

$$A\widehat{a}_i = \widehat{b}_i, \quad i = 0, \dots, n,$$

and that A is given by

$$A = \begin{pmatrix} \widehat{b}_0 & \widehat{b}_1 & \cdots & \widehat{b}_n \end{pmatrix} \begin{pmatrix} \widehat{a}_0 & \widehat{a}_1 & \cdots & \widehat{a}_n \end{pmatrix}^{-1}.$$

Make sure to prove that the bottom row of A is $(0, \dots, 0, 1)$.

In the special case where $(a_0, \dots, a_n)$ is the canonical affine frame with $a_i = e_{i+1}$ for $i = 0, \dots, n-1$ and $a_n = (0, \dots, 0)$ (where e_i is the ith canonical basis vector), show that

$$\begin{pmatrix} \widehat{a}_0 & \widehat{a}_1 & \cdots & \widehat{a}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} \widehat{a}_0 & \widehat{a}_1 & \cdots & \widehat{a}_n \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \\ -1 & -1 & \cdots & -1 & 1 \end{pmatrix}.$$

For example, when $n = 2$, if we write $b_i = (x_i, y_i)$, then we have

$$A = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} x_1 - x_3 & x_2 - x_3 & x_3 \\ y_1 - y_3 & y_2 - y_3 & y_3 \\ 0 & 0 & 1 \end{pmatrix}.$$

(6) Recall that a nonempty affine subspace $\mathcal{A}$ of $\mathbb{R}^n$ is any nonempty subset of $\mathbb{R}^n$ closed under affine combinations. For any affine map $f\colon \mathbb{R}^n \to \mathbb{R}^m$, for any affine subspace $\mathcal{A}$ of $\mathbb{R}^n$, and any affine subspace $\mathcal{B}$ of $\mathbb{R}^m$, prove that $f(\mathcal{A})$ is an affine subspace of $\mathbb{R}^m$, and that $f^{-1}(\mathcal{B})$ is an affine subspace of $\mathbb{R}^n$.

Chapter 7

Determinants

In this chapter all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$.

7.1 Permutations, Signature of a Permutation

This chapter contains a review of determinants and their use in linear algebra. We begin with permutations and the signature of a permutation. Next, we define multilinear maps and alternating multilinear maps. Determinants are introduced as alternating multilinear maps taking the value 1 on the unit matrix (following Emil Artin). It is then shown how to compute a determinant using the Laplace expansion formula, and the connection with the usual definition is made. It is shown how determinants can be used to invert matrices and to solve (at least in theory!) systems of linear equations (the Cramer formulae). The determinant of a linear map is defined. We conclude by defining the characteristic polynomial of a matrix (and of a linear map) and by proving the celebrated Cayley-Hamilton theorem which states that every matrix is a "zero" of its characteristic polynomial (we give two proofs; one computational, the other one more conceptual).

Determinants can be defined in several ways. For example, determinants can be defined in a fancy way in terms of the exterior algebra (or alternating algebra) of a vector space. We will follow a more algorithmic approach due to Emil Artin. No matter which approach is followed, we need a few preliminaries about permutations on a finite set. We need to show that every permutation on n elements is a product of transpositions, and that the parity of the number of transpositions involved is an invariant of the permutation. Let $[n] = \{1, 2 \ldots, n\}$, where $n \in \mathbb{N}$, and $n > 0$.

Definition 7.1. A *permutation on n elements* is a bijection $\pi\colon [n] \to [n]$. When $n = 1$, the only function from [1] to [1] is the constant map: $1 \mapsto 1$. Thus, we will assume that $n \geq 2$. A *transposition* is a permutation $\tau\colon [n] \to [n]$ such that, for some $i < j$ (with $1 \leq i < j \leq n$), $\tau(i) = j$, $\tau(j) = i$, and $\tau(k) = k$, for all $k \in [n] - \{i, j\}$. In other words, a transposition exchanges two distinct elements $i, j \in [n]$. A *cyclic permutation of order k (or k-cycle)* is a permutation $\sigma\colon [n] \to [n]$ such that, for some sequence $(i_1, i_2, \ldots, i_k)$ of distinct elements of $[n]$ with $2 \leq k \leq n$,

$$\sigma(i_1) = i_2, \sigma(i_2) = i_3, \ldots, \sigma(i_{k-1}) = i_k,\ \sigma(i_k) = i_1,$$

and $\sigma(j) = j$, for $j \in [n] - \{i_1, \ldots, i_k\}$. The set $\{i_1, \ldots, i_k\}$ is called the *domain* of the cyclic permutation, and the cyclic permutation is usually denoted by $(i_1\, i_2\, \ldots\, i_k)$.

If τ is a transposition, clearly, $\tau \circ \tau = \mathrm{id}$. Also, a cyclic permutation of order 2 is a transposition, and for a cyclic permutation σ of order k, we have $\sigma^k = \mathrm{id}$. Clearly, the composition of two permutations is a permutation and every permutation has an inverse which is also a permutation. Therefore, the set of permutations on $[n]$ is a *group* often denoted $\mathfrak{S}_n$. It is easy to show by induction that the group $\mathfrak{S}_n$ has $n!$ elements. We will also use the terminology product of permutations (or transpositions), as a synonym for composition of permutations.

A permutation σ on n elements, say $\sigma(i) = k_i$ for $i = 1, \ldots, n$, can be represented in functional notation by the $2 \times n$ array

$$\begin{pmatrix} 1 & \cdots & i & \cdots & n \\ k_1 & \cdots & k_i & \cdots & k_n \end{pmatrix}$$

known as *Cauchy two-line notation*. For example, we have the permutation σ denoted by

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 3 & 6 & 5 & 1 \end{pmatrix}.$$

A more concise notation often used in computer science and in combinatorics is to represent a permutation by its image, namely by the sequence

$$\sigma(1)\ \sigma(2)\ \cdots\ \sigma(n)$$

written as a row vector without commas separating the entries. The above is known as the *one-line notation*. For example, in the one-line notation, our previous permutation σ is represented by

$$2\ 4\ 3\ 6\ 5\ 1.$$

The reason for not enclosing the above sequence within parentheses is avoid confusion with the notation for cycles, for which is it customary to include parentheses.

The following proposition shows the importance of cyclic permutations and transpositions.

Proposition 7.1. *For every $n \geq 2$, for every permutation $\pi\colon [n] \to [n]$, there is a partition of $[n]$ into r subsets called the orbits of π, with $1 \leq r \leq n$, where each set J in this partition is either a singleton $\{i\}$, or it is of the form*

$$J = \{i, \pi(i), \pi^2(i), \ldots, \pi^{r_i - 1}(i)\},$$

where r_i is the smallest integer, such that, $\pi^{r_i}(i) = i$ and $2 \leq r_i \leq n$. If π is not the identity, then it can be written in a unique way (up to the order) as a composition $\pi = \sigma_1 \circ \ldots \circ \sigma_s$ of cyclic permutations with disjoint domains, where s is the number of orbits with at least two elements. Every permutation $\pi \colon [n] \to [n]$ can be written as a nonempty composition of transpositions.

Proof. Consider the relation R_π defined on $[n]$ as follows: $iR_\pi j$ iff there is some $k \geq 1$ such that $j = \pi^k(i)$. We claim that R_π is an equivalence relation. Transitivity is obvious. We claim that for every $i \in [n]$, there is some least r $(1 \leq r \leq n)$ such that $\pi^r(i) = i$.

Indeed, consider the following sequence of $n + 1$ elements:

$$\langle i, \pi(i), \pi^2(i), \ldots, \pi^n(i) \rangle.$$

Since $[n]$ only has n distinct elements, there are some h, k with $0 \leq h < k \leq n$ such that

$$\pi^h(i) = \pi^k(i),$$

and since π is a bijection, this implies $\pi^{k-h}(i) = i$, where $0 \leq k - h \leq n$. Thus, we proved that there is some integer $m \geq 1$ such that $\pi^m(i) = i$, so there is such a smallest integer r.

Consequently, R_π is reflexive. It is symmetric, since if $j = \pi^k(i)$, letting r be the least $r \geq 1$ such that $\pi^r(i) = i$, then

$$i = \pi^{kr}(i) = \pi^{k(r-1)}(\pi^k(i)) = \pi^{k(r-1)}(j).$$

Now, for every $i \in [n]$, the equivalence class (orbit) of i is a subset of $[n]$, either the singleton $\{i\}$ or a set of the form

$$J = \{i, \pi(i), \pi^2(i), \ldots, \pi^{r_i - 1}(i)\},$$

where r_i is the smallest integer such that $\pi^{r_i}(i) = i$ and $2 \leq r_i \leq n$, and in the second case, the restriction of π to J induces a cyclic permutation σ_i, and $\pi = \sigma_1 \circ \ldots \circ \sigma_s$, where s is the number of equivalence classes having at least two elements.

For the second part of the proposition, we proceed by induction on n. If $n = 2$, there are exactly two permutations on $[2]$, the transposition τ exchanging 1 and 2, and the identity. However, $\mathrm{id}_2 = \tau^2$. Now, let $n \geq 3$. If $\pi(n) = n$, since by the induction hypothesis, the restriction of π to $[n - 1]$ can be written as a product of transpositions, π itself can be written as a product of transpositions. If $\pi(n) = k \neq n$, letting τ be the transposition such that $\tau(n) = k$ and $\tau(k) = n$, it is clear that $\tau \circ \pi$ leaves n invariant, and by the induction hypothesis, we have $\tau \circ \pi = \tau_m \circ \ldots \circ \tau_1$ for some transpositions, and thus

$$\pi = \tau \circ \tau_m \circ \ldots \circ \tau_1,$$

a product of transpositions (since $\tau \circ \tau = \mathrm{id}_n$). $\square$

Remark: When $\pi = \mathrm{id}_n$ is the identity permutation, we can agree that the composition of 0 transpositions is the identity. The second part of Proposition 7.1 shows that the transpositions generate the group of permutations $\mathfrak{S}_n$.

In writing a permutation π as a composition $\pi = \sigma_1 \circ \ldots \circ \sigma_s$ of cyclic permutations, it is clear that the order of the σ_i does not matter, since their domains are disjoint. Given a permutation written as a product of transpositions, we now show that the parity of the number of transpositions is an invariant.

Definition 7.2. For every $n \geq 2$, since every permutation $\pi\colon [n] \to [n]$ defines a partition of r subsets over which π acts either as the identity or as a cyclic permutation, let $\epsilon(\pi)$, called the *signature* of π, be defined by $\epsilon(\pi) = (-1)^{n-r}$, where r is the number of sets in the partition.

If τ is a transposition exchanging i and j, it is clear that the partition associated with τ consists of $n-1$ equivalence classes, the set $\{i, j\}$, and the $n-2$ singleton sets $\{k\}$, for $k \in [n] - \{i, j\}$, and thus, $\epsilon(\tau) = (-1)^{n-(n-1)} = (-1)^1 = -1$.

Proposition 7.2. *For every $n \geq 2$, for every permutation $\pi\colon [n] \to [n]$, for every transposition τ, we have*

$$\epsilon(\tau \circ \pi) = -\epsilon(\pi).$$

Consequently, for every product of transpositions such that $\pi = \tau_m \circ \ldots \circ \tau_1$, we have

$$\epsilon(\pi) = (-1)^m,$$

which shows that the parity of the number of transpositions is an invariant.

Proof. Assume that $\tau(i) = j$ and $\tau(j) = i$, where $i < j$. There are two cases, depending whether i and j are in the same equivalence class J_l of R_π, or if they are in distinct equivalence classes. If i and j are in the same class J_l, then if

$$J_l = \{i_1, \ldots, i_p, \ldots i_q, \ldots i_k\},$$

where $i_p = i$ and $i_q = j$, since

$$\tau(\pi(\pi^{-1}(i_p))) = \tau(i_p) = \tau(i) = j = i_q$$

and

$$\tau(\pi(i_{q-1})) = \tau(i_q) = \tau(j) = i = i_p,$$

it is clear that J_l splits into two subsets, one of which is $\{i_p, \ldots, i_{q-1}\}$, and thus, the number of classes associated with $\tau \circ \pi$ is $r+1$, and $\epsilon(\tau \circ \pi) = (-1)^{n-r-1} = -(-1)^{n-r} = -\epsilon(\pi)$. If i and j are in distinct equivalence classes J_l and J_m, say

$$\{i_1, \ldots, i_p, \ldots i_h\}$$

and

$$\{j_1, \ldots, j_q, \ldots j_k\},$$

where $i_p = i$ and $j_q = j$, since

$$\tau(\pi(\pi^{-1}(i_p))) = \tau(i_p) = \tau(i) = j = j_q$$

and

$$\tau(\pi(\pi^{-1}(j_q))) = \tau(j_q) = \tau(j) = i = i_p,$$

we see that the classes J_l and J_m merge into a single class, and thus, the number of classes associated with $\tau \circ \pi$ is $r - 1$, and $\epsilon(\tau \circ \pi) = (-1)^{n-r+1} = -(-1)^{n-r} = -\epsilon(\pi)$.

Now, let $\pi = \tau_m \circ \ldots \circ \tau_1$ be any product of transpositions. By the first part of the proposition, we have

$$\epsilon(\pi) = (-1)^{m-1}\epsilon(\tau_1) = (-1)^{m-1}(-1) = (-1)^m,$$

since $\epsilon(\tau_1) = -1$ for a transposition. $\square$

Remark: When $\pi = \mathrm{id}_n$ is the identity permutation, since we agreed that the composition of 0 transpositions is the identity, it it still correct that $(-1)^0 = \epsilon(\mathrm{id}) = +1$. From the proposition, it is immediate that $\epsilon(\pi' \circ \pi) = \epsilon(\pi')\epsilon(\pi)$. In particular, since $\pi^{-1} \circ \pi = \mathrm{id}_n$, we get $\epsilon(\pi^{-1}) = \epsilon(\pi)$.

We can now proceed with the definition of determinants.

7.2 Alternating Multilinear Maps

First we define multilinear maps, symmetric multilinear maps, and alternating multilinear maps.

Remark: Most of the definitions and results presented in this section also hold when K is a commutative ring and when we consider modules over K (free modules, when bases are needed).

Let $E_1, \ldots, E_n$, and F, be vector spaces over a field K, where $n \geq 1$.

Definition 7.3. A function $f\colon E_1 \times \ldots \times E_n \to F$ is a *multilinear map (or an n-linear map)* if it is linear in each argument, holding the others fixed. More explicitly, for every i, $1 \leq i \leq n$, for all $x_1 \in E_1, \ldots, x_{i-1} \in E_{i-1}$, $x_{i+1} \in E_{i+1}, \ldots, x_n \in E_n$, for all $x, y \in E_i$, for all $\lambda \in K$,

$$\begin{aligned} f(x_1, \ldots, x_{i-1}, x + y, x_{i+1}, \ldots, x_n) &= f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \\ &\quad + f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n), \\ f(x_1, \ldots, x_{i-1}, \lambda x, x_{i+1}, \ldots, x_n) &= \lambda f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n). \end{aligned}$$

When $F = K$, we call f an *n-linear form (or multilinear form)*. If $n \geq 2$ and $E_1 = E_2 = \ldots = E_n$, an n-linear map $f\colon E \times \ldots \times E \to F$ is called *symmetric*, if $f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)})$ for every permutation π on $\{1, \ldots, n\}$. An n-linear map $f\colon E \times \ldots \times E \to F$ is called *alternating*, if $f(x_1, \ldots, x_n) = 0$ whenever $x_i = x_{i+1}$ for some i, $1 \leq i \leq n-1$ (in other words, when two adjacent arguments are equal). It does no harm to agree that when $n = 1$, a linear map is considered to be both symmetric and alternating, and we will do so.

When $n = 2$, a 2-linear map $f\colon E_1 \times E_2 \to F$ is called a *bilinear map*. We have already seen several examples of bilinear maps. Multiplication $\cdot\colon K \times K \to K$ is a bilinear map, treating K as a vector space over itself.

The operation $\langle -, - \rangle\colon E^* \times E \to K$ applying a linear form to a vector is a bilinear map.

Symmetric bilinear maps (and multilinear maps) play an important role in geometry (inner products, quadratic forms) and in differential calculus (partial derivatives).

A bilinear map is symmetric if $f(u, v) = f(v, u)$, for all $u, v \in E$.

Alternating multilinear maps satisfy the following simple but crucial properties.

Proposition 7.3. *Let $f\colon E \times \ldots \times E \to F$ be an n-linear alternating map, with $n \geq 2$. The following properties hold:*

(1)
$$f(\ldots, x_i, x_{i+1}, \ldots) = -f(\ldots, x_{i+1}, x_i, \ldots)$$

(2)
$$f(\ldots, x_i, \ldots, x_j, \ldots) = 0,$$
where $x_i = x_j$, and $1 \leq i < j \leq n$.

(3)
$$f(\ldots, x_i, \ldots, x_j, \ldots) = -f(\ldots, x_j, \ldots, x_i, \ldots),$$
where $1 \leq i < j \leq n$.

(4)
$$f(\ldots, x_i, \ldots) = f(\ldots, x_i + \lambda x_j, \ldots),$$
for any $\lambda \in K$, and where $i \neq j$.

Proof. (1) By multilinearity applied twice, we have

$$\begin{aligned} f(\ldots, x_i + x_{i+1}, x_i + x_{i+1}, \ldots) &= f(\ldots, x_i, x_i, \ldots) + f(\ldots, x_i, x_{i+1}, \ldots) \\ &\quad + f(\ldots, x_{i+1}, x_i, \ldots) + f(\ldots, x_{i+1}, x_{i+1}, \ldots), \end{aligned}$$

and since f is alternating, this yields

$$0 = f(\ldots, x_i, x_{i+1}, \ldots) + f(\ldots, x_{i+1}, x_i, \ldots),$$

that is, $f(\ldots, x_i, x_{i+1}, \ldots) = -f(\ldots, x_{i+1}, x_i, \ldots)$.

(2) If $x_i = x_j$ and i and j are not adjacent, we can interchange x_i and x_{i+1}, and then x_i and x_{i+2}, etc, until x_i and x_j become adjacent. By (1),

$$f(\ldots, x_i, \ldots, x_j, \ldots) = \epsilon f(\ldots, x_i, x_j, \ldots),$$

where $\epsilon = +1$ or -1, but $f(\ldots, x_i, x_j, \ldots) = 0$, since $x_i = x_j$, and (2) holds.

(3) follows from (2) as in (1). (4) is an immediate consequence of (2). $\square$

Proposition 7.3 will now be used to show a fundamental property of alternating multilinear maps. First we need to extend the matrix notation a little bit. Let E be a vector space over K. Given an $n \times n$ matrix $A = (a_{i\,j})$ over K, we can define a map $L(A)\colon E^n \to E^n$ as follows:

$$\begin{aligned} L(A)_1(u) &= a_{1\,1}u_1 + \cdots + a_{1\,n}u_n, \\ &\cdots \\ L(A)_n(u) &= a_{n\,1}u_1 + \cdots + a_{n\,n}u_n, \end{aligned}$$

for all $u_1, \ldots, u_n \in E$ and with $u = (u_1, \ldots, u_n)$. It is immediately verified that $L(A)$ is linear. Then given two $n \times n$ matrices $A = (a_{i\,j})$ and $B = (b_{i\,j})$, by repeating the calculations establishing the product of matrices (just before Definition 3.12), we can show that

$$L(AB) = L(A) \circ L(B).$$

It is then convenient to use the matrix notation to describe the effect of the linear map $L(A)$, as

$$\begin{pmatrix} L(A)_1(u) \\ L(A)_2(u) \\ \vdots \\ L(A)_n(u) \end{pmatrix} = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Lemma 7.4. *Let $f\colon E \times \ldots \times E \to F$ be an n-linear alternating map. Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two families of n vectors, such that,*

$$\begin{aligned} v_1 &= a_{1\,1}u_1 + \cdots + a_{n\,1}u_n, \\ &\cdots \\ v_n &= a_{1\,n}u_1 + \cdots + a_{n\,n}u_n. \end{aligned}$$

Equivalently, letting

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix},$$

assume that we have

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = A^{\top} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Then,

$$f(v_1, \ldots, v_n) = \Big(\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}\Big) f(u_1, \ldots, u_n),$$

where the sum ranges over all permutations π *on* $\{1, \ldots, n\}$*.*

Proof. Expanding $f(v_1, \ldots, v_n)$ by multilinearity, we get a sum of terms of the form

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_{\pi(1)}, \ldots, u_{\pi(n)}),$$

for all possible functions $\pi\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. However, because f is alternating, only the terms for which π is a permutation are nonzero. By Proposition 7.1, every permutation π is a product of transpositions, and by Proposition 7.2, the parity $\epsilon(\pi)$ of the number of transpositions only depends on π. Then applying Proposition 7.3 (3) to each transposition in π, we get

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_{\pi(1)}, \ldots, u_{\pi(n)}) = \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} f(u_1, \ldots, u_n).$$

Thus, we get the expression of the lemma. ∎

For the case of $n = 2$, the proof details of Lemma 7.4 become

$$\begin{aligned}
f(v_1, v_2) &= f(a_{11}u_1 + a_{21}u_2, a_{12}u_1 + a_{22}u_2) \\
&= f(a_{11}u_1 + a_{21}u_2, a_{12}u_1) + f(a_{11}u_1 + a_{21}u_2, a_{22}u_2) \\
&= f(a_{11}u_1, a_{12}u_1) + f(a_{21}u_2, a_{12}u_1) + f(a_{11}u_a, a_{22}u_2) + f(a_{21}u_2, a_{22}u_2) \\
&= a_{11}a_{12}f(u_1, u_1) + a_{21}a_{12}f(u_2, u_1) + a_{11}a_{22}f(u_1, u_2) + a_{21}a_{22}f(u_2, u_2) \\
&= a_{21}a_{12}f(u_2, u_1)a_{11}a_{22}f(u_1, u_2) \\
&= (a_{11}a_{22} - a_{12}a_{22})\, f(u_1, u_2).
\end{aligned}$$

Hopefully the reader will recognize the quantity $a_{11}a_{22} - a_{12}a_{22}$. It is the determinant of the 2×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

This is no accident. The quantity

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}$$

is in fact the value of the determinant of A (which, as we shall see shortly, is also equal to the determinant of $A^\top$). However, working directly with the above definition is quite awkward, and we will proceed via a slightly indirect route

Remark: The reader might have been puzzled by the fact that it is the transpose matrix $A^\top$ rather than A itself that appears in Lemma 7.4. The reason is that if we want the generic term in the determinant to be

$$\epsilon(\pi)a_{\pi(1)\,1}\cdots a_{\pi(n)\,n},$$

where the permutation applies to the first index, then we have to express the v_js in terms of the u_is in terms of $A^\top$ as we did. Furthermore, since

$$v_j = a_{1\,j}u_1 + \cdots + a_{i\,j}u_i + \cdots + a_{n\,j}u_n,$$

we see that v_j corresponds to the jth column of the matrix A, and so the determinant is viewed as a function of the *columns* of A.

The literature is split on this point. Some authors prefer to define a determinant as we did. Others use A itself, which amounts to viewing det as a function of the rows, in which case we get the expression

$$\sum_{\sigma\in\mathfrak{S}_n}\epsilon(\sigma)a_{1\,\sigma(1)}\cdots a_{n\,\sigma(n)}.$$

Corollary 7.7 show that these two expressions are equal, so it doesn't matter which is chosen. This is a matter of taste.

7.3 Definition of a Determinant

Recall that the set of all square $n\times n$-matrices with coefficients in a field K is denoted by $\mathrm{M}_n(K)$.

Definition 7.4. A *determinant* is defined as any map

$$D\colon \mathrm{M}_n(K)\to K,$$

which, when viewed as a map on $(K^n)^n$, i.e., a map of the n columns of a matrix, is n-linear alternating and such that $D(I_n)=1$ for the identity matrix I_n. Equivalently, we can consider a vector space E of dimension n, some fixed basis $(e_1,\ldots,e_n)$, and define

$$D\colon E^n\to K$$

as an n-linear alternating map such that $D(e_1,\ldots,e_n)=1$.

First we will show that such maps D exist, using an inductive definition that also gives a recursive method for computing determinants. Actually, we will define a family $(\mathcal{D}_n)_{n\geq 1}$ of (finite) sets of maps $D\colon \mathrm{M}_n(K) \to K$. Second we will show that determinants are in fact uniquely defined, that is, we will show that each $\mathcal{D}_n$ consists of a *single map*. This will show the equivalence of the direct definition $\det(A)$ of Lemma 7.4 with the inductive definition $D(A)$. Finally, we will prove some basic properties of determinants, using the uniqueness theorem.

Given a matrix $A \in \mathrm{M}_n(K)$, we denote its n columns by $A^1, \ldots, A^n$. In order to describe the recursive process to define a determinant we need the notion of a minor.

Definition 7.5. Given any $n\times n$ matrix with $n \geq 2$, for any two indices i, j with $1 \leq i, j \leq n$, let A_{ij} be the $(n-1)\times(n-1)$ matrix obtained by deleting Row i and Column j from A and called a *minor*:

$$A_{ij} = \begin{pmatrix} & & & & \times & & \\ & & & & \times & & \\ \times & \times & \times & \times & \times & \times & \times \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \\ & & & & \times & & \end{pmatrix}.$$

For example, if

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

then

$$A_{2\,3} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Definition 7.6. For every $n \geq 1$, we define a finite set $\mathcal{D}_n$ of maps $D\colon \mathrm{M}_n(K) \to K$ inductively as follows:

When $n = 1$, $\mathcal{D}_1$ consists of the single map D such that, $D(A) = a$, where $A = (a)$, with $a \in K$.

Assume that $\mathcal{D}_{n-1}$ has been defined, where $n \geq 2$. Then $\mathcal{D}_n$ consists of all the maps D such that, for some i, $1 \leq i \leq n$,

$$D(A) = (-1)^{i+1} a_{i\,1} D(A_{i\,1}) + \cdots + (-1)^{i+n} a_{i\,n} D(A_{i\,n}),$$

where for every j, $1 \leq j \leq n$, $D(A_{i\,j})$ is the result of applying any D in $\mathcal{D}_{n-1}$ to the minor $A_{i\,j}$.

We confess that the use of the same letter D for the member of $\mathcal{D}_n$ being defined, and for members of $\mathcal{D}_{n-1}$, may be slightly confusing. We considered using subscripts to distinguish, but this seems to complicate things unnecessarily. One should not worry too much anyway, since it will turn out that each $\mathcal{D}_n$ contains just one map.

Each $(-1)^{i+j}D(A_{i\,j})$ is called the *cofactor* of $a_{i\,j}$, and the inductive expression for $D(A)$ is called a *Laplace expansion of D according to the i-th Row*. Given a matrix $A \in \mathrm{M}_n(K)$, each $D(A)$ is called a *determinant* of A.

We can think of each member of $\mathcal{D}_n$ as an *algorithm* to evaluate "the" determinant of A. The main point is that these algorithms, which recursively evaluate a determinant using all possible Laplace row expansions, all yield the *same result*, $\det(A)$.

We will prove shortly that $D(A)$ is uniquely defined (at the moment, it is not clear that $\mathcal{D}_n$ consists of a single map). Assuming this fact, given a $n \times n$-matrix $A = (a_{i\,j})$,

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \dots & a_{n\,n} \end{pmatrix},$$

its determinant is denoted by $D(A)$ or $\det(A)$, or more explicitly by

$$\det(A) = \begin{vmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \dots & a_{n\,n} \end{vmatrix}.$$

Let us first consider some examples.

Example 7.1.

1. When $n = 2$, if
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$
then by expanding according to any row, we have
$$D(A) = ad - bc.$$

2. When $n = 3$, if
$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} \\ a_{2\,1} & a_{2\,2} & a_{2\,3} \\ a_{3\,1} & a_{3\,2} & a_{3\,3} \end{pmatrix},$$

then by expanding according to the first row, we have

$$D(A) = a_{1\,1} \begin{vmatrix} a_{2\,2} & a_{2\,3} \\ a_{3\,2} & a_{3\,3} \end{vmatrix} - a_{1\,2} \begin{vmatrix} a_{2\,1} & a_{2\,3} \\ a_{3\,1} & a_{3\,3} \end{vmatrix} + a_{1\,3} \begin{vmatrix} a_{2\,1} & a_{2\,2} \\ a_{3\,1} & a_{3\,2} \end{vmatrix},$$

that is,

$$D(A) = a_{1\,1}(a_{2\,2}a_{3\,3} - a_{3\,2}a_{2\,3}) - a_{1\,2}(a_{2\,1}a_{3\,3} - a_{3\,1}a_{2\,3}) + a_{1\,3}(a_{2\,1}a_{3\,2} - a_{3\,1}a_{2\,2}),$$

which gives the explicit formula

$$D(A) = a_{1\,1}a_{2\,2}a_{3\,3} + a_{2\,1}a_{3\,2}a_{1\,3} + a_{3\,1}a_{1\,2}a_{2\,3} - a_{1\,1}a_{3\,2}a_{2\,3} - a_{2\,1}a_{1\,2}a_{3\,3} - a_{3\,1}a_{2\,2}a_{1\,3}.$$

We now show that each $D \in \mathcal{D}_n$ is a determinant (map).

Lemma 7.5. *For every $n \geq 1$, for every $D \in \mathcal{D}_n$ as defined in Definition 7.6, D is an alternating multilinear map such that $D(I_n) = 1$.*

Proof. By induction on n, it is obvious that $D(I_n) = 1$. Let us now prove that D is multilinear. Let us show that D is linear in each column. Consider any Column k. Since

$$D(A) = (-1)^{i+1}a_{i\,1}D(A_{i\,1}) + \cdots + (-1)^{i+j}a_{i\,j}D(A_{i\,j}) + \cdots + (-1)^{i+n}a_{i\,n}D(A_{i\,n}),$$

if $j \neq k$, then by induction, $D(A_{i\,j})$ is linear in Column k, and $a_{i\,j}$ does not belong to Column k, so $(-1)^{i+j}a_{i\,j}D(A_{i\,j})$ is linear in Column k. If $j = k$, then $D(A_{i\,j})$ does not depend on Column $k = j$, since $A_{i\,j}$ is obtained from A by deleting Row i and Column $j = k$, and $a_{i\,j}$ belongs to Column $j = k$. Thus, $(-1)^{i+j}a_{i\,j}D(A_{i\,j})$ is linear in Column k. Consequently, in all cases, $(-1)^{i+j}a_{i\,j}D(A_{i\,j})$ is linear in Column k, and thus, $D(A)$ is linear in Column k.

Let us now prove that D is alternating. Assume that two adjacent columns of A are equal, say $A^k = A^{k+1}$. Assume that $j \neq k$ and $j \neq k+1$. Then the matrix $A_{i\,j}$ has two identical adjacent columns, and by the induction hypothesis, $D(A_{i\,j}) = 0$. The remaining terms of $D(A)$ are

$$(-1)^{i+k}a_{i\,k}D(A_{i\,k}) + (-1)^{i+k+1}a_{i\,k+1}D(A_{i\,k+1}).$$

However, the two matrices $A_{i\,k}$ and $A_{i\,k+1}$ are equal, since we are assuming that Columns k and $k+1$ of A are identical and $A_{i\,k}$ is obtained from A by deleting Row i and Column k while $A_{i\,k+1}$ is obtained from A by deleting Row i and Column $k+1$. Similarly, $a_{i\,k} = a_{i\,k+1}$, since Columns k and $k+1$ of A are equal. But then,

$$(-1)^{i+k}a_{i\,k}D(A_{i\,k}) + (-1)^{i+k+1}a_{i\,k+1}D(A_{i\,k+1}) = (-1)^{i+k}a_{i\,k}D(A_{i\,k}) - (-1)^{i+k}a_{i\,k}D(A_{i\,k}) = 0.$$

This shows that D is alternating and completes the proof. $\square$

Lemma 7.5 shows the existence of determinants. We now prove their uniqueness.

Theorem 7.6. *For every $n \geq 1$, for every $D \in \mathcal{D}_n$, for every matrix $A \in \mathrm{M}_n(K)$, we have*

$$D(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. As a consequence, $\mathcal{D}_n$ consists of a single map for every $n \geq 1$, and this map is given by the above explicit formula.

Proof. Consider the standard basis $(e_1, \ldots, e_n)$ of K^n, where $(e_i)_i = 1$ and $(e_i)_j = 0$, for $j \neq i$. Then each column A^j of A corresponds to a vector v_j whose coordinates over the basis $(e_1, \ldots, e_n)$ are the components of A^j, that is, we can write

$$\begin{aligned} v_1 &= a_{1\,1} e_1 + \cdots + a_{n\,1} e_n, \\ &\cdots \\ v_n &= a_{1\,n} e_1 + \cdots + a_{n\,n} e_n. \end{aligned}$$

Since by Lemma 7.5, each D is a multilinear alternating map, by applying Lemma 7.4, we get

$$D(A) = D(v_1, \ldots, v_n) = \Big(\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} \Big) D(e_1, \ldots, e_n),$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. But $D(e_1, \ldots, e_n) = D(I_n)$, and by Lemma 7.5, we have $D(I_n) = 1$. Thus,

$$D(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. $\square$

From now on we will favor the notation $\det(A)$ over $D(A)$ for the determinant of a square matrix.

Remark: There is a geometric interpretation of determinants which we find quite illuminating. Given n linearly independent vectors $(u_1, \ldots, u_n)$ in $\mathbb{R}^n$, the set

$$P_n = \{\lambda_1 u_1 + \cdots + \lambda_n u_n \mid 0 \leq \lambda_i \leq 1,\ 1 \leq i \leq n\}$$

is called a *parallelotope*. If $n = 2$, then P_2 is a *parallelogram* and if $n = 3$, then P_3 is a *parallelepiped*, a skew box having u_1, u_2, u_3 as three of its corner sides. See Figures 7.1 and 7.2.

Then it turns out that $\det(u_1, \ldots, u_n)$ is the *signed volume* of the parallelotope P_n (where volume means n-dimensional volume). The sign of this volume accounts for the orientation of P_n in $\mathbb{R}^n$.

We can now prove some properties of determinants.

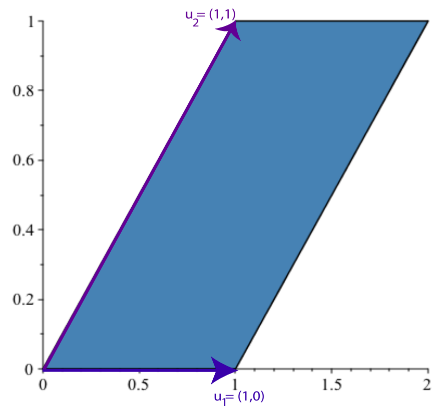


Figure 7.1: The parallelogram in $\mathbb{R}^w$ spanned by the vectors $u_1 = (1, 0)$ and $u_2 = (1, 1)$.

Corollary 7.7. *For every matrix $A \in \mathrm{M}_n(K)$, we have* $\det(A) = \det(A^\top)$.

Proof. By Theorem 7.6, we have

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n},$$

where the sum ranges over all permutations π on $\{1, \ldots, n\}$. Since a permutation is invertible, every product

$$a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}$$

can be rewritten as

$$a_{1\,\pi^{-1}(1)} \cdots a_{n\,\pi^{-1}(n)},$$

and since $\epsilon(\pi^{-1}) = \epsilon(\pi)$ and the sum is taken over all permutations on $\{1, \ldots, n\}$, we have

$$\sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n} = \sum_{\sigma \in \mathfrak{S}_n} \epsilon(\sigma) a_{1\,\sigma(1)} \cdots a_{n\,\sigma(n)},$$

where π and σ range over all permutations. But it is immediately verified that

$$\det(A^\top) = \sum_{\sigma \in \mathfrak{S}_n} \epsilon(\sigma) a_{1\,\sigma(1)} \cdots a_{n\,\sigma(n)}. \qquad \square$$

A useful consequence of Corollary 7.7 is that the determinant of a matrix is also a multilinear alternating map of its *rows*. This fact, combined with the fact that the determinant of a matrix is a multilinear alternating map of its columns, is often useful for finding short-cuts in computing determinants. We illustrate this point on the following example which shows up in polynomial interpolation.

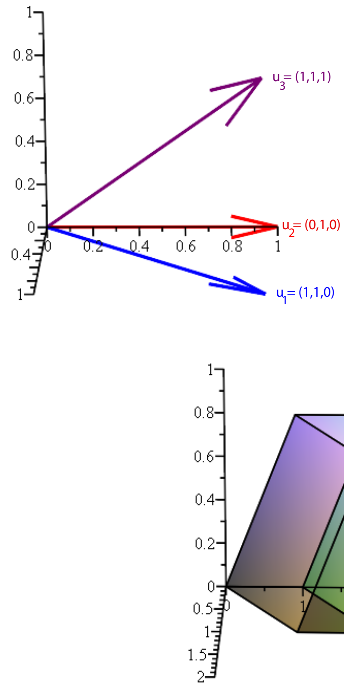


Figure 7.2: The parallelepiped in $\mathbb{R}^3$ spanned by the vectors $u_1 = (1, 1, 0)$, $u_2 = (0, 1, 0)$, and $u_3 = (0, 0, 1)$.

Example 7.2. Consider the so-called *Vandermonde determinant*

$$V(x_1, \ldots, x_n) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{vmatrix}.$$

We claim that

$$V(x_1, \ldots, x_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i),$$

with $V(x_1, \ldots, x_n) = 1$, when $n = 1$. We prove it by induction on $n \geq 1$. The case $n = 1$ is obvious. Assume $n \geq 2$. We proceed as follows: multiply Row $n - 1$ by x_1 and subtract it from Row n (the last row), then multiply Row $n - 2$ by x_1 and subtract it from Row $n - 1$, etc, multiply Row $i - 1$ by x_1 and subtract it from row i, until we reach Row 1. We obtain the following determinant:

$$V(x_1, \ldots, x_n) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ 0 & x_2 - x_1 & \ldots & x_n - x_1 \\ 0 & x_2(x_2 - x_1) & \ldots & x_n(x_n - x_1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_2^{n-2}(x_2 - x_1) & \ldots & x_n^{n-2}(x_n - x_1) \end{vmatrix}.$$

Now expanding this determinant according to the first column and using multilinearity, we can factor $(x_i - x_1)$ from the column of index $i - 1$ of the matrix obtained by deleting the first row and the first column, and thus

$$V(x_1, \ldots, x_n) = (x_2 - x_1)(x_3 - x_1) \cdots (x_n - x_1) V(x_2, \ldots, x_n),$$

which establishes the induction step.

Example 7.3. The determinant of upper triangular matrices and more generally of block matrices that are block upper triangular has a remarkable form. Recall that an $n \times n$ matrix $A = (a_{ij})$ is upper-triangular if it is of the form

$$A = \begin{pmatrix} a_{11} & \times & \times & \cdots & \times \\ 0 & a_{22} & \times & \cdots & \times \\ 0 & 0 & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a_{nn} \end{pmatrix},$$

that is, $a_{ij} = 0$ for all $i > j$, $1 \leq i, j \leq n$. Using $n - 1$ times Laplace expansion with respect to the first column we obain

$$\det(A) = a_{11} a_{22} \cdots a_{nn}.$$

Similarly, if A is an $n \times n$ block matrix which is *block upper triangular*,

$$A = \begin{pmatrix} A_{11} & \times & \times & \cdots & \times \\ 0 & A_{22} & \times & \cdots & \times \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & A_{pp} \end{pmatrix},$$

where each A_{ii} is an $n_i \times n_i$ matrix, with $n_1 + \cdots + n_p = n$, each block $\times$ above the diagonal in position (i, j) for $i < j$ is an $n_i \times n_j$ matrix, and each block in position (i, j) for $i > j$ is the $n_i \times n_j$ zero matrix, then it can be shown by induction on $p \geq 1$ that

$$\det(A) = \det(A_{11}) \det(A_{22}) \cdots \det(A_{pp}).$$

Lemma 7.4 can be reformulated nicely as follows.

Proposition 7.8. *Let $f\colon E \times \ldots \times E \to F$ be an n-linear alternating map. Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two families of n vectors, such that*

$$\begin{aligned} v_1 &= a_{1\,1}u_1 + \cdots + a_{1\,n}u_n, \\ &\cdots \\ v_n &= a_{n\,1}u_1 + \cdots + a_{n\,n}u_n. \end{aligned}$$

Equivalently, letting

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix},$$

assume that we have

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = A \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Then,

$$f(v_1, \ldots, v_n) = \det(A) f(u_1, \ldots, u_n).$$

Proof. The only difference with Lemma 7.4 is that here we are using $A^\top$ instead of A. Thus, by Lemma 7.4 and Corollary 7.7, we get the desired result. ☐

As a consequence, we get the very useful property that the determinant of a product of matrices is the product of the determinants of these matrices.

Proposition 7.9. *For any two $n \times n$-matrices A and B, we have* $\det(AB) = \det(A)\det(B)$.

Proof. We use Proposition 7.8 as follows: let $(e_1, \ldots, e_n)$ be the standard basis of K^n, and let

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = AB \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Then we get

$$\det(w_1, \ldots, w_n) = \det(AB)\det(e_1, \ldots, e_n) = \det(AB),$$

since $\det(e_1, \ldots, e_n) = 1$. Now letting

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = B \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

we get

$$\det(v_1, \ldots, v_n) = \det(B),$$

and since

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = A \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

we get

$$\det(w_1, \ldots, w_n) = \det(A) \det(v_1, \ldots, v_n) = \det(A) \det(B).$$ □

It should be noted that all the results of this section, up to now, also hold when K is a commutative ring and not necessarily a field. We can now characterize when an $n \times n$-matrix A is invertible in terms of its determinant $\det(A)$.

7.4 Inverse Matrices and Determinants

In the next two sections, K is a commutative ring and when needed a field.

Definition 7.7. Let K be a commutative ring. Given a matrix $A \in \mathrm{M}_n(K)$, let $\widetilde{A} = (b_{i\,j})$ be the matrix defined such that

$$b_{i\,j} = (-1)^{i+j} \det(A_{j\,i}),$$

the cofactor of $a_{j\,i}$. The matrix $\widetilde{A}$ is called the *adjugate* of A, and each matrix $A_{j\,i}$ is called a *minor* of the matrix A.

For example, if

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -2 & -2 \\ 3 & 3 & -3 \end{pmatrix},$$

we have

$$b_{11} = \det(A_{11}) = \begin{vmatrix} -2 & -2 \\ 3 & -3 \end{vmatrix} = 12 \qquad b_{12} = -\det(A_{21}) = -\begin{vmatrix} 1 & 1 \\ 3 & -3 \end{vmatrix} = 6$$

$$b_{13} = \det(A_{31}) = \begin{vmatrix} 1 & 1 \\ -2 & -2 \end{vmatrix} = 0 \qquad b_{21} = -\det(A_{12}) = -\begin{vmatrix} 2 & -2 \\ 3 & -3 \end{vmatrix} = 0$$

$$b_{22} = \det(A_{22}) = \begin{vmatrix} 1 & 1 \\ 3 & -3 \end{vmatrix} = -6 \qquad b_{23} = -\det(A_{32}) = -\begin{vmatrix} 1 & 1 \\ 2 & -2 \end{vmatrix} = 4$$

$$b_{31} = \det(A_{13}) = \begin{vmatrix} 2 & -2 \\ 3 & 3 \end{vmatrix} = 12 \qquad b_{32} = -\det(A_{23}) = -\begin{vmatrix} 1 & 1 \\ 3 & 3 \end{vmatrix} = 0$$

$$b_{33} = \det(A_{33}) = \begin{vmatrix} 1 & 1 \\ 2 & -2 \end{vmatrix} = -4,$$

we find that

$$\widetilde{A} = \begin{pmatrix} 12 & 6 & 0 \\ 0 & -6 & 4 \\ 12 & 0 & -4 \end{pmatrix}.$$

Note the reversal of the indices in

$$b_{i\,j} = (-1)^{i+j} \det(A_{j\,i}).$$

Thus, $\widetilde{A}$ is the *transpose* of the matrix of cofactors of elements of A.

We have the following proposition.

Proposition 7.10. *Let K be a commutative ring. For every matrix $A \in \mathrm{M}_n(K)$, we have*

$$A\widetilde{A} = \widetilde{A}A = \det(A)I_n.$$

As a consequence, A is invertible iff $\det(A)$ is invertible, and if so, $A^{-1} = (\det(A))^{-1}\widetilde{A}$.

Proof. If $\widetilde{A} = (b_{i\,j})$ and $A\widetilde{A} = (c_{i\,j})$, we know that the entry $c_{i\,j}$ in row i and column j of $A\widetilde{A}$ is

$$c_{i\,j} = a_{i\,1}b_{1\,j} + \cdots + a_{i\,k}b_{k\,j} + \cdots + a_{i\,n}b_{n\,j},$$

which is equal to

$$a_{i\,1}(-1)^{j+1}\det(A_{j\,1}) + \cdots + a_{i\,n}(-1)^{j+n}\det(A_{j\,n}).$$

If $j = i$, then we recognize the expression of the expansion of $\det(A)$ according to the i-th row:

$$c_{i\,i} = \det(A) = a_{i\,1}(-1)^{i+1}\det(A_{i\,1}) + \cdots + a_{i\,n}(-1)^{i+n}\det(A_{i\,n}).$$

If $j \neq i$, we can form the matrix A' by replacing the j-th row of A by the i-th row of A. Now the matrix $A_{j\,k}$ obtained by deleting row j and column k from A is equal to the matrix

$A'_{j\,k}$ obtained by deleting row j and column k from A', since A and A' only differ by the j-th row. Thus,

$$\det(A_{j\,k}) = \det(A'_{j\,k}),$$

and we have

$$c_{i\,j} = a_{i\,1}(-1)^{j+1}\det(A'_{j\,1}) + \cdots + a_{i\,n}(-1)^{j+n}\det(A'_{j\,n}).$$

However, this is the expansion of $\det(A')$ according to the j-th row, since the j-th row of A' is equal to the i-th row of A. Furthermore, since A' has two identical rows i and j, because det is an alternating map of the rows (see an earlier remark), we have $\det(A') = 0$. Thus, we have shown that $c_{i\,i} = \det(A)$, and $c_{i\,j} = 0$, when $j \neq i$, and so

$$A\widetilde{A} = \det(A)I_n.$$

It is also obvious from the definition of $\widetilde{A}$, that

$$\widetilde{A}^\top = \widetilde{A^\top}.$$

Then applying the first part of the argument to $A^\top$, we have

$$A^\top \widetilde{A^\top} = \det(A^\top)I_n,$$

and since $\det(A^\top) = \det(A)$, $\widetilde{A}^\top = \widetilde{A^\top}$, and $(\widetilde{A}A)^\top = A^\top \widetilde{A}^\top$, we get

$$\det(A)I_n = A^\top \widetilde{A^\top} = A^\top \widetilde{A}^\top = (\widetilde{A}A)^\top,$$

that is,

$$(\widetilde{A}A)^\top = \det(A)I_n,$$

which yields

$$\widetilde{A}A = \det(A)I_n,$$

since $I_n^\top = I_n$. This proves that

$$A\widetilde{A} = \widetilde{A}A = \det(A)I_n.$$

As a consequence, if $\det(A)$ is invertible, we have $A^{-1} = (\det(A))^{-1}\widetilde{A}$. Conversely, if A is invertible, from $AA^{-1} = I_n$, by Proposition 7.9, we have $\det(A)\det(A^{-1}) = 1$, and $\det(A)$ is invertible. $\square$

For example, we saw earlier that

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -2 & -2 \\ 3 & 3 & -3 \end{pmatrix} \quad \text{and} \quad \widetilde{A} = \begin{pmatrix} 12 & 6 & 0 \\ 0 & -6 & 4 \\ 12 & 0 & -4 \end{pmatrix},$$

and we have

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & -2 & -2 \\ 3 & 3 & -3 \end{pmatrix} \begin{pmatrix} 12 & 6 & 0 \\ 0 & -6 & 4 \\ 12 & 0 & -4 \end{pmatrix} = 24 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

with $\det(A) = 24$.

When K is a field, an element $a \in K$ is invertible iff $a \neq 0$. In this case, the second part of the proposition can be stated as A is invertible iff $\det(A) \neq 0$. Note in passing that this method of computing the inverse of a matrix is usually not practical.

7.5 Systems of Linear Equations and Determinants

We now consider some applications of determinants to linear independence and to solving systems of linear equations. Although these results hold for matrices over certain rings, their proofs require more sophisticated methods. Therefore, we assume again that K is a field (usually, $K = \mathbb{R}$ or $K = \mathbb{C}$).

Let A be an $n \times n$-matrix, x a column vectors of variables, and b another column vector, and let $A^1, \ldots, A^n$ denote the columns of A. Observe that the system of equations $Ax = b$,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n\,1} & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

is equivalent to

$$x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n = b,$$

since the equation corresponding to the i-th row is in both cases

$$a_{i\,1} x_1 + \cdots + a_{i\,j} x_j + \cdots + a_{i\,n} x_n = b_i.$$

First we characterize linear independence of the column vectors of a matrix A in terms of its determinant.

Proposition 7.11. *Given an $n \times n$-matrix A over a field K, the columns $A^1, \ldots, A^n$ of A are linearly dependent iff* $\det(A) = \det(A^1, \ldots, A^n) = 0$. *Equivalently, A has rank n iff* $\det(A) \neq 0$.

Proof. First assume that the columns $A^1, \ldots, A^n$ of A are linearly dependent. Then there are $x_1, \ldots, x_n \in K$, such that

$$x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n = 0,$$

where $x_j \neq 0$ for some j. If we compute

$$\det(A^1, \ldots, x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n, \ldots, A^n) = \det(A^1, \ldots, 0, \ldots, A^n) = 0,$$

where 0 occurs in the j-th position. By multilinearity, all terms containing two identical columns A^k for $k \neq j$ vanish, and we get

$$\det(A^1, \ldots, x_1 A^1 + \cdots + x_j A^j + \cdots + x_n A^n, \ldots, A^n) = x_j \det(A^1, \ldots, A^n) = 0.$$

Since $x_j \neq 0$ and K is a field, we must have $\det(A^1, \ldots, A^n) = 0$.

Conversely, we show that if the columns $A^1, \ldots, A^n$ of A are linearly independent, then $\det(A^1, \ldots, A^n) \neq 0$. If the columns $A^1, \ldots, A^n$ of A are linearly independent, then they form a basis of K^n, and we can express the standard basis $(e_1, \ldots, e_n)$ of K^n in terms of $A^1, \ldots, A^n$. Thus, we have

$$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} b_{1\,1} & b_{1\,2} & \ldots & b_{1\,n} \\ b_{2\,1} & b_{2\,2} & \ldots & b_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n\,1} & b_{n\,2} & \ldots & b_{n\,n} \end{pmatrix} \begin{pmatrix} A^1 \\ A^2 \\ \vdots \\ A^n \end{pmatrix},$$

for some matrix $B = (b_{i\,j})$, and by Proposition 7.8, we get

$$\det(e_1, \ldots, e_n) = \det(B) \det(A^1, \ldots, A^n),$$

and since $\det(e_1, \ldots, e_n) = 1$, this implies that $\det(A^1, \ldots, A^n) \neq 0$ (and $\det(B) \neq 0$). For the second assertion, recall that the rank of a matrix is equal to the maximum number of linearly independent columns, and the conclusion is clear. $\square$

We now characterize when a system of linear equations of the form $Ax = b$ has a unique solution.

Proposition 7.12. *Given an $n \times n$-matrix A over a field K, the following properties hold:*

(1) For every column vector b, there is a unique column vector x such that $Ax = b$ iff the only solution to $Ax = 0$ is the trivial vector $x = 0$, iff $\det(A) \neq 0$.

(2) If $\det(A) \neq 0$, the unique solution of $Ax = b$ is given by the expressions

$$x_j = \frac{\det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n)}{\det(A^1, \ldots, A^{j-1}, A^j, A^{j+1}, \ldots, A^n)},$$

known as Cramer's rules.

(3) The system of linear equations $Ax = 0$ has a nonzero solution iff $\det(A) = 0$.

Proof. (1) Assume that $Ax = b$ has a single solution x_0, and assume that $Ay = 0$ with $y \neq 0$. Then,

$$A(x_0 + y) = Ax_0 + Ay = Ax_0 + 0 = b,$$

and $x_0 + y \neq x_0$ is another solution of $Ax = b$, contradicting the hypothesis that $Ax = b$ has a single solution x_0. Thus, $Ax = 0$ only has the trivial solution. Now assume that $Ax = 0$ only has the trivial solution. This means that the columns $A^1, \ldots, A^n$ of A are linearly independent, and by Proposition 7.11, we have $\det(A) \neq 0$. Finally, if $\det(A) \neq 0$, by Proposition 7.10, this means that A is invertible, and then for every b, $Ax = b$ is equivalent to $x = A^{-1}b$, which shows that $Ax = b$ has a single solution.

(2) Assume that $Ax = b$. If we compute

$$\det(A^1, \ldots, x_1A^1 + \cdots + x_jA^j + \cdots + x_nA^n, \ldots, A^n) = \det(A^1, \ldots, b, \ldots, A^n),$$

where b occurs in the j-th position, by multilinearity, all terms containing two identical columns A^k for $k \neq j$ vanish, and we get

$$x_j \det(A^1, \ldots, A^n) = \det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n),$$

for every j, $1 \leq j \leq n$. Since we assumed that $\det(A) = \det(A^1, \ldots, A^n) \neq 0$, we get the desired expression.

(3) Note that $Ax = 0$ has a nonzero solution iff $A^1, \ldots, A^n$ are linearly dependent (as observed in the proof of Proposition 7.11), which, by Proposition 7.11, is equivalent to $\det(A) = 0$. $\square$

As pleasing as Cramer's rules are, it is usually impractical to solve systems of linear equations using the above expressions. However, these formula imply an interesting fact, which is that the solution of the system $Ax = b$ are continuous in A and b. If we assume that the entries in A are continuous functions $a_{ij}(t)$ and the entries in b are are also continuous functions $b_j(t)$ of a real parameter t, since determinants are polynomial functions of their entries, the expressions

$$x_j(t) = \frac{\det(A^1, \ldots, A^{j-1}, b, A^{j+1}, \ldots, A^n)}{\det(A^1, \ldots, A^{j-1}, A^j, A^{j+1}, \ldots, A^n)}$$

are ratios of polynomials, and thus are also continuous as long as $\det(A(t))$ is nonzero. Similarly, if the functions $a_{ij}(t)$ and $b_j(t)$ are differentiable, so are the $x_j(t)$.

7.6 Determinant of a Linear Map

Given a vector space E of finite dimension n, given a basis $(u_1, \ldots, u_n)$ of E, for every linear map $f \colon E \to E$, if $M(f)$ is the matrix of f w.r.t. the basis $(u_1, \ldots, u_n)$, we can define $\det(f) = \det(M(f))$. If $(v_1, \ldots, v_n)$ is any other basis of E, and if P is the change of basis matrix, by Corollary 4.6, the matrix of f with respect to the basis $(v_1, \ldots, v_n)$ is $P^{-1}M(f)P$. By Proposition 7.9, we have

$$\det(P^{-1}M(f)P) = \det(P^{-1})\det(M(f))\det(P) = \det(P^{-1})\det(P)\det(M(f)) = \det(M(f)).$$

Thus, $\det(f)$ is indeed independent of the basis of E.

Definition 7.8. Given a vector space E of finite dimension, for any linear map $f\colon E \to E$, we define the *determinant* $\det(f)$ *of* f as the determinant $\det(M(f))$ of the matrix of f in any basis (since, from the discussion just before this definition, this determinant does not depend on the basis).

Then we have the following proposition.

Proposition 7.13. *Given any vector space E of finite dimension n, a linear map $f\colon E \to E$ is invertible iff $\det(f) \neq 0$.*

Proof. The linear map $f\colon E \to E$ is invertible iff its matrix $M(f)$ in any basis is invertible (by Proposition 4.2), iff $\det(M(f)) \neq 0$, by Proposition 7.10. □

Given a vector space of finite dimension n, it is easily seen that the set of bijective linear maps $f\colon E \to E$ such that $\det(f) = 1$ is a group under composition. This group is a subgroup of the general linear group $\mathbf{GL}(E)$. It is called the *special linear group (of E)*, and it is denoted by $\mathbf{SL}(E)$, or when $E = K^n$, by $\mathbf{SL}(n, K)$, or even by $\mathbf{SL}(n)$.

7.7 The Cayley–Hamilton Theorem

We next discuss an interesting and important application of Proposition 7.10, the *Cayley–Hamilton theorem*. The results of this section apply to matrices over any commutative ring K. First we need the concept of the characteristic polynomial of a matrix.

Definition 7.9. If K is any commutative ring, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, the *characteristic polynomial* $P_A(X)$ of A is the determinant

$$P_A(X) = \det(XI - A).$$

The characteristic polynomial $P_A(X)$ is a polynomial in $K[X]$, the ring of polynomials in the indeterminate X with coefficients in the ring K. For example, when $n = 2$, if

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$P_A(X) = \begin{vmatrix} X - a & -b \\ -c & X - d \end{vmatrix} = X^2 - (a + d)X + ad - bc.$$

We can substitute the matrix A for the variable X in the polynomial $P_A(X)$, obtaining a *matrix* P_A. If we write

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n,$$

then

$$P_A = A^n + c_1 A^{n-1} + \cdots + c_n I.$$

We have the following remarkable theorem.

Theorem 7.14. *(Cayley–Hamilton) If K is any commutative ring, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, if we let*

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n$$

be the characteristic polynomial of A, then

$$P_A = A^n + c_1 A^{n-1} + \cdots + c_n I = 0.$$

Proof. We can view the matrix $B = XI - A$ as a matrix with coefficients in the polynomial ring $K[X]$, and then we can form the matrix $\widetilde{B}$ which is the transpose of the matrix of cofactors of elements of B. Each entry in $\widetilde{B}$ is an $(n-1) \times (n-1)$ determinant, and thus a polynomial of degree a most $n-1$, so we can write $\widetilde{B}$ as

$$\widetilde{B} = X^{n-1} B_0 + X^{n-2} B_1 + \cdots + B_{n-1},$$

for some $n \times n$ matrices $B_0, \ldots, B_{n-1}$ with coefficients in K. For example, when $n = 2$, we have

$$B = \begin{pmatrix} X-a & -b \\ -c & X-d \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} X-d & b \\ c & X-a \end{pmatrix} = X \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -d & b \\ c & -a \end{pmatrix}.$$

By Proposition 7.10, we have

$$B\widetilde{B} = \det(B) I = P_A(X) I.$$

On the other hand, we have

$$B\widetilde{B} = (XI - A)(X^{n-1} B_0 + X^{n-2} B_1 + \cdots + X^{n-j-1} B_j + \cdots + B_{n-1}),$$

and by multiplying out the right-hand side, we get

$$B\widetilde{B} = X^n D_0 + X^{n-1} D_1 + \cdots + X^{n-j} D_j + \cdots + D_n,$$

with

$$\begin{aligned} D_0 &= B_0 \\ D_1 &= B_1 - AB_0 \\ &\vdots \\ D_j &= B_j - AB_{j-1} \\ &\vdots \\ D_{n-1} &= B_{n-1} - AB_{n-2} \\ D_n &= -AB_{n-1}. \end{aligned}$$

Since

$$P_A(X)I = (X^n + c_1X^{n-1} + \cdots + c_n)I,$$

the equality

$$X^nD_0 + X^{n-1}D_1 + \cdots + D_n = (X^n + c_1X^{n-1} + \cdots + c_n)I$$

is an equality between two matrices, so it requires that all corresponding entries are equal, and since these are polynomials, the coefficients of these polynomials must be identical, which is equivalent to the set of equations

$$\begin{aligned}
I &= B_0 \\
c_1I &= B_1 - AB_0 \\
&\vdots \\
c_jI &= B_j - AB_{j-1} \\
&\vdots \\
c_{n-1}I &= B_{n-1} - AB_{n-2} \\
c_nI &= -AB_{n-1},
\end{aligned}$$

for all j, with $1 \leq j \leq n-1$. If, as in the table below,

$$\begin{aligned}
A^n &= A^nB_0 \\
c_1A^{n-1} &= A^{n-1}(B_1 - AB_0) \\
&\vdots \\
c_jA^{n-j} &= A^{n-j}(B_j - AB_{j-1}) \\
&\vdots \\
c_{n-1}A &= A(B_{n-1} - AB_{n-2}) \\
c_nI &= -AB_{n-1},
\end{aligned}$$

we multiply the first equation by A^n, the last by I, and generally the $(j+1)$th by A^{n-j}, when we add up all these new equations, we see that the right-hand side adds up to 0, and we get our desired equation

$$A^n + c_1A^{n-1} + \cdots + c_nI = 0,$$

as claimed. $\square$

As a concrete example, when $n = 2$, the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

satisfies the equation

$$A^2 - (a+d)A + (ad - bc)I = 0.$$

Most readers will probably find the proof of Theorem 7.14 rather clever but very mysterious and unmotivated. The conceptual difficulty is that we really need to understand how polynomials in one variable "act" on vectors in terms of the matrix A. This can be done and yields a more "natural" proof. Actually, the reasoning is simpler and more general if we free ourselves from matrices and instead consider a finite-dimensional vector space E and some given linear map $f\colon E \to E$. Given any polynomial $p(X) = a_0X^n + a_1X^{n-1} + \cdots + a_n$ with coefficients in the field K, we define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0f^n + a_1f^{n-1} + \cdots + a_n\mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0f^n(u) + a_1f^{n-1}(u) + \cdots + a_nu,$$

for every vector $u \in E$. Then we define a new kind of scalar multiplication $\cdot\colon K[X] \times E \to E$ by polynomials as follows: for every polynomial $p(X) \in K[X]$, for every $u \in E$,

$$p(X) \cdot u = p(f)(u).$$

It is easy to verify that this is a "good action," which means that

$$\begin{aligned}
p \cdot (u + v) &= p \cdot u + p \cdot v\\
(p + q) \cdot u &= p \cdot u + q \cdot u\\
(pq) \cdot u &= p \cdot (q \cdot u)\\
1 \cdot u &= u,
\end{aligned}$$

for all $p, q \in K[X]$ and all $u, v \in E$. With this new scalar multiplication, E is a $K[X]$-module.

If $p = \lambda$ is just a scalar in K (a polynomial of degree 0), then

$$\lambda \cdot u = (\lambda\mathrm{id})(u) = \lambda u,$$

which means that K acts on E by scalar multiplication as before. If $p(X) = X$ (the monomial X), then

$$X \cdot u = f(u).$$

Now if we pick a basis $(e_1, \ldots, e_n)$ of E, if a polynomial $p(X) \in K[X]$ has the property that

$$p(X) \cdot e_i = 0, \quad i = 1, \ldots, n,$$

then this means that $p(f)(e_i) = 0$ for $i = 1, \ldots, n$, which means that the linear map $p(f)$ vanishes on E. We can also check, as we did in Section 7.2, that if A and B are two $n \times n$ matrices and if $(u_1, \ldots, u_n)$ are any n vectors, then

$$A \cdot \left(B \cdot \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}\right) = (AB) \cdot \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

This suggests the plan of attack for our second proof of the Cayley–Hamilton theorem. For simplicity, we prove the theorem for vector spaces over a field. The proof goes through for a free module over a commutative ring.

Theorem 7.15. *(Cayley–Hamilton) For every finite-dimensional vector space over a field K, for every linear map $f\colon E \to E$, for every basis $(e_1, \ldots, e_n)$, if A is the matrix over f over the basis $(e_1, \ldots, e_n)$ and if*

$$P_A(X) = X^n + c_1 X^{n-1} + \cdots + c_n$$

is the characteristic polynomial of A, then

$$P_A(f) = f^n + c_1 f^{n-1} + \cdots + c_n \mathrm{id} = 0.$$

Proof. Since the columns of A consist of the vector $f(e_j)$ expressed over the basis $(e_1, \ldots, e_n)$, we have

$$f(e_j) = \sum_{i=1}^{n} a_{i\,j} e_i, \quad 1 \leq j \leq n.$$

Using our action of $K[X]$ on E, the above equations can be expressed as

$$X \cdot e_j = \sum_{i=1}^{n} a_{i\,j} \cdot e_i, \quad 1 \leq j \leq n,$$

which yields

$$\sum_{i=1}^{j-1} -a_{i\,j} \cdot e_i + (X - a_{j\,j}) \cdot e_j + \sum_{i=j+1}^{n} -a_{i\,j} \cdot e_i = 0, \qquad 1 \leq j \leq n.$$

Observe that the transpose of the characteristic polynomial shows up, so the above system can be written as

$$\begin{pmatrix} X - a_{1\,1} & -a_{2\,1} & \cdots & -a_{n\,1} \\ -a_{1\,2} & X - a_{2\,2} & \cdots & -a_{n\,2} \\ \vdots & \vdots & \vdots & \vdots \\ -a_{1\,n} & -a_{2\,n} & \cdots & X - a_{n\,n} \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

If we let $B = XI - A^\top$, then as in the previous proof, if $\widetilde{B}$ is the transpose of the matrix of cofactors of B, we have

$$\widetilde{B}B = \det(B)I = \det(XI - A^\top)I = \det(XI - A)I = P_A I.$$

But since

$$B \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and since $\widetilde{B}$ is matrix whose entries are polynomials in $K[X]$, it makes sense to multiply on the left by $\widetilde{B}$ and we get

$$\widetilde{B} \cdot B \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = (\widetilde{B}B) \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = P_A I \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \widetilde{B} \cdot \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix};$$

that is,

$$P_A \cdot e_j = 0, \quad j = 1, \ldots, n,$$

which proves that $P_A(f) = 0$, as claimed. $\square$

If K is a field, then the characteristic polynomial of a linear map $f \colon E \to E$ is independent of the basis $(e_1, \ldots, e_n)$ chosen in E. To prove this, observe that the matrix of f over another basis will be of the form $P^{-1}AP$, for some inverible matrix P, and then

$$\begin{aligned} \det(XI - P^{-1}AP) &= \det(XP^{-1}IP - P^{-1}AP) \\ &= \det(P^{-1}(XI - A)P) \\ &= \det(P^{-1}) \det(XI - A) \det(P) \\ &= \det(XI - A). \end{aligned}$$

Therefore, the characteristic polynomial of a linear map is intrinsic to f, and it is denoted by P_f.

The zeros (roots) of the characteristic polynomial of a linear map f are called the *eigenvalues* of f. They play an important role in theory and applications. We will come back to this topic later on.

7.8 Permanents

Recall that the explicit formula for the determinant of an $n \times n$ matrix is

$$\det(A) = \sum_{\pi \in \mathfrak{S}_n} \epsilon(\pi) a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}.$$

If we drop the sign $\epsilon(\pi)$ of every permutation from the above formula, we obtain a quantity known as the *permanent*:

$$\operatorname{per}(A) = \sum_{\pi \in \mathfrak{S}_n} a_{\pi(1)\,1} \cdots a_{\pi(n)\,n}.$$

Permanents and determinants were investigated as early as 1812 by Cauchy. It is clear from the above definition that the permanent is a multilinear symmetric form. We also have

$$\operatorname{per}(A) = \operatorname{per}(A^{\top}),$$

and the following unsigned version of the Laplace expansion formula:

$$\mathrm{per}(A) = a_{i\,1}\mathrm{per}(A_{i\,1}) + \cdots + a_{i\,j}\mathrm{per}(A_{i\,j}) + \cdots + a_{i\,n}\mathrm{per}(A_{i\,n}),$$

for $i = 1, \ldots, n$. However, unlike determinants which have a clear geometric interpretation as signed volumes, permanents do not have any natural geometric interpretation. Furthermore, determinants can be evaluated efficiently, for example using the conversion to row reduced echelon form, but computing the permanent is hard.

Permanents turn out to have various combinatorial interpretations. One of these is in terms of perfect matchings of bipartite graphs which we now discuss.

See Definition 20.5 for the definition of an undirected graph. A *bipartite* (undirected) graph $G = (V, E)$ is a graph whose set of nodes V can be partitioned into two nonempty disjoint subsets V_1 and V_2, such that every edge $e \in E$ has one endpoint in V_1 and one endpoint in V_2.

An example of a bipartite graph with 14 nodes is shown in Figure 7.3; its nodes are partitioned into the two sets $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and $\{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$.

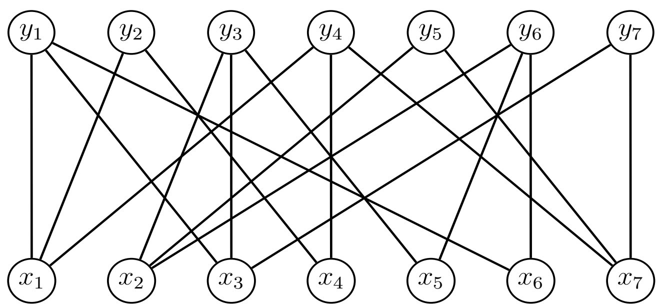


Figure 7.3: A bipartite graph G.

A *matching* in a graph $G = (V, E)$ (bipartite or not) is a set M of pairwise non-adjacent edges, which means that no two edges in M share a common vertex. A *perfect matching* is a matching such that every node in V is incident to some edge in the matching M (every node in V is an endpoint of some edge in M). Figure 7.4 shows a perfect matching (in red) in the bipartite graph G.

Obviously, a perfect matching in a bipartite graph can exist only if its set of nodes has a partition in two blocks of equal size, say $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_m\}$. Then there is a bijection between perfect matchings and bijections $\pi \colon \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_m\}$ such that $\pi(x_i) = y_j$ iff there is an edge between x_i and y_j.

Now every bipartite graph G with a partition of its nodes into two sets of equal size as above is represented by an $m \times m$ matrix $A = (a_{ij})$ such that $a_{ij} = 1$ iff there is an edge between x_i and y_j, and $a_{ij} = 0$ otherwise. Using the interpretation of perfect matchings as

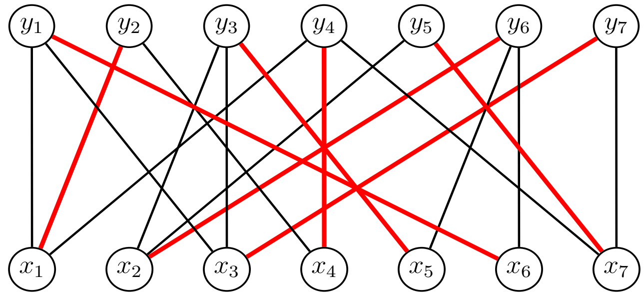


Figure 7.4: A perfect matching in the bipartite graph G.

bijections $\pi\colon \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_m\}$, we see that *the permanent* $\mathrm{per}(A)$ *of the* $(0,1)$-*matrix* A *representing the bipartite graph* G *counts the number of perfect matchings in* G.

In a famous paper published in 1979, Leslie Valiant proves that computing the permanent is a #P-complete problem. Such problems are suspected to be intractable. It is known that if a polynomial-time algorithm existed to solve a #P-complete problem, then we would have $P = NP$, which is believed to be very unlikely.

Another combinatorial interpretation of the permanent can be given in terms of systems of distinct representatives. Given a finite set S, let $(A_1, \ldots, A_n)$ be any sequence of nonempty subsets of S (not necessarily distinct). A *system of distinct representatives* (for short *SDR*) of the sets $A_1, \ldots, A_n$ is a sequence of n distinct elements $(a_1, \ldots, a_n)$, with $a_i \in A_i$ for $i = 1, \ldots, n$. The number of SDR's of a sequence of sets plays an important role in combinatorics. Now, if $S = \{1, 2, \ldots, n\}$ and if we associate to any sequence $(A_1, \ldots, A_n)$ of nonempty subsets of S the matrix $A = (a_{ij})$ defined such that $a_{ij} = 1$ if $j \in A_i$ and $a_{ij} = 0$ otherwise, then *the permanent* $\mathrm{per}(A)$ *counts the number of SDR's of the sets* $A_1, \ldots, A_n$.

This interpretation of permanents in terms of SDR's can be used to prove bounds for the permanents of various classes of matrices. Interested readers are referred to van Lint and Wilson [180] (Chapters 11 and 12). In particular, a proof of a theorem known as *Van der Waerden conjecture* is given in Chapter 12. This theorem states that for any $n \times n$ matrix A with nonnegative entries in which all row-sums and column-sums are 1 (doubly stochastic matrices), we have

$$\mathrm{per}(A) \geq \frac{n!}{n^n},$$

with equality for the matrix in which all entries are equal to $1/n$.

7.9 Summary

The main concepts and results of this chapter are listed below:

- *Permutations*, *transpositions*, *basics transpositions*.
- Every permutation can be written as a composition of permutations.
- The *parity* of the number of transpositions involved in any decomposition of a permutation σ is an invariant; it is the *signature* $\epsilon(\sigma)$ of the permutation σ.
- *Multilinear maps* (also called *n-linear maps*); *bilinear maps*.
- *Symmetric* and *alternating* multilinear maps.
- A basic property of alternating multilinear maps (Lemma 7.4) and the introduction of the formula expressing a determinant.
- Definition of a *determinant* as a multlinear alternating map $D\colon \mathrm{M}_n(K) \to K$ such that $D(I) = 1$.
- We define the set of algorithms $\mathcal{D}_n$, to compute the determinant of an $n \times n$ matrix.
- *Laplace expansion according to the ith row*; *cofactors*.
- We prove that the algorithms in $\mathcal{D}_n$ compute determinants (Lemma 7.5).
- We prove that all algorithms in $\mathcal{D}_n$ compute the same determinant (Theorem 7.6).
- We give an interpretation of determinants as *signed volumes*.
- We prove that $\det(A) = \det(A^\top)$.
- We prove that $\det(AB) = \det(A)\det(B)$.
- The *adjugate* $\widetilde{A}$ of a matrix A.
- Formula for the inverse in terms of the adjugate.
- A matrix A is invertible iff $\det(A) \neq 0$.
- Solving linear equations using *Cramer's rules*.
- Determinant of a linear map.
- The *characteristic polynomial* of a matrix.
- The *Cayley–Hamilton theorem*.
- The action of the polynomial ring induced by a linear map on a vector space.
- *Permanents*.
- Permanents count the number of perfect matchings in bipartite graphs.

- Computing the permanent is a #P-perfect problem (L. Valiant).
- Permanents count the number of SDRs of sequences of subsets of a given set.

7.10 Further Readings

Thorough expositions of the material covered in Chapter 3–6 and 7 can be found in Strang [170, 169], Lax [113], Lang [109], Artin [7], Mac Lane and Birkhoff [118], Hoffman and Kunze [94], Dummit and Foote [54], Bourbaki [25, 26], Van Der Waerden [179], Serre [156], Horn and Johnson [95], and Bertin [15]. These notions of linear algebra are nicely put to use in classical geometry, see Berger [11, 12], Tisseron [175] and Dieudonné [49].

7.11 Problems

Problem 7.1. Prove that every transposition can be written as a product of basic transpositions.

Problem 7.2. (1) Given two vectors in $\mathbb{R}^2$ of coordinates (c_1-a_1, c_2-a_2) and (b_1-a_1, b_2-a_2), prove that they are linearly dependent iff

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

(2) Given three vectors in $\mathbb{R}^3$ of coordinates $(d_1-a_1, d_2-a_2, d_3-a_3)$, $(c_1-a_1, c_2-a_2, c_3-a_3)$, and $(b_1-a_1, b_2-a_2, b_3-a_3)$, prove that they are linearly dependent iff

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ 1 & 1 & 1 & 1 \end{vmatrix} = 0.$$

Problem 7.3. Let A be the $(m+n)\times(m+n)$ block matrix (over any field K) given by

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_4 \end{pmatrix},$$

where A_1 is an $m\times m$ matrix, A_2 is an $m\times n$ matrix, and A_4 is an $n\times n$ matrix. Prove that $\det(A) = \det(A_1)\det(A_4)$.

Use the above result to prove that if A is an upper triangular $n\times n$ matrix, then $\det(A) = a_{11}a_{22}\cdots a_{nn}$.

Problem 7.4. Prove that if $n \geq 3$, then

$$\det \begin{pmatrix} 1+x_1y_1 & 1+x_1y_2 & \cdots & 1+x_1y_n \\ 1+x_2y_1 & 1+x_2y_2 & \cdots & 1+x_2y_n \\ \vdots & \vdots & \vdots & \vdots \\ 1+x_ny_1 & 1+x_ny_2 & \cdots & 1+x_ny_n \end{pmatrix} = 0.$$

Problem 7.5. Prove that

$$\begin{vmatrix} 1 & 4 & 9 & 16 \\ 4 & 9 & 16 & 25 \\ 9 & 16 & 25 & 36 \\ 16 & 25 & 36 & 49 \end{vmatrix} = 0.$$

Problem 7.6. Consider the $n \times n$ symmetric matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \ldots & 0 & 0 \\ 2 & 5 & 2 & 0 & \ldots & 0 & 0 \\ 0 & 2 & 5 & 2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 2 & 5 & 2 & 0 \\ 0 & 0 & \ldots & 0 & 2 & 5 & 2 \\ 0 & 0 & \ldots & 0 & 0 & 2 & 5 \end{pmatrix}.$$

(1) Find an upper-triangular matrix R such that $A = R^\top R$.

(2) Prove that $\det(A) = 1$.

(3) Consider the sequence

$$\begin{aligned} p_0(\lambda) &= 1 \\ p_1(\lambda) &= 1 - \lambda \\ p_k(\lambda) &= (5-\lambda)p_{k-1}(\lambda) - 4p_{k-2}(\lambda) \quad 2 \leq k \leq n. \end{aligned}$$

Prove that

$$\det(A - \lambda I) = p_n(\lambda).$$

Remark: It can be shown that $p_n(\lambda)$ has n distinct (real) roots and that the roots of $p_k(\lambda)$ separate the roots of $p_{k+1}(\lambda)$.

Problem 7.7. Let B be the $n \times n$ matrix ($n \geq 3$) given by

$$B = \begin{pmatrix} 1 & -1 & -1 & -1 & \cdots & -1 & -1 \\ 1 & -1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & -1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & -1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & -1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & -1 \end{pmatrix}.$$

Prove that

$$\det(B) = (-1)^n(n-2)2^{n-1}.$$

Problem 7.8. Given a field K (say $K = \mathbb{R}$ or $K = \mathbb{C}$), given any two polynomials $p(X), q(X) \in K[X]$, we says that $q(X)$ *divides* $p(X)$ (and that $p(X)$ *is a multiple of* $q(X)$) iff there is some polynomial $s(X) \in K[X]$ such that

$$p(X) = q(X)s(X).$$

In this case we say that $q(X)$ *is a factor of* $p(X)$, and if $q(X)$ has degree at least one, we say that $q(X)$ *is a nontrivial factor of* $p(X)$.

Let $f(X)$ and $g(X)$ be two polynomials in $K[X]$ with

$$f(X) = a_0X^m + a_1X^{m-1} + \cdots + a_m$$

of degree $m \geq 1$ and

$$g(X) = b_0X^n + b_1X^{n-1} + \cdots + b_n$$

of degree $n \geq 1$ (with $a_0, b_0 \neq 0$).

You will need the following result which you need not prove:

Two polynomials $f(X)$ *and* $g(X)$ *with* $\deg(f) = m \geq 1$ *and* $\deg(g) = n \geq 1$ *have some common nontrivial factor iff there exist two nonzero polynomials* $p(X)$ *and* $q(X)$ *such that*

$$fp = gq,$$

with $\deg(p) \leq n-1$ *and* $\deg(q) \leq m-1$.

(1) Let $\mathcal{P}_m$ denote the vector space of all polynomials in $K[X]$ of degree at most $m-1$, and let $T\colon \mathcal{P}_n \times \mathcal{P}_m \to \mathcal{P}_{m+n}$ be the map given by

$$T(p,q) = fp + gq, \quad p \in \mathcal{P}_n,\ q \in \mathcal{P}_m,$$

where f and g are some fixed polynomials of degree $m \geq 1$ and $n \geq 1$.

Prove that the map T is linear.

(2) Prove that T is not injective iff f and g have a common nontrivial factor.

(3) Prove that f and g have a nontrivial common factor iff $R(f,g) = 0$, where $R(f,g)$ is the determinant given by

$$R(f,g) = \begin{vmatrix}
a_0 & a_1 & \cdots & \cdots & a_m & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & a_0 & a_1 & \cdots & \cdots & a_m & 0 & \cdots & \cdots & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & a_0 & a_1 & \cdots & \cdots & a_m \\
b_0 & b_1 & \cdots & \cdots & \cdots & \cdots & \cdots & b_n & 0 & \cdots & 0 \\
0 & b_0 & b_1 & \cdots & \cdots & \cdots & \cdots & \cdots & b_n & 0 & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & b_0 & b_1 & \cdots & \cdots & \cdots & \cdots & \cdots & b_n
\end{vmatrix}.$$

The above determinant is called the *resultant of f and g*.

Note that the matrix of the resultant is an $(n+m) \times (n+m)$ matrix, with the first row (involving the a_is) occurring n times, each time shifted over to the right by one column, and the $(n+1)$th row (involving the b_js) occurring m times, each time shifted over to the right by one column.

Hint. Express the matrix of T over some suitable basis.

(4) Compute the resultant in the following three cases:

(a) $m = n = 1$, and write $f(X) = aX + b$ and $g(X) = cX + d$.

(b) $m = 1$ and $n \geq 2$ arbitrary.

(c) $f(X) = aX^2 + bX + c$ and $g(X) = 2aX + b$.

(5) Compute the resultant of $f(X) = X^3 + pX + q$ and $g(X) = 3X^2 + p$, and

$$\begin{aligned} f(X) &= a_0X^2 + a_1X + a_2 \\ g(X) &= b_0X^2 + b_1X + b_2. \end{aligned}$$

In the second case, you should get

$$4R(f, g) = (2a_0b_2 - a_1b_1 + 2a_2b_0)^2 - (4a_0a_2 - a_1^2)(4b_0b_2 - b_1^2).$$

Problem 7.9. Let A, B, C, D be $n \times n$ real or complex matrices.

(1) Prove that if A is invertible and if $AC = CA$, then

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(AD - CB).$$

(2) Prove that if H is an $n \times n$ Hadamard matrix ($n \geq 2$), then $|\det(H)| = n^{n/2}$.

(3) Prove that if H is an $n \times n$ Hadamard matrix ($n \geq 2$), then

$$\det \begin{pmatrix} H & H \\ H & -H \end{pmatrix} = (2n)^n.$$

Problem 7.10. Compute the product of the following determinants

$$\begin{vmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{vmatrix} \begin{vmatrix} x & -y & -z & -t \\ y & x & -t & z \\ z & t & x & -y \\ t & -z & y & x \end{vmatrix}$$

to prove the following identity (due to Euler):

$$\begin{aligned} (a^2 + b^2 + c^2 + d^2)(x^2 + y^2 + z^2 + t^2) &= (ax + by + cz + dt)^2 + (ay - bx + ct - dz)^2 \\ &\quad + (az - bt - cx + dy)^2 + (at + bz - cy - dx)^2. \end{aligned}$$

Problem 7.11. Let A be an $n \times n$ matrix with integer entries. Prove that A^{-1} exists and has integer entries if and only if $\det(A) = \pm 1$.

Problem 7.12. Let A be an $n \times n$ real or complex matrix.

(1) Prove that if $A^\top = -A$ (A is *skew-symmetric*) and if n is odd, then $\det(A) = 0$.

(2) Prove that

$$\begin{vmatrix} 0 & a & b & c \\ -a & 0 & d & e \\ -b & -d & 0 & f \\ -c & -e & -f & 0 \end{vmatrix} = (af - be + dc)^2.$$

Problem 7.13. A *Cauchy matrix* is a matrix of the form

$$\begin{pmatrix} \dfrac{1}{\lambda_1 - \sigma_1} & \dfrac{1}{\lambda_1 - \sigma_2} & \cdots & \dfrac{1}{\lambda_1 - \sigma_n} \\ \dfrac{1}{\lambda_2 - \sigma_1} & \dfrac{1}{\lambda_2 - \sigma_2} & \cdots & \dfrac{1}{\lambda_2 - \sigma_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_n - \sigma_1} & \dfrac{1}{\lambda_n - \sigma_2} & \cdots & \dfrac{1}{\lambda_n - \sigma_n} \end{pmatrix}$$

where $\lambda_i \neq \sigma_j$, for all i, j, with $1 \leq i, j \leq n$. Prove that the determinant C_n of a Cauchy matrix as above is given by

$$C_n = \frac{\prod_{i=2}^{n} \prod_{j=1}^{i-1} (\lambda_i - \lambda_j)(\sigma_j - \sigma_i)}{\prod_{i=1}^{n} \prod_{j=1}^{n} (\lambda_i - \sigma_j)}.$$

Problem 7.14. Let $(\alpha_1, \ldots, \alpha_{m+1})$ be a sequence of pairwise distinct scalars in $\mathbb{R}$ and let $(\beta_1, \ldots, \beta_{m+1})$ be any sequence of scalars in $\mathbb{R}$, not necessarily distinct.

(1) Prove that there is a unique polynomial P of degree at most m such that

$$P(\alpha_i) = \beta_i, \quad 1 \leq i \leq m+1.$$

Hint. Remember Vandermonde!

(2) Let $L_i(X)$ be the polynomial of degree m given by

$$L_i(X) = \frac{(X - \alpha_1) \cdots (X - \alpha_{i-1})(X - \alpha_{i+1}) \cdots (X - \alpha_{m+1})}{(\alpha_i - \alpha_1) \cdots (\alpha_i - \alpha_{i-1})(\alpha_i - \alpha_{i+1}) \cdots (\alpha_i - \alpha_{m+1})}, \quad 1 \leq i \leq m+1.$$

The polynomials $L_i(X)$ are known as *Lagrange polynomial interpolants*. Prove that

$$L_i(\alpha_j) = \delta_{i\,j} \quad 1 \leq i, j \leq m+1.$$

Prove that

$$P(X) = \beta_1 L_1(X) + \cdots + \beta_{m+1} L_{m+1}(X)$$

is the unique polynomial of degree at most m such that

$$P(\alpha_i) = \beta_i, \quad 1 \leq i \leq m+1.$$

(3) Prove that $L_1(X), \ldots, L_{m+1}(X)$ are linearly independent, and that they form a basis of all polynomials of degree at most m.

How is 1 (the constant polynomial 1) expressed over the basis $(L_1(X), \ldots, L_{m+1}(X))$?

Give the expression of every polynomial $P(X)$ of degree at most m over the basis $(L_1(X), \ldots, L_{m+1}(X))$.

(4) Prove that the dual basis $(L_1^*, \ldots, L_{m+1}^*)$ of the basis $(L_1(X), \ldots, L_{m+1}(X))$ consists of the linear forms L_i^* given by

$$L_i^*(P) = P(\alpha_i),$$

for every polynomial P of degree at most m; this is simply *evaluation at* α_i.

Chapter 8

Gaussian Elimination, LU-Factorization, Cholesky Factorization, Reduced Row Echelon Form

In this chapter we assume that all vector spaces are over the field $\mathbb{R}$. All results that do not rely on the ordering on $\mathbb{R}$ or on taking square roots hold for arbitrary fields.

8.1 Motivating Example: Curve Interpolation

Curve interpolation is a problem that arises frequently in computer graphics and in robotics (path planning). There are many ways of tackling this problem and in this section we will describe a solution using *cubic splines*. Such splines consist of cubic Bézier curves. They are often used because they are cheap to implement and give more flexibility than quadratic Bézier curves.

A *cubic Bézier curve* $C(t)$ (in $\mathbb{R}^2$ or $\mathbb{R}^3$) is specified by a list of four *control points* (b_0, b_1, b_2, b_3) and is given parametrically by the equation

$$C(t) = (1-t)^3\, b_0 + 3(1-t)^2 t\, b_1 + 3(1-t)t^2\, b_2 + t^3\, b_3.$$

Clearly, $C(0) = b_0$, $C(1) = b_3$, and for $t \in [0,1]$, the point $C(t)$ belongs to the convex hull of the control points b_0, b_1, b_2, b_3. The polynomials

$$(1-t)^3, \quad 3(1-t)^2 t, \quad 3(1-t)t^2, \quad t^3$$

are the *Bernstein polynomials* of degree 3.

Typically, we are only interested in the curve segment corresponding to the values of t in the interval $[0,1]$. Still, the placement of the control points drastically affects the shape of the curve segment, which can even have a self-intersection; See Figures 8.1, 8.2, 8.3 illustrating various configurations.

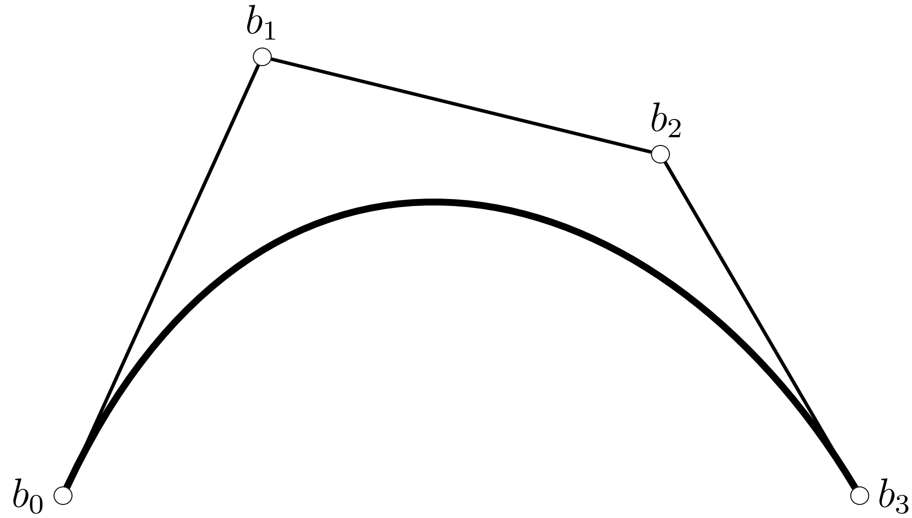


Figure 8.1: A "standard" Bézier curve.

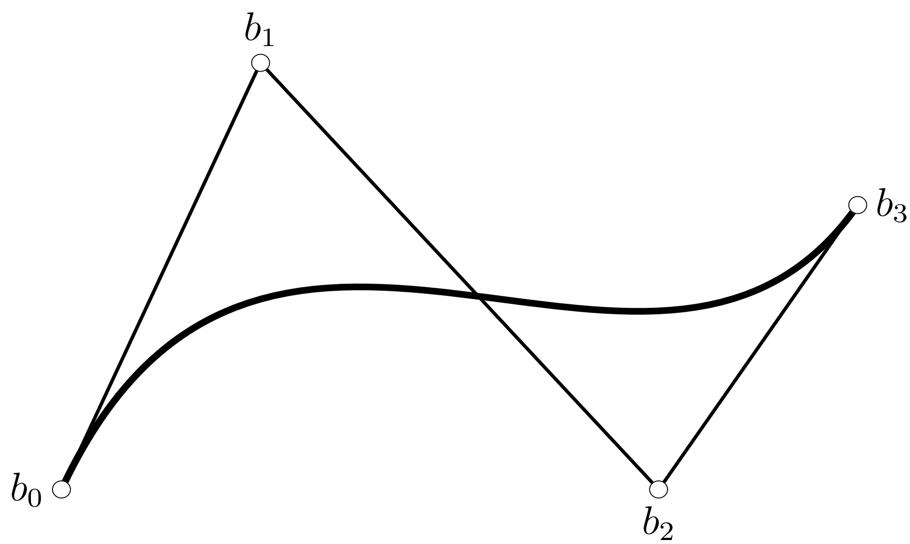


Figure 8.2: A Bézier curve with an inflection point.

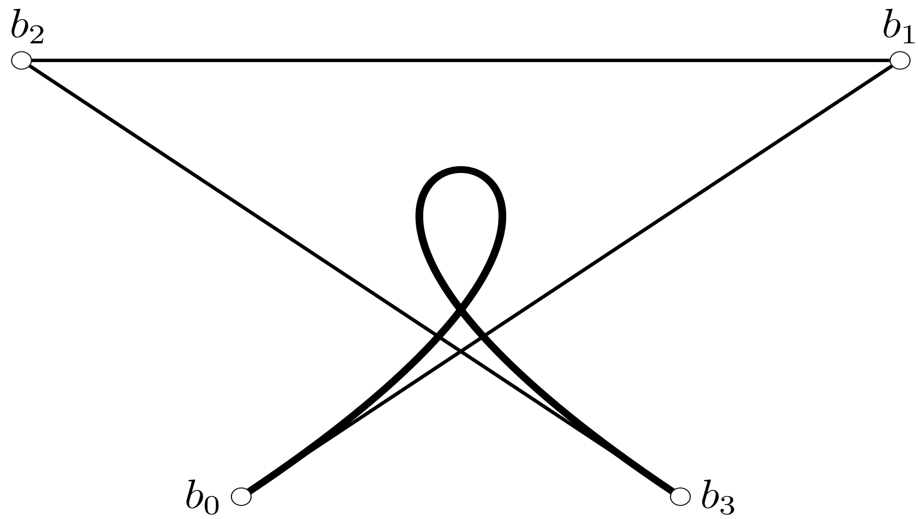


Figure 8.3: A self-intersecting Bézier curve.

Interpolation problems require finding curves passing through some given data points and possibly satisfying some extra constraints.

A *Bézier spline curve* F is a curve which is made up of curve segments which are Bézier curves, say $C_1, \ldots, C_m$ $(m \geq 2)$. We will assume that F defined on $[0, m]$, so that for $i = 1, \ldots, m$,

$$F(t) = C_i(t - i + 1), \quad i - 1 \leq t \leq i.$$

Typically, some smoothness is required between any two junction points, that is, between any two points $C_i(1)$ and $C_{i+1}(0)$, for $i = 1, \ldots, m-1$. We require that $C_i(1) = C_{i+1}(0)$ (C^0-*continuity*), and typically that the derivatives of C_i at 1 and of C_{i+1} at 0 agree up to second order derivatives. This is called C^2-*continuity*, and it ensures that the tangents agree as well as the curvatures.

There are a number of interpolation problems, and we consider one of the most common problems which can be stated as follows:

Problem: Given $N+1$ data points $x_0, \ldots, x_N$, find a C^2 cubic spline curve F such that $F(i) = x_i$ for all i, $0 \leq i \leq N$ $(N \geq 2)$.

A way to solve this problem is to find $N+3$ auxiliary points $d_{-1}, \ldots, d_{N+1}$, called *de Boor control points*, from which N Bézier curves can be found. Actually,

$$d_{-1} = x_0 \quad \text{and} \quad d_{N+1} = x_N$$

so we only need to find $N+1$ points $d_0, \ldots, d_N$.

It turns out that the C^2-continuity constraints on the N Bézier curves yield only $N-1$ equations, so d_0 and d_N can be chosen arbitrarily. In practice, d_0 and d_N are chosen according to various *end conditions*, such as prescribed velocities at x_0 and x_N. For the time being, we will assume that d_0 and d_N are given.

Figure 8.4 illustrates an interpolation problem involving $N+1 = 7+1 = 8$ data points. The control points d_0 and d_7 were chosen arbitrarily.

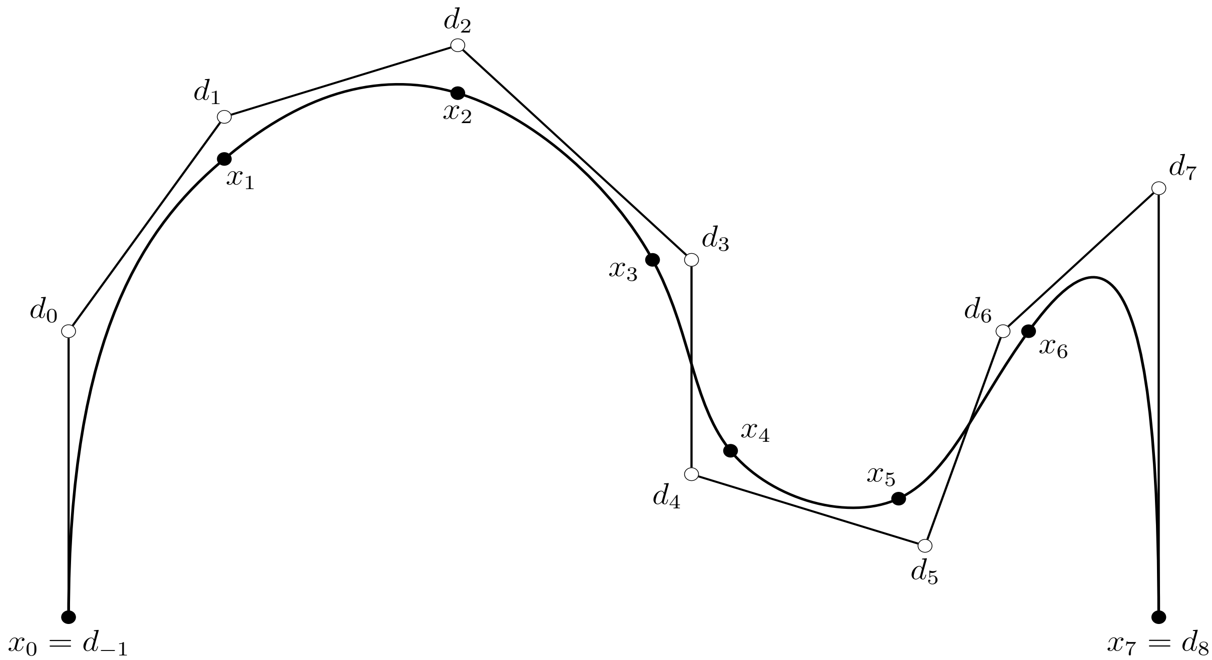


Figure 8.4: A C^2 cubic interpolation spline curve passing through the points $x_0, x_1,\ x_2, x_3,\ x_4, x_5,\ x_6, x_7$.

It can be shown that $d_1, \ldots, d_{N-1}$ are given by the linear system

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix} = \begin{pmatrix} 6x_1 - \frac{3}{2}d_0 \\ 6x_2 \\ \vdots \\ 6x_{N-2} \\ 6x_{N-1} - \frac{3}{2}d_N \end{pmatrix}.$$

We will show later that the above matrix is invertible because it is strictly diagonally dominant.

Once the above system is solved, the Bézier cubics $C_1, \ldots, C_N$ are determined as follows (we assume $N \geq 2$): For $2 \leq i \leq N-1$, the control points $(b_0^i, b_1^i, b_2^i, b_3^i)$ of C_i are given by

$$\begin{aligned}
b_0^i &= x_{i-1} \\
b_1^i &= \frac{2}{3}d_{i-1} + \frac{1}{3}d_i \\
b_2^i &= \frac{1}{3}d_{i-1} + \frac{2}{3}d_i \\
b_3^i &= x_i.
\end{aligned}$$

The control points $(b_0^1, b_1^1, b_2^1, b_3^1)$ of C_1 are given by

$$\begin{aligned}
b_0^1 &= x_0 \\
b_1^1 &= d_0 \\
b_2^1 &= \frac{1}{2}d_0 + \frac{1}{2}d_1 \\
b_3^1 &= x_1,
\end{aligned}$$

and the control points $(b_0^N, b_1^N, b_2^N, b_3^N)$ of C_N are given by

$$\begin{aligned}
b_0^N &= x_{N-1} \\
b_1^N &= \frac{1}{2}d_{N-1} + \frac{1}{2}d_N \\
b_2^N &= d_N \\
b_3^N &= x_N.
\end{aligned}$$

Figure 8.5 illustrates this process spline interpolation for $N = 7$.

We will now describe various methods for solving linear systems. Since the matrix of the above system is tridiagonal, there are specialized methods which are more efficient than the general methods. We will discuss a few of these methods.

8.2 Gaussian Elimination

Let A be an $n \times n$ matrix, let $b \in \mathbb{R}^n$ be an n-dimensional vector and assume that A is invertible. Our goal is to solve the system $Ax = b$. Since A is assumed to be invertible, we know that this system has a unique solution $x = A^{-1}b$. Experience shows that two counter-intuitive facts are revealed:

(1) One should avoid computing the inverse A^{-1} of A explicitly. This is inefficient since it would amount to solving the n linear systems $Au^{(j)} = e_j$ for $j = 1, \ldots, n$, where $e_j = (0, \ldots, 1, \ldots, 0)$ is the jth canonical basis vector of $\mathbb{R}^n$ (with a 1 is the jth slot). By doing so, we would replace the resolution of a single system by the resolution of n systems, and we would still have to multiply A^{-1} by b.

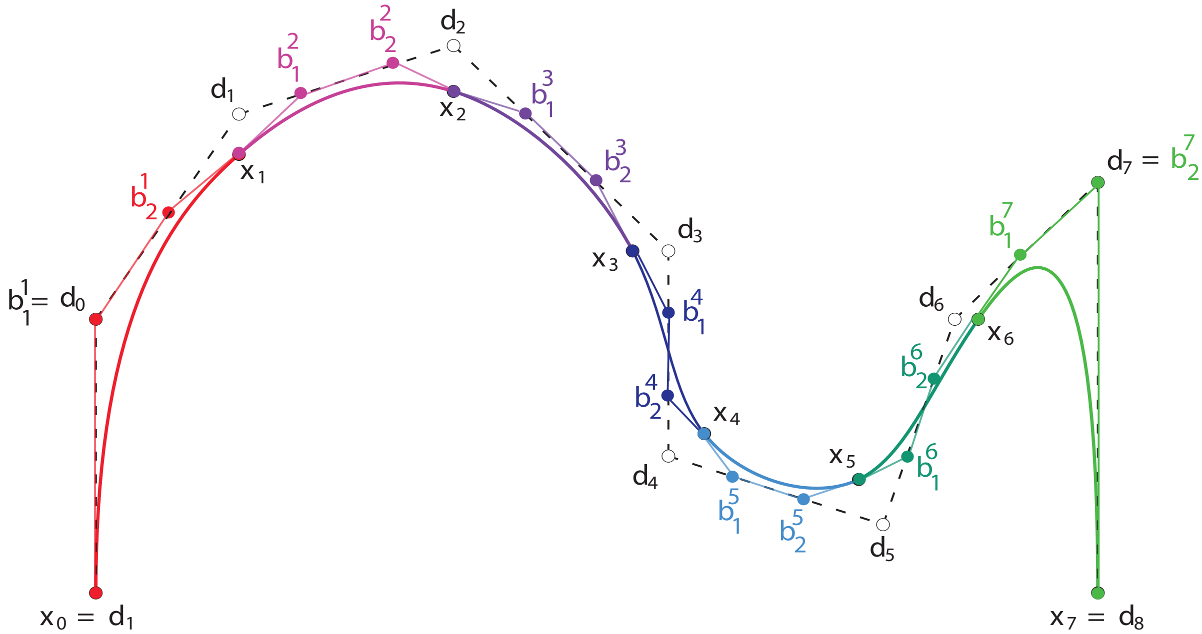


Figure 8.5: A C^2 cubic interpolation of $x_0, x_1,\ x_2, x_3,\ x_4, x_5,\ x_6, x_7$ with associated color coded Bézier cubics.

(2) One does not solve (large) linear systems by computing determinants (using Cramer's formulae) since this method requires a number of additions (resp. multiplications) proportional to $(n+1)!$ (resp. $(n+2)!$).

The key idea on which most direct methods (as opposed to iterative methods, that look for an approximation of the solution) are based is that if A is an upper-triangular matrix, which means that $a_{ij} = 0$ for $1 \leq j < i \leq n$ (resp. lower-triangular, which means that $a_{ij} = 0$ for $1 \leq i < j \leq n$), then computing the solution x is trivial. Indeed, say A is an upper-triangular matrix

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \cdots & a_{1\,n-2} & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & \cdots & a_{2\,n-2} & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & \cdots & 0 & 0 & a_{n\,n} \end{pmatrix}.$$

Then $\det(A) = a_{1\,1}a_{2\,2}\cdots a_{n\,n} \neq 0$, which implies that $a_{i\,i} \neq 0$ for $i = 1, \ldots, n$, and we can solve the system $Ax = b$ from bottom-up by *back-substitution*. That is, first we compute

x_n from the last equation, next plug this value of x_n into the next to the last equation and compute x_{n-1} from it, *etc.* This yields

$$\begin{aligned} x_n &= a_{n\,n}^{-1} b_n \\ x_{n-1} &= a_{n-1\,n-1}^{-1}(b_{n-1} - a_{n-1\,n} x_n) \\ &\vdots \\ x_1 &= a_{1\,1}^{-1}(b_1 - a_{1\,2}x_2 - \cdots - a_{1\,n}x_n). \end{aligned}$$

Note that the use of determinants can be avoided to prove that if A is invertible then $a_{i\,i} \neq 0$ for $i = 1, \ldots, n$. Indeed, it can be shown directly (by induction) that an upper (or lower) triangular matrix is invertible iff all its diagonal entries are nonzero.

If A is lower-triangular, we solve the system from top-down by *forward-substitution*.

Thus, what we need is a method for transforming a matrix to an equivalent one in upper-triangular form. This can be done by *elimination*. Let us illustrate this method on the following example:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ 4x & - & 6y & & & = & -2 \\ -2x & + & 7y & + & 2z & = & 9. \end{array}$$

We can eliminate the variable x from the second and the third equation as follows: Subtract twice the first equation from the second and add the first equation to the third. We get the new system

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & 8y & + & 3z & = & 14. \end{array}$$

This time we can eliminate the variable y from the third equation by adding the second equation to the third:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & & & z & = & 2. \end{array}$$

This last system is upper-triangular. Using back-substitution, we find the solution: $z = 2$, $y = 1$, $x = 1$.

Observe that we have performed only *row operations*. The general method is to iteratively eliminate variables using simple row operations (namely, adding or subtracting a multiple of a row to another row of the matrix) while simultaneously applying these operations to the vector b, to obtain a system, $MAx = Mb$, where MA is upper-triangular. Such a method is called *Gaussian elimination*. However, one extra twist is needed for the method to work in all cases: It may be necessary to permute rows, as illustrated by the following example:

$$\begin{array}{rcrcrl} x & + & y & + & z & = 1 \\ x & + & y & + & 3z & = 1 \\ 2x & + & 5y & + & 8z & = 1. \end{array}$$

In order to eliminate x from the second and third row, we subtract the first row from the second and we subtract twice the first row from the third:

$$\begin{array}{ccccccc} x & + & y & + & z & = 1 \\ & & & & 2z & = 0 \\ & & 3y & + & 6z & = -1. \end{array}$$

Now the trouble is that y does not occur in the second row; so, we can't eliminate y from the third row by adding or subtracting a multiple of the second row to it. The remedy is simple: Permute the second and the third row! We get the system:

$$\begin{array}{ccccccc} x & + & y & + & z & = 1 \\ & & 3y & + & 6z & = -1 \\ & & & & 2z & = 0, \end{array}$$

which is already in triangular form. Another example where some permutations are needed is:

$$\begin{array}{ccccccc} & & & & z & = & 1 \\ -2x & + & 7y & + & 2z & = & 1 \\ 4x & - & 6y & & & = & -1. \end{array}$$

First we permute the first and the second row, obtaining

$$\begin{array}{ccccccc} -2x & + & 7y & + & 2z & = & 1 \\ & & & & z & = & 1 \\ 4x & - & 6y & & & = & -1, \end{array}$$

and then we add twice the first row to the third, obtaining:

$$\begin{array}{ccccccc} -2x & + & 7y & + & 2z & = & 1 \\ & & & & z & = & 1 \\ & & 8y & + & 4z & = & 1. \end{array}$$

Again we permute the second and the third row, getting

$$\begin{array}{ccccccc} -2x & + & 7y & + & 2z & = & 1 \\ & & 8y & + & 4z & = & 1 \\ & & & & z & = & 1, \end{array}$$

an upper-triangular system. Of course, in this example, z is already solved and we could have eliminated it first, but for the general method, we need to proceed in a systematic fashion.

We now describe the method of *Gaussian elimination* applied to a linear system $Ax = b$, where A is assumed to be invertible. We use the variable k to keep track of the stages of elimination. Initially, $k = 1$.

(1) The first step is to pick some nonzero entry $a_{i\,1}$ in the first column of A. Such an entry must exist, since A is invertible (otherwise, the first column of A would be the zero vector, and the columns of A would not be linearly independent. Equivalently, we would have $\det(A) = 0$). The actual choice of such an element has some impact on the numerical stability of the method, but this will be examined later. For the time being, we assume that some arbitrary choice is made. This chosen element is called the *pivot* of the elimination step and is denoted π_1 (so, in this first step, $\pi_1 = a_{i\,1}$).

(2) Next we permute the row (i) corresponding to the pivot with the first row. Such a step is called *pivoting*. So after this permutation, the first element of the first row is nonzero.

(3) We now eliminate the variable x_1 from all rows except the first by adding suitable multiples of the first row to these rows. More precisely we add $-a_{i\,1}/\pi_1$ times the first row to the ith row for $i = 2, \ldots, n$. At the end of this step, all entries in the first column are zero except the first.

(4) Increment k by 1. If $k = n$, stop. Otherwise, $k < n$, and then iteratively repeat Steps (1), (2), (3) on the $(n-k+1) \times (n-k+1)$ subsystem obtained by deleting the first $k-1$ rows and $k-1$ columns from the current system.

If we let $A_1 = A$ and $A_k = (a^{(k)}_{i\,j})$ be the matrix obtained after $k-1$ elimination steps ($2 \leq k \leq n$), then the kth elimination step is applied to the matrix A_k of the form

$$A_k = \begin{pmatrix} a^{(k)}_{1\,1} & a^{(k)}_{1\,2} & \cdots & \cdots & \cdots & a^{(k)}_{1\,n} \\ 0 & a^{(k)}_{2\,2} & \cdots & \cdots & \cdots & a^{(k)}_{2\,n} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & a^{(k)}_{k\,k} & \cdots & a^{(k)}_{k\,n} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & a^{(k)}_{n\,k} & \cdots & a^{(k)}_{n\,n} \end{pmatrix}.$$

Actually, note that

$$a^{(k)}_{i\,j} = a^{(i)}_{i\,j}$$

for all i, j with $1 \leq i \leq k-2$ and $i \leq j \leq n$, since the first $k-1$ rows remain unchanged after the $(k-1)$th step.

We will prove later that $\det(A_k) = \pm \det(A)$. Consequently, A_k is invertible. The fact that A_k is invertible iff A is invertible can also be shown without determinants from the fact that there is some invertible matrix M_k such that $A_k = M_k A$, as we will see shortly.

Since A_k is invertible, some entry $a^{(k)}_{i\,k}$ with $k \leq i \leq n$ is nonzero. Otherwise, the last $n-k+1$ entries in the first k columns of A_k would be zero, and the first k columns of A_k would yield k vectors in $\mathbb{R}^{k-1}$. But then the first k columns of A_k would be linearly dependent and A_k would not be invertible, a contradiction. This situation is illustrated by the following matrix for $n = 5$ and $k = 3$:

$$\begin{pmatrix} a_{1\,1}^{(3)} & a_{1\,2}^{(3)} & a_{1\,3}^{(3)} & a_{1\,3}^{(3)} & a_{1\,5}^{(3)} \\ 0 & a_{2\,2}^{(3)} & a_{2\,3}^{(3)} & a_{2\,4}^{(3)} & a_{2\,5}^{(3)} \\ 0 & 0 & 0 & a_{3\,4}^{(3)} & a_{3\,5}^{(3)} \\ 0 & 0 & 0 & a_{4\,4}^{(3)} & a_{4\,n}^{(3)} \\ 0 & 0 & 0 & a_{5\,4}^{(3)} & a_{5\,5}^{(3)} \end{pmatrix}.$$

The first three columns of the above matrix are linearly dependent.

So one of the entries $a_{i\,k}^{(k)}$ with $k \leq i \leq n$ can be chosen as pivot, and we permute the kth row with the ith row, obtaining the matrix $\alpha^{(k)} = (\alpha_{j\,l}^{(k)})$. The new pivot is $\pi_k = \alpha_{k\,k}^{(k)}$, and we zero the entries $i = k+1, \ldots, n$ in column k by adding $-\alpha_{i\,k}^{(k)}/\pi_k$ times row k to row i. At the end of this step, we have A_{k+1}. Observe that the first $k-1$ rows of A_k are identical to the first $k-1$ rows of A_{k+1}.

The process of Gaussian elimination is illustrated in schematic form below:

$$\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times \\ 0 & \mathbf{0} & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \mathbf{0} & \times \end{pmatrix}.$$

8.3 Elementary Matrices and Row Operations

It is easy to figure out what kind of matrices perform the elementary row operations used during Gaussian elimination. The key point is that if $A = PB$, where A, B are $m \times n$ matrices and P is a square matrix of dimension m, if (as usual) we denote the rows of A and B by $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$, then the formula

$$a_{ij} = \sum_{k=1}^{m} p_{ik} b_{kj}$$

giving the (i, j)th entry in A shows that the ith row of A is a *linear combination* of the rows of B:

$$A_i = p_{i1} B_1 + \cdots + p_{im} B_m.$$

Therefore, *multiplication of a matrix on the left by a square matrix performs row operations*. Similarly, multiplication of a matrix on the right by a square matrix performs column operations

The permutation of the kth row with the ith row is achieved by multiplying A on the left by the *transposition matrix* $P(i, k)$, which is the matrix obtained from the identity matrix by permuting rows i and k, *i.e.*,

$$P(i,k) = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 0 & & & & 1 & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & 1 & & & & 0 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix}.$$

For example, if $m = 3$,

$$P(1,3) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

then

$$P(1,3)B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \end{pmatrix} = \begin{pmatrix} b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \end{pmatrix}.$$

Observe that $\det(P(i,k)) = -1$. Furthermore, $P(i,k)$ is *symmetric* ($P(i,k)^\top = P(i,k)$), and

$$P(i,k)^{-1} = P(i,k).$$

During the permutation Step (2), if row k and row i need to be permuted, the matrix A is multiplied on the left by the matrix P_k such that $P_k = P(i,k)$, else we set $P_k = I$.

Adding β times row j to row i (with $i \neq j$) is achieved by multiplying A on the left by the *elementary matrix*,

$$E_{i,j;\beta} = I + \beta e_{i\,j},$$

where

$$(e_{i\,j})_{k\,l} = \begin{cases} 1 & \text{if } k = i \text{ and } l = j \\ 0 & \text{if } k \neq i \text{ or } l \neq j, \end{cases}$$

i.e.,

$$E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & \beta & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix} \quad \text{or} \quad E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & \beta & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix},$$

on the left, $i > j$, and on the right, $i < j$. The index i is the index of the row that is *changed* by the multiplication. For example, if $m = 3$ and we want to add twice row 1 to row 3, since $\beta = 2$, $j = 1$ and $i = 3$, we form

$$E_{3,1;2} = I + 2e_{31} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix},$$

and calculate

$$\begin{aligned} E_{3,1;2}B &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \end{pmatrix} \\ &= \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ 2b_{11} + b_{31} & 2b_{12} + b_{32} & \cdots & \cdots & \cdots 2b_{1n} + b_{3n} \end{pmatrix}. \end{aligned}$$

Observe that the inverse of $E_{i,j;\beta} = I + \beta e_{i\,j}$ is $E_{i,j;-\beta} = I - \beta e_{i\,j}$ and that $\det(E_{i,j;\beta}) = 1$. Therefore, during Step 3 (the elimination step), the matrix A is multiplied on the left by a product E_k of matrices of the form $E_{i,k;\beta_{i,k}}$, with $i > k$.

Consequently, we see that

$$A_{k+1} = E_k P_k A_k,$$

and then

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A.$$

This justifies the claim made earlier that $A_k = M_k A$ for some invertible matrix M_k; we can pick

$$M_k = E_{k-1} P_{k-1} \cdots E_1 P_1,$$

a product of invertible matrices.

The fact that $\det(P(i,k)) = -1$ and that $\det(E_{i,j;\beta}) = 1$ implies immediately the fact claimed above: We always have

$$\det(A_k) = \pm \det(A).$$

Furthermore, since

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A$$

and since Gaussian elimination stops for $k = n$, the matrix

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A$$

is upper-triangular. Also note that if we let $M = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1$, then $\det(M) = \pm 1$, and

$$\det(A) = \pm \det(A_n).$$

The matrices $P(i,k)$ and $E_{i,j;\beta}$ are called *elementary matrices*. We can summarize the above discussion in the following theorem:

Theorem 8.1. *(Gaussian elimination) Let A be an $n \times n$ matrix (invertible or not). Then there is some invertible matrix M so that $U = MA$ is upper-triangular. The pivots are all nonzero iff A is invertible.*

Proof. We already proved the theorem when A is invertible, as well as the last assertion. Now A is singular iff some pivot is zero, say at Stage k of the elimination. If so, we must have $a_{i\,k}^{(k)} = 0$ for $i = k, \ldots, n$; but in this case, $A_{k+1} = A_k$ and we may pick $P_k = E_k = I$. $\square$

Remark: Obviously, the matrix M can be computed as

$$M = E_{n-1}P_{n-1} \cdots E_2P_2E_1P_1,$$

but this expression is of no use. Indeed, what we need is M^{-1}; when no permutations are needed, it turns out that M^{-1} can be obtained immediately from the matrices E_k's, in fact, from their inverses, and no multiplications are necessary.

Remark: Instead of looking for an invertible matrix M so that MA is upper-triangular, we can look for an invertible matrix M so that MA is a diagonal matrix. Only a simple change to Gaussian elimination is needed. At every Stage k, after the pivot has been found and pivoting been performed, if necessary, in addition to adding suitable multiples of the kth row to the rows *below* row k in order to zero the entries in column k for $i = k+1, \ldots, n$, also add suitable multiples of the kth row to the rows *above* row k in order to zero the entries in column k for $i = 1, \ldots, k-1$. Such steps are also achieved by multiplying on the left by elementary matrices $E_{i,k;\beta_{i,k}}$, except that $i < k$, so that these matrices are not lower-triangular matrices. Nevertheless, at the end of the process, we find that $A_n = MA$, is a diagonal matrix.

This method is called the *Gauss-Jordan factorization*. Because it is more expensive than Gaussian elimination, this method is not used much in practice. However, Gauss-Jordan factorization can be used to compute the inverse of a matrix A. Indeed, we find the jth column of A^{-1} by solving the system $Ax^{(j)} = e_j$ (where e_j is the jth canonical basis vector of $\mathbb{R}^n$). By applying Gauss-Jordan, we are led to a system of the form $D_jx^{(j)} = M_je_j$, where D_j is a diagonal matrix, and we can immediately compute $x^{(j)}$.

It remains to discuss the choice of the pivot, and also conditions that guarantee that no permutations are needed during the Gaussian elimination process. We begin by stating a necessary and sufficient condition for an invertible matrix to have an LU-factorization (*i.e.*, Gaussian elimination does not require pivoting).

8.4 LU-Factorization

Definition 8.1. We say that an invertible matrix A has an *LU-factorization* if it can be written as $A = LU$, where U is upper-triangular invertible and L is lower-triangular, with $L_{i\,i} = 1$ for $i = 1, \ldots, n$.

A lower-triangular matrix with diagonal entries equal to 1 is called a *unit lower-triangular* matrix. Given an $n \times n$ matrix $A = (a_{i\,j})$, for any k with $1 \leq k \leq n$, let $A(1:k, 1:k)$ denote the submatrix of A whose entries are $a_{i\,j}$, where $1 \leq i, j \leq k$.[1] For example, if A is the 5×5 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{pmatrix},$$

then

$$A(1:3, 1:3) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Proposition 8.2. *Let A be an invertible $n \times n$-matrix. Then A has an LU-factorization $A = LU$ iff every matrix $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$. Furthermore, when A has an LU-factorization, we have*

$$\det(A(1:k, 1:k)) = \pi_1 \pi_2 \cdots \pi_k, \quad k = 1, \ldots, n,$$

where π_k is the pivot obtained after $k-1$ elimination steps. Therefore, the kth pivot is given by

$$\pi_k = \begin{cases} a_{11} = \det(A(1:1, 1:1)) & \text{if } k = 1 \\ \dfrac{\det(A(1:k, 1:k))}{\det(A(1:k-1, 1:k-1))} & \text{if } k = 2, \ldots, n. \end{cases}$$

Proof. First assume that $A = LU$ is an LU-factorization of A. We can write

$$A = \begin{pmatrix} A(1:k, 1:k) & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} U_1 & U_2 \\ 0 & U_4 \end{pmatrix} = \begin{pmatrix} L_1 U_1 & L_1 U_2 \\ L_3 U_1 & L_3 U_2 + L_4 U_4 \end{pmatrix},$$

where L_1, L_4 are unit lower-triangular and U_1, U_4 are upper-triangular. (Note, $A(1:k, 1:k)$, L_1, and U_1 are $k \times k$ matrices; A_2 and U_2 are $k \times (n-k)$ matrices; A_3 and L_3 are $(n-k) \times k$ matrices; A_4, L_4, and U_4 are $(n-k) \times (n-k)$ matrices.) Thus,

$$A(1:k, 1:k) = L_1 U_1,$$

and since U is invertible, U_1 is also invertible (the determinant of U is the product of the diagonal entries in U, which is the product of the diagonal entries in U_1 and U_4). As L_1 is invertible (since its diagonal entries are equal to 1), we see that $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$.

Conversely, assume that $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$. We just need to show that Gaussian elimination does not need pivoting. We prove by induction on k that the kth step does not need pivoting.

[1] We are using `Matlab`'s notation.

This holds for $k = 1$, since $A(1:1, 1:1) = (a_{1\,1})$, so $a_{1\,1} \neq 0$. Assume that no pivoting was necessary for the first $k-1$ steps ($2 \leq k \leq n-1$). In this case, we have

$$E_{k-1} \cdots E_2 E_1 A = A_k,$$

where $L = E_{k-1} \cdots E_2 E_1$ is a unit lower-triangular matrix and $A_k(1:k, 1:k)$ is upper-triangular, so that $LA = A_k$ can be written as

$$\begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} A(1:k, 1:k) & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} U_1 & B_2 \\ 0 & B_4 \end{pmatrix},$$

where L_1 is unit lower-triangular and U_1 is upper-triangular. (Once again $A(1:k, 1:k)$, L_1, and U_1 are $k \times k$ matrices; A_2 and B_2 are $k \times (n-k)$ matrices; A_3 and L_3 are $(n-k) \times k$ matrices; A_4, L_4, and B_4 are $(n-k) \times (n-k)$ matrices.) But then,

$$L_1 A(1:k, 1:k)) = U_1,$$

where L_1 is invertible (in fact, $\det(L_1) = 1$), and since by hypothesis $A(1:k, 1:k)$ is invertible, U_1 is also invertible, which implies that $(U_1)_{kk} \neq 0$, since U_1 is upper-triangular. Therefore, no pivoting is needed in Step k, establishing the induction step. Since $\det(L_1) = 1$, we also have

$$\det(U_1) = \det(L_1 A(1:k, 1:k)) = \det(L_1)\det(A(1:k, 1:k)) = \det(A(1:k, 1:k)),$$

and since U_1 is upper-triangular and has the pivots $\pi_1, \ldots, \pi_k$ on its diagonal, we get

$$\det(A(1:k, 1:k)) = \pi_1 \pi_2 \cdots \pi_k, \quad k = 1, \ldots, n,$$

as claimed. ☐

Remark: The use of determinants in the first part of the proof of Proposition 8.2 can be avoided if we use the fact that a triangular matrix is invertible iff all its diagonal entries are nonzero.

Corollary 8.3. *(LU-Factorization) Let A be an invertible $n \times n$-matrix. If every matrix $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$, then Gaussian elimination requires no pivoting and yields an LU-factorization $A = LU$.*

Proof. We proved in Proposition 8.2 that in this case Gaussian elimination requires no pivoting. Then since every elementary matrix $E_{i,k;\beta}$ is lower-triangular (since we always arrange that the pivot π_k occurs above the rows that it operates on), since $E_{i,k;\beta}^{-1} = E_{i,k;-\beta}$ and the $E_k s$ are products of $E_{i,k;\beta_{i,k}} s$, from

$$E_{n-1} \cdots E_2 E_1 A = U,$$

where U is an upper-triangular matrix, we get

$$A = LU,$$

where $L = E_1^{-1} E_2^{-1} \cdots E_{n-1}^{-1}$ is a lower-triangular matrix. Furthermore, as the diagonal entries of each $E_{i,k;\beta}$ are 1, the diagonal entries of each E_k are also 1. ☐

Example 8.1. The reader should verify that

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

is an LU-factorization.

One of the main reasons why the existence of an LU-factorization for a matrix A is interesting is that if we need to solve *several* linear systems $Ax = b$ corresponding to the same matrix A, we can do this cheaply by solving the two triangular systems

$$Lw = b, \quad \text{and} \quad Ux = w.$$

There is a certain asymmetry in the LU-decomposition $A = LU$ of an invertible matrix A. Indeed, the diagonal entries of L are all 1, but this is generally false for U. This asymmetry can be eliminated as follows: if

$$D = \mathrm{diag}(u_{11}, u_{22}, \ldots, u_{nn})$$

is the diagonal matrix consisting of the diagonal entries in U (the pivots), then if we let $U' = D^{-1}U$, we can write

$$A = LDU',$$

where L is lower- triangular, U' is upper-triangular, all diagonal entries of both L and U' are 1, and D is a diagonal matrix of pivots. Such a decomposition leads to the following definition.

Definition 8.2. We say that an invertible $n \times n$ matrix A has an LDU-factorization if it can be written as $A = LDU'$, where L is lower- triangular, U' is upper-triangular, all diagonal entries of both L and U' are 1, and D is a diagonal matrix.

We will see shortly than if A is real symmetric, then $U' = L^\top$.

As we will see a bit later, real symmetric positive definite matrices satisfy the condition of Proposition 8.2. *Therefore, linear systems involving real symmetric positive definite matrices can be solved by Gaussian elimination without pivoting.* Actually, it is possible to do better: this is the Cholesky factorization.

If a square invertible matrix A has an LU-factorization, then it is possible to find L and U while performing Gaussian elimination. Recall that at Step k, we pick a pivot $\pi_k = a_{ik}^{(k)} \neq 0$ in the portion consisting of the entries of index $j \geq k$ of the k-th column of the matrix A_k obtained so far, we swap rows i and k if necessary (the pivoting step), and then we zero the entries of index $j = k+1, \ldots, n$ in column k. Schematically, we have the following steps:

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \end{pmatrix}.$$

More precisely, after permuting row k and row i (the pivoting step), if the entries in column k below row k are $\alpha_{k+1k}, \ldots, \alpha_{nk}$, then we add $-\alpha_{jk}/\pi_k$ times row k to row j; this process is illustrated below:

$$\begin{pmatrix} a_{kk}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{ik}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} a_{ik}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{kk}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} = \begin{pmatrix} \pi_k \\ \alpha_{k+1k} \\ \vdots \\ \alpha_{ik} \\ \vdots \\ \alpha_{nk} \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \pi_k \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

Then if we write $\ell_{jk} = \alpha_{jk}/\pi_k$ for $j = k+1, \ldots, n$, the kth column of L is

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \ell_{k+1k} \\ \vdots \\ \ell_{nk} \end{pmatrix}.$$

Observe that the signs of the multipliers $-\alpha_{jk}/\pi_k$ have been flipped. Thus, we obtain the unit lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix}.$$

It is easy to see (and this is proven in Theorem 8.5) that if the result of Gaussian elimination (without pivoting) is $U = E_{n-1} \cdots E_1 A$, so that $L = E_1^{-1} E_2^{-1} \cdots E_{n-1}^{-1}$, then

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

so the kth column of E_k^{-1} is the kth column of L.

Unfortunately, even though $L^{-1} = E_{n-1} \cdots E_2 E_1$, the matrices E_k occur in the wrong order and the kth column of L^{-1} is *not* the kth column of E_k.

Here is an example illustrating the method.

Example 8.2. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: The first pivot is $\pi_1 = 1$ in row 1, and we subtract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 2, and we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \qquad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

It is easy to check that indeed

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix} = A.$$

We now show how to extend the above method to deal with pivoting efficiently. This is the $PA = LU$ factorization.

8.5 $PA = LU$ Factorization

The following easy proposition shows that, in principle, A can be premultiplied by some permutation matrix P, so that PA can be converted to upper-triangular form without using any pivoting. Permutations are discussed in some detail in Section 30.3, but for now we just need this definition. For the precise connection between the notion of permutation (as discussed in Section 30.3) and permutation matrices, see Problem 8.16.

Definition 8.3. A *permutation matrix* is a square matrix that has a single 1 in every row and every column and zeros everywhere else.

It is shown in Section 30.3 that every permutation matrix is a product of transposition matrices (the $P(i,k)$s), and that P is invertible with inverse $P^\top$.

Proposition 8.4. *Let A be an invertible $n \times n$-matrix. There is some permutation matrix P so that $(PA)(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$.*

Proof. The case $n = 1$ is trivial, and so is the case $n = 2$ (we swap the rows if necessary). If $n \geq 3$, we proceed by induction. Since A is invertible, its columns are linearly independent; in particular, its first $n-1$ columns are also linearly independent. Delete the last column of A. Since the remaining $n-1$ columns are linearly independent, there are also $n-1$ linearly independent rows in the corresponding $n \times (n-1)$ matrix. Thus, there is a permutation of these n rows so that the $(n-1) \times (n-1)$ matrix consisting of the first $n-1$ rows is invertible. But then there is a corresponding permutation matrix P_1, so that the first $n-1$ rows and columns of $P_1 A$ form an invertible matrix A'. Applying the induction hypothesis to the $(n-1) \times (n-1)$ matrix A', we see that there some permutation matrix P_2 (leaving the nth row fixed), so that $(P_2 P_1 A)(1:k, 1:k)$ is invertible, for $k = 1, \ldots, n-1$. Since A is invertible in the first place and P_1 and P_2 are invertible, $P_1 P_2 A$ is also invertible, and we are done. ☐

Remark: One can also prove Proposition 8.4 using a clever reordering of the Gaussian elimination steps suggested by Trefethen and Bau [176] (Lecture 21). Indeed, we know that if A is invertible, then there are permutation matrices P_i and products of elementary matrices E_i, so that

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A,$$

where $U = A_n$ is upper-triangular. For example, when $n = 4$, we have $E_3 P_3 E_2 P_2 E_1 P_1 A = U$. We can define new matrices E_1', E_2', E_3' which are still products of elementary matrices so that we have

$$E_3' E_2' E_1' P_3 P_2 P_1 A = U.$$

Indeed, if we let $E_3' = E_3$, $E_2' = P_3 E_2 P_3^{-1}$, and $E_1' = P_3 P_2 E_1 P_2^{-1} P_3^{-1}$, we easily verify that each E_k' is a product of elementary matrices and that

$$E_3' E_2' E_1' P_3 P_2 P_1 = E_3 (P_3 E_2 P_3^{-1})(P_3 P_2 E_1 P_2^{-1} P_3^{-1}) P_3 P_2 P_1 = E_3 P_3 E_2 P_2 E_1 P_1.$$

It can also be proven that E'_1, E'_2, E'_3 are lower triangular (see Theorem 8.5).

In general, we let

$$E'_k = P_{n-1} \cdots P_{k+1} E_k P_{k+1}^{-1} \cdots P_{n-1}^{-1},$$

and we have

$$E'_{n-1} \cdots E'_1 P_{n-1} \cdots P_1 A = U,$$

where each E'_j is a lower triangular matrix (see Theorem 8.5).

It is remarkable that if pivoting steps are necessary during Gaussian elimination, a very simple modification of the algorithm for finding an LU-factorization yields the matrices L, U, and P, such that $PA = LU$. To describe this new method, since the diagonal entries of L are 1s, it is convenient to write

$$L = I + \Lambda.$$

Then in assembling the matrix Λ while performing Gaussian elimination with pivoting, we make the same transposition on the rows of Λ (really Λ_{k-1}) that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method.

Example 8.3. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: We initialize $\Lambda_0 = 0$ and $P_0 = I_4$. The first pivot is $\pi_1 = 1$ in row 1, and we subtract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 3, so we permute row 2 and 3; we also apply this permutation to Λ and P:

$$A'_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda'_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = \Lambda_3 + I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \quad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad P = P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is easy to check that indeed

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

and

$$PA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Using the idea in the remark before the above example, we can prove the theorem below which shows the correctness of the algorithm for computing P, L and U using a simple adaptation of Gaussian elimination.

We are not aware of a detailed proof of Theorem 8.5 in the standard texts. Although Golub and Van Loan [80] state a version of this theorem as their Theorem 3.1.4, they say that "The proof is a messy subscripting argument." Meyer [125] also provides a sketch of proof (see the end of Section 3.10). In view of this situation, we offer a complete proof. It does involve a lot of subscripts and superscripts, but in our opinion, it contains some techniques that go far beyond symbol manipulation.

Theorem 8.5. *For every invertible $n \times n$-matrix A, the following hold:*

(1) There is some permutation matrix P, some upper-triangular matrix U, and some unit lower-triangular matrix L, so that $PA = LU$ (recall, $L_{i\,i} = 1$ for $i = 1, \ldots, n$). Furthermore, if $P = I$, then L and U are unique and they are produced as a result of Gaussian elimination without pivoting.

(2) *If $E_{n-1} \dots E_1 A = U$ is the result of Gaussian elimination without pivoting, write as usual $A_k = E_{k-1} \dots E_1 A$ (with $A_k = (a_{ij}^{(k)})$), and let $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, with $1 \leq k \leq n-1$ and $k+1 \leq i \leq n$. Then*

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix},$$

where the kth column of L is the kth column of E_k^{-1}, for $k = 1, \dots, n-1$.

(3) *If $E_{n-1}P_{n-1} \cdots E_1 P_1 A = U$ is the result of Gaussian elimination with some pivoting, write $A_k = E_{k-1}P_{k-1} \cdots E_1 P_1 A$, and define E_j^k, with $1 \leq j \leq n-1$ and $j \leq k \leq n-1$, such that, for $j = 1, \dots, n-2$,*

$$\begin{aligned} E_j^j &= E_j \\ E_j^k &= P_k E_j^{k-1} P_k, \quad \text{for } k = j+1, \dots, n-1, \end{aligned}$$

and

$$E_{n-1}^{n-1} = E_{n-1}.$$

Then,

$$\begin{aligned} E_j^k &= P_k P_{k-1} \cdots P_{j+1} E_j P_{j+1} \cdots P_{k-1} P_k \\ U &= E_{n-1}^{n-1} \cdots E_1^{n-1} P_{n-1} \cdots P_1 A, \end{aligned}$$

and if we set

$$\begin{aligned} P &= P_{n-1} \cdots P_1 \\ L &= (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}, \end{aligned}$$

then

$$PA = LU. \tag{$\dagger_1$}$$

Furthermore,

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k, \quad 1 \leq j \leq n-1,\ j \leq k \leq n-1,$$

where $\mathcal{E}_j^k$ is a lower triangular matrix of the form

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

we have

$$E_j^k = I - \mathcal{E}_j^k,$$

and

$$\mathcal{E}_j^k = P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

where $P_k = I$ or else $P_k = P(k,i)$ for some i such that $k+1 \leq i \leq n$; if $P_k \neq I$, this means that $(E_j^k)^{-1}$ is obtained from $(E_j^{k-1})^{-1}$ by permuting the entries on rows i and k in column j. Because the matrices $(E_j^k)^{-1}$ are all lower triangular, the matrix L is also lower triangular.

In order to find L, define lower triangular $n \times n$ matrices Λ_k of the form

$$\Lambda_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{(k)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{(k)} & \lambda_{32}^{(k)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k+11}^{(k)} & \lambda_{k+12}^{(k)} & \cdots & \lambda_{k+1k}^{(k)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+21}^{(k)} & \lambda_{k+22}^{(k)} & \cdots & \lambda_{k+2k}^{(k)} & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{(k)} & \lambda_{n2}^{(k)} & \cdots & \lambda_{nk}^{(k)} & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

to assemble the columns of L iteratively as follows: let

$$(-\ell_{k+1k}^{(k)}, \ldots, -\ell_{nk}^{(k)})$$

be the last $n-k$ elements of the kth column of E_k, and define Λ_k inductively by setting

$$\Lambda_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 0 \end{pmatrix},$$

then for $k = 2, \ldots, n-1$, define

$$\Lambda_k' = P_k \Lambda_{k-1}, \tag{$\dagger_2$}$$

and

$$\Lambda_k = (I + \Lambda'_k)E_k^{-1} - I = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{\prime(k-1)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{\prime(k-1)} & \lambda_{32}^{\prime(k-1)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k1}^{\prime(k-1)} & \lambda_{k2}^{\prime(k-1)} & \cdots & \lambda_{k\,(k-1)}^{\prime(k-1)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+1 1}^{\prime(k-1)} & \lambda_{k+1 2}^{\prime(k-1)} & \cdots & \lambda_{k+1\,(k-1)}^{\prime(k-1)} & \ell_{k+1k}^{(k)} & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{\prime(k-1)} & \lambda_{n2}^{\prime(k-1)} & \cdots & \lambda_{n\,k-1}^{\prime(k-1)} & \ell_{nk}^{(k)} & \cdots & \cdots & 0 \end{pmatrix},$$

with $P_k = I$ or $P_k = P(k, i)$ for some $i > k$. This means that in assembling L, row k and row i of Λ_{k-1} need to be permuted when a pivoting step permuting row k and row i of A_k is required. Then

$$\begin{aligned} I + \Lambda_k &= (E_1^k)^{-1} \cdots (E_k^k)^{-1} \\ \Lambda_k &= \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k, \end{aligned}$$

for $k = 1, \ldots, n-1$, and therefore

$$L = I + \Lambda_{n-1}.$$

The proof of Theorem 8.5, which is very technical, is given in Section 8.6.

We emphasize again that Part (3) of Theorem 8.5 shows the remarkable fact that in assembling the matrix L while performing Gaussian elimination with pivoting, the only change to the algorithm is to make the same transposition on the rows of Λ_{k-1} that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We can also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method.

Example 8.4. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}.$$

We set $P_0 = I_4$, and we can also set $\Lambda_0 = 0$. The first step is to permute row 1 and row 2, using the pivot 4. We also apply this permutation to P_0:

$$A'_1 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we subtract 1/4 times row 1 from row 2, 1/2 times row 1 from row 3, and add 3/4 times row 1 to row 4, and start assembling Λ:

$$A_2 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 0 & -6 & 6 \\ 0 & -1 & -4 & 5 \\ 0 & 5 & 10 & -10 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we permute row 2 and row 4, using the pivot 5. We also apply this permutation to Λ and P:

$$A_3' = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & -1 & -4 & 5 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda_2' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we add 1/5 times row 2 to row 3, and update Λ_2':

$$A_3 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we permute row 3 and row 4, using the pivot -6. We also apply this permutation to Λ and P:

$$A_4' = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & -2 & 3 \end{pmatrix} \quad \Lambda_3' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finally we subtract 1/3 times row 3 from row 4, and update Λ_3':

$$A_4 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 1/3 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Consequently, adding the identity to Λ_3, we obtain

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We check that

$$PA = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix},$$

and that

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix} = PA.$$

Note that if one willing to overwrite the lower triangular part of the evolving matrix A, one can store the evolving Λ there, since these entries will eventually be zero anyway! There is also no need to save explicitly the permutation matrix P. One could instead record the permutation steps in an extra column (record the vector $(\pi(1), \ldots, \pi(n))$ corresponding to the permutation π applied to the rows). We let the reader write such a bold and space-efficient version of LU-decomposition!

Remark: In `Matlab` the function `lu` returns the matrices P, L, U involved in the $PA = LU$ factorization using the call $[L, U, P] = \mathrm{lu}(A)$.

As a corollary of Theorem 8.5(1), we can show the following result.

Proposition 8.6. *If an invertible real symmetric matrix A has an LU-decomposition, then A has a factorization of the form*

$$A = LDL^\top,$$

where L is a lower-triangular matrix whose diagonal entries are equal to 1, and where D consists of the pivots. Furthermore, such a decomposition is unique.

Proof. If A has an LU-factorization, then it has an LDU factorization

$$A = LDU,$$

where L is lower-triangular, U is upper-triangular, and the diagonal entries of both L and U are equal to 1. Since A is symmetric, we have

$$LDU = A = A^\top = U^\top DL^\top,$$

with $U^\top$ lower-triangular and $DL^\top$ upper-triangular. By the uniqueness of LU-factorization (Part (1) of Theorem 8.5), we must have $L = U^\top$ (and $DU = DL^\top$), thus $U = L^\top$, as claimed. $\square$

Remark: It can be shown that Gaussian elimination plus back-substitution requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications and $n^2/2 + O(n)$ divisions.

8.6 Proof of Theorem 8.5 ⊛

Proof. (1) The only part that has not been proven is the uniqueness part (when $P = I$). Assume that A is invertible and that $A = L_1U_1 = L_2U_2$, with L_1, L_2 unit lower-triangular and U_1, U_2 upper-triangular. Then we have

$$L_2^{-1}L_1 = U_2U_1^{-1}.$$

However, it is obvious that L_2^{-1} is lower-triangular and that U_1^{-1} is upper-triangular, and so $L_2^{-1}L_1$ is lower-triangular and $U_2U_1^{-1}$ is upper-triangular. Since the diagonal entries of L_1 and L_2 are 1, the above equality is only possible if $U_2U_1^{-1} = I$, that is, $U_1 = U_2$, and so $L_1 = L_2$.

(2) When $P = I$, we have $L = E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}$, where E_k is the product of $n-k$ elementary matrices of the form $E_{i,k;-\ell_i}$, where $E_{i,k;-\ell_i}$ subtracts ℓ_i times row k from row i, with $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, $1 \leq k \leq n-1$, and $k+1 \leq i \leq n$. Then it is immediately verified that

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

and that

$$E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}.$$

If we define L_k by

$$L_k = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{21} & 1 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{31} & \ell_{32} & \ddots & 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \ddots & 1 & 0 & \vdots & 0 \\ \ell_{k+11} & \ell_{k+12} & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}$$

for $k = 1, \ldots, n-1$, we easily check that $L_1 = E_1^{-1}$, and that

$$L_k = L_{k-1}E_k^{-1}, \quad 2 \leq k \leq n-1,$$

because multiplication on the right by E_k^{-1} adds ℓ_i times column i to column k (of the matrix L_{k-1}) with $i > k$, and column i of L_{k-1} has only the nonzero entry 1 as its ith element. Since

$$L_k = E_1^{-1} \cdots E_k^{-1}, \quad 1 \leq k \leq n-1,$$

we conclude that $L = L_{n-1}$, proving our claim about the shape of L.

(3)

Step 1. Prove ($\dagger_1$).

First we prove by induction on k that

$$A_{k+1} = E_k^k \cdots E_1^k P_k \cdots P_1 A, \quad k = 1, \ldots, n-2.$$

For $k = 1$, we have $A_2 = E_1 P_1 A = E_1^1 P_1 A$, since $E_1^1 = E_1$, so our assertion holds trivially.

Now if $k \geq 2$,

$$A_{k+1} = E_k P_k A_k,$$

and by the induction hypothesis,

$$A_k = E_{k-1}^{k-1} \cdots E_2^{k-1} E_1^{k-1} P_{k-1} \cdots P_1 A.$$

Because P_k is either the identity or a transposition, $P_k^2 = I$, so by inserting occurrences of $P_k P_k$ as indicated below we can write

$$\begin{aligned}
A_{k+1} &= E_k P_k A_k \\
&= E_k P_k E_{k-1}^{k-1} \cdots E_2^{k-1} E_1^{k-1} P_{k-1} \cdots P_1 A \\
&= E_k P_k E_{k-1}^{k-1} (P_k P_k) \cdots (P_k P_k) E_2^{k-1} (P_k P_k) E_1^{k-1} (P_k P_k) P_{k-1} \cdots P_1 A \\
&= E_k (P_k E_{k-1}^{k-1} P_k) \cdots (P_k E_2^{k-1} P_k)(P_k E_1^{k-1} P_k) P_k P_{k-1} \cdots P_1 A.
\end{aligned}$$

Observe that P_k has been "moved" to the right of the elimination steps. However, by definition,

$$\begin{aligned}
E_j^k &= P_k E_j^{k-1} P_k, \quad j = 1, \ldots, k-1 \\
E_k^k &= E_k,
\end{aligned}$$

so we get

$$A_{k+1} = E_k^k E_{k-1}^k \cdots E_2^k E_1^k P_k \cdots P_1 A,$$

establishing the induction hypothesis. For $k = n-2$, we get

$$U = A_{n-1} = E_{n-1}^{n-1} \cdots E_1^{n-1} P_{n-1} \cdots P_1 A,$$

as claimed, and the factorization $PA = LU$ with

$$\begin{aligned}
P &= P_{n-1} \cdots P_1 \\
L &= (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}
\end{aligned}$$

is clear.

Step 2. Prove that the matrices $(E_j^k)^{-1}$ are lower-triangular. To achieve this, we prove that the matrices $\mathcal{E}_j^k$ are strictly lower triangular matrices of a very special form.

Since for $j = 1, \ldots, n-2$, we have $E_j^j = E_j$,

$$E_j^k = P_k E_j^{k-1} P_k, \quad k = j+1, \ldots, n-1,$$

since $E_{n-1}^{n-1} = E_{n-1}$ and $P_k^{-1} = P_k$, we get $(E_j^j)^{-1} = E_j^{-1}$ for $j = 1, \ldots, n-1$, and for $j = 1, \ldots, n-2$, we have

$$(E_j^k)^{-1} = P_k (E_j^{k-1})^{-1} P_k, \quad k = j+1, \ldots, n-1.$$

Since

$$(E_j^{k-1})^{-1} = I + \mathcal{E}_j^{k-1}$$

and $P_k = P(k, i)$ is a transposition or $P_k = I$, so $P_k^2 = I$, and we get

$$(E_j^k)^{-1} = P_k (E_j^{k-1})^{-1} P_k = P_k (I + \mathcal{E}_j^{k-1}) P_k = P_k^2 + P_k \, \mathcal{E}_j^{k-1} \, P_k = I + P_k \, \mathcal{E}_j^{k-1} \, P_k.$$

Therefore, we have

$$(E_j^k)^{-1} = I + P_k \, \mathcal{E}_j^{k-1} \, P_k, \quad 1 \leq j \leq n-2, \; j+1 \leq k \leq n-1.$$

We prove for $j = 1, \ldots, n-1$, that for $k = j, \ldots, n-1$, each $\mathcal{E}_j^k$ is a lower triangular matrix of the form

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

and that

$$\mathcal{E}_j^k = P_k \, \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2, \; j+1 \leq k \leq n-1,$$

with $P_k = I$ or $P_k = P(k, i)$ for some i such that $k+1 \leq i \leq n$.

For each j ($1 \leq j \leq n-1$) we proceed by induction on $k = j, \ldots, n-1$. Since $(E_j^j)^{-1} = E_j^{-1}$ and since E_j^{-1} is of the above form, the base case holds.

For the induction step, we only need to consider the case where $P_k = P(k, i)$ is a transposition, since the case where $P_k = I$ is trivial. We have to figure out what $P_k \, \mathcal{E}_j^{k-1} \, P_k =$

$P(k, i)\, \mathcal{E}_j^{k-1}\, P(k, i)$ is. However, since

$$\mathcal{E}_j^{k-1} = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k-1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k-1)} & 0 & \cdots & 0 \end{pmatrix},$$

and because $k + 1 \leq i \leq n$ and $j \leq k - 1$, multiplying $\mathcal{E}_j^{k-1}$ on the right by $P(k, i)$ will permute *columns* i and k, which are columns of zeros, so

$$P(k, i)\, \mathcal{E}_j^{k-1}\, P(k, i) = P(k, i)\, \mathcal{E}_j^{k-1},$$

and thus,

$$(E_j^k)^{-1} = I + P(k, i)\, \mathcal{E}_j^{k-1}.$$

But since

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k,$$

we deduce that

$$\mathcal{E}_j^k = P(k, i)\, \mathcal{E}_j^{k-1}.$$

We also know that multiplying $\mathcal{E}_j^{k-1}$ on the left by $P(k, i)$ will permute *rows* i and k, which shows that $\mathcal{E}_j^k$ has the desired form, as claimed. Since all $\mathcal{E}_j^k$ are strictly lower triangular, all $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ are lower triangular, so the product

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}$$

is also lower triangular.

Step 3. Express L as $L = I + \Lambda_{n-1}$, with $\Lambda_{n-1} = \mathcal{E}_1^1 + \cdots + \mathcal{E}_{n-1}^{n-1}$.

From Step 1 of Part (3), we know that

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}.$$

We prove by induction on k that

$$\begin{aligned} I + \Lambda_k &= (E_1^k)^{-1} \cdots (E_k^k)^{-1} \\ \Lambda_k &= \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k, \end{aligned}$$

for $k = 1, \ldots, n - 1$.

If $k = 1$, we have $E_1^1 = E_1$ and

$$E_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix}.$$

We also get

$$(E_1^{-1})^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix} = I + \Lambda_1.$$

Since $(E_1^{-1})^{-1} = I + \mathcal{E}_1^1$, we find that we get $\Lambda_1 = \mathcal{E}_1^1$, and the base step holds.

Since $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ with

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix}$$

and $\mathcal{E}_i^k \mathcal{E}_j^k = 0$ if $i < j$, as in part (2) for the computation involving the products of L_k's, we get

$$(E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1} = I + \mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}, \quad 2 \leq k \leq n. \tag{$*$}$$

Similarly, from the fact that $\mathcal{E}_j^{k-1} P(k, i) = \mathcal{E}_j^{k-1}$ if $i \geq k+1$ and $j \leq k-1$ and since

$$(E_j^k)^{-1} = I + P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

we get

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k(\mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}), \quad 2 \leq k \leq n-1. \tag{$**$}$$

By the induction hypothesis,

$$I + \Lambda_{k-1} = (E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1},$$

and from $(*)$, we get

$$\Lambda_{k-1} = \mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}.$$

Using $(**)$, we deduce that

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k \Lambda_{k-1}.$$

Since $E_k^k = E_k$, we obtain

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1} = (I + P_k\Lambda_{k-1})E_k^{-1}.$$

However, by definition

$$I + \Lambda_k = (I + P_k\Lambda_{k-1})E_k^{-1},$$

which proves that

$$I + \Lambda_k = (E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1}, \qquad (\dagger)$$

and finishes the induction step for the proof of this formula.

If we apply Equation $(*)$ again with $k+1$ in place of k, we have

$$(E_1^k)^{-1} \cdots (E_k^k)^{-1} = I + \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k,$$

and together with $(\dagger)$, we obtain,

$$\Lambda_k = \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k,$$

also finishing the induction step for the proof of this formula. For $k = n-1$ in $(\dagger)$, we obtain the desired equation: $L = I + \Lambda_{n-1}$. $\square$

8.7 Dealing with Roundoff Errors; Pivoting Strategies

Let us now briefly comment on the choice of a pivot. Although theoretically, any pivot can be chosen, the possibility of roundoff errors implies that it is not a good idea to pick very small pivots. The following example illustrates this point. Consider the linear system

$$\begin{array}{rcrcl} 10^{-4}x & + & y & = & 1 \\ x & + & y & = & 2. \end{array}$$

Since 10^{-4} is nonzero, it can be taken as pivot, and we get

$$\begin{array}{rcrcl} 10^{-4}x & + & y & = & 1 \\ & & (1-10^4)y & = & 2-10^4. \end{array}$$

Thus, the exact solution is

$$x = \frac{10^4}{10^4-1}, \quad y = \frac{10^4-2}{10^4-1}.$$

However, if roundoff takes place on the fourth digit, then $10^4 - 1 = 9999$ and $10^4 - 2 = 9998$ will be rounded off both to 9990, and then the solution is $x = 0$ and $y = 1$, very far from the exact solution where $x \approx 1$ and $y \approx 1$. The problem is that we picked a very small pivot. If instead we permute the equations, the pivot is 1, and after elimination we get the system

$$\begin{array}{rcrcl} x & + & y & = & 2 \\ & & (1-10^{-4})y & = & 1-2\times 10^{-4}. \end{array}$$

This time, $1 - 10^{-4} = 0.9999$ and $1 - 2 \times 10^{-4} = 0.9998$ are rounded off to 0.999 and the solution is $x = 1, y = 1$, much closer to the exact solution.

To remedy this problem, one may use the strategy of *partial pivoting*. This consists of choosing during Step k ($1 \leq k \leq n-1$) one of the entries $a^{(k)}_{i\,k}$ such that

$$|a^{(k)}_{i\,k}| = \max_{k \leq p \leq n} |a^{(k)}_{p\,k}|.$$

By maximizing the value of the pivot, we avoid dividing by undesirably small pivots.

Remark: A matrix, A, is called *strictly column diagonally dominant* iff

$$|a_{j\,j}| > \sum_{i=1,\, i \neq j}^{n} |a_{i\,j}|, \quad \text{for } j = 1, \ldots, n$$

(resp. *strictly row diagonally dominant* iff

$$|a_{i\,i}| > \sum_{j=1,\, j \neq i}^{n} |a_{i\,j}|, \quad \text{for } i = 1, \ldots, n.)$$

For example, the matrix

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix}$$

of the curve interpolation problem discussed in Section 8.1 is strictly column (and row) diagonally dominant.

It has been known for a long time (before 1900, say by Hadamard) that if a matrix A is strictly column diagonally dominant (resp. strictly row diagonally dominant), then it is invertible. It can also be shown that if A is strictly column diagonally dominant, then Gaussian elimination with partial pivoting does not actually require pivoting (see Problem 8.12).

Another strategy, called *complete pivoting*, consists in choosing some entry $a^{(k)}_{i\,j}$, where $k \leq i, j \leq n$, such that

$$|a^{(k)}_{i\,j}| = \max_{k \leq p,q \leq n} |a^{(k)}_{p\,q}|.$$

However, in this method, if the chosen pivot is not in column k, it is also necessary to permute columns. This is achieved by multiplying on the right by a permutation matrix. However, complete pivoting tends to be too expensive in practice, and partial pivoting is the method of choice.

A special case where the LU-factorization is particularly efficient is the case of tridiagonal matrices, which we now consider.

8.8 Gaussian Elimination of Tridiagonal Matrices

Consider the tridiagonal matrix

$$A = \begin{pmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & & & \\ & a_3 & b_3 & c_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & a_{n-2} & b_{n-2} & c_{n-2} & \\ & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & & a_n & b_n \end{pmatrix}.$$

Define the sequence

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \delta_k = b_k\delta_{k-1} - a_kc_{k-1}\delta_{k-2}, \quad 2 \leq k \leq n.$$

Proposition 8.7. *If A is the tridiagonal matrix above, then $\delta_k = \det(A(1 : k, 1 : k))$ for $k = 1, \ldots, n$.*

Proof. By expanding $\det(A(1 : k, 1 : k))$ with respect to its last row, the proposition follows by induction on k. $\square$

Theorem 8.8. *If A is the tridiagonal matrix above and $\delta_k \neq 0$ for $k = 1, \ldots, n$, then A has the following LU-factorization:*

$$A = \begin{pmatrix} 1 & & & & & \\ a_2\dfrac{\delta_0}{\delta_1} & 1 & & & & \\ & a_3\dfrac{\delta_1}{\delta_2} & 1 & & & \\ & & \ddots & \ddots & & \\ & & & a_{n-1}\dfrac{\delta_{n-3}}{\delta_{n-2}} & 1 & \\ & & & & a_n\dfrac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} \dfrac{\delta_1}{\delta_0} & c_1 & & & & \\ & \dfrac{\delta_2}{\delta_1} & c_2 & & & \\ & & \dfrac{\delta_3}{\delta_2} & c_3 & & \\ & & & \ddots & \ddots & \\ & & & & \dfrac{\delta_{n-1}}{\delta_{n-2}} & c_{n-1} \\ & & & & & \dfrac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

Proof. Since $\delta_k = \det(A(1 : k, 1 : k)) \neq 0$ for $k = 1, \ldots, n$, by Theorem 8.5 (and Proposition 8.2), we know that A has a unique LU-factorization. Therefore, it suffices to check that the proposed factorization works. We easily check that

$$\begin{aligned} (LU)_{k\,k+1} &= c_k, \quad 1 \leq k \leq n-1 \\ (LU)_{k\,k-1} &= a_k, \quad 2 \leq k \leq n \\ (LU)_{k\,l} &= 0, \quad |k-l| \geq 2 \\ (LU)_{1\,1} &= \frac{\delta_1}{\delta_0} = b_1 \\ (LU)_{k\,k} &= \frac{a_kc_{k-1}\delta_{k-2} + \delta_k}{\delta_{k-1}} = b_k, \quad 2 \leq k \leq n, \end{aligned}$$

since $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$. $\square$

It follows that there is a simple method to solve a linear system $Ax = d$ where A is tridiagonal (and $\delta_k \neq 0$ for $k = 1, \ldots, n$). For this, it is convenient to "squeeze" the diagonal matrix Δ defined such that $\Delta_{k\,k} = \delta_k/\delta_{k-1}$ into the factorization so that $A = (L\Delta)(\Delta^{-1}U)$, and if we let

$$z_1 = \frac{c_1}{b_1}, \quad z_k = c_k\,\frac{\delta_{k-1}}{\delta_k}, \quad 2 \leq k \leq n-1, \quad z_n = \frac{\delta_n}{\delta_{n-1}} = b_n - a_n z_{n-1},$$

$A = (L\Delta)(\Delta^{-1}U)$ is written as

$$A = \begin{pmatrix} \frac{c_1}{z_1} \\ a_2 & \frac{c_2}{z_2} \\ & a_3 & \frac{c_3}{z_3} \\ & & \ddots & \ddots \\ & & & a_{n-1} & \frac{c_{n-1}}{z_{n-1}} \\ & & & & a_n & z_n \end{pmatrix} \begin{pmatrix} 1 & z_1 \\ & 1 & z_2 \\ & & 1 & z_3 \\ & & & \ddots & \ddots \\ & & & & 1 & z_{n-2} \\ & & & & & 1 & z_{n-1} \\ & & & & & & 1 \end{pmatrix}.$$

As a consequence, the system $Ax = d$ can be solved by constructing three sequences: First, the sequence

$$z_1 = \frac{c_1}{b_1}, \quad z_k = \frac{c_k}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n-1, \quad z_n = b_n - a_n z_{n-1},$$

corresponding to the recurrence $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$ and obtained by dividing both sides of this equation by δ_{k-1}, next

$$w_1 = \frac{d_1}{b_1}, \quad w_k = \frac{d_k - a_k w_{k-1}}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n,$$

corresponding to solving the system $L\Delta w = d$, and finally

$$x_n = w_n, \quad x_k = w_k - z_k x_{k+1}, \quad k = n-1, n-2, \ldots, 1,$$

corresponding to solving the system $\Delta^{-1}Ux = w$.

Remark: It can be verified that this requires $3(n-1)$ additions, $3(n-1)$ multiplications, and $2n$ divisions, a total of $8n-6$ operations, which is much less that the $O(2n^3/3)$ required by Gaussian elimination in general.

We now consider the special case of symmetric positive definite matrices (SPD matrices).

8.9 SPD Matrices and the Cholesky Decomposition

Definition 8.4. A real $n \times n$ matrix A is *symmetric positive definite*, for short *SPD*, iff it is symmetric and if

$$x^\top Ax > 0 \quad \text{for all } x \in \mathbb{R}^n \text{ with } x \neq 0.$$

The following facts about a symmetric positive definite matrix A are easily established (some left as an exercise):

(1) The matrix A is invertible. (Indeed, if $Ax = 0$, then $x^\top Ax = 0$, which implies $x = 0$.)

(2) We have $a_{i\,i} > 0$ for $i = 1, \ldots, n$. (Just observe that for $x = e_i$, the ith canonical basis vector of $\mathbb{R}^n$, we have $e_i^\top Ae_i = a_{i\,i} > 0$.)

(3) For every $n \times n$ real invertible matrix Z, the matrix $Z^\top AZ$ is real symmetric positive definite iff A is real symmetric positive definite.

(4) The set of $n \times n$ real symmetric positive definite matrices is convex. This means that if A and B are two $n \times n$ symmetric positive definite matrices, then for any $\lambda \in \mathbb{R}$ such that $0 \leq \lambda \leq 1$, the matrix $(1 - \lambda)A + \lambda B$ is also symmetric positive definite. Clearly since A and B are symmetric, $(1 - \lambda)A + \lambda B$ is also symmetric. For any nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$ and $x^\top Bx > 0$, so

$$x^\top((1 - \lambda)A + \lambda B)x = (1 - \lambda)x^\top Ax + \lambda x^\top Bx > 0,$$

because $0 \leq \lambda \leq 1$, so $1-\lambda \geq 0$ and $\lambda \geq 0$, and $1-\lambda$ and λ can't be zero simultaneously.

(5) The set of $n \times n$ real symmetric positive definite matrices is a cone. This means that if A is symmetric positive definite and if $\lambda > 0$ is any real, then λA is symmetric positive definite. Clearly λA is symmetric, and for nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$, and since $\lambda > 0$, we have $x^\top \lambda Ax = \lambda x^\top Ax > 0$.

Remark: Given a complex $m \times n$ matrix A, we define the matrix $\overline{A}$ as the $m \times n$ matrix $\overline{A} = (\overline{a_{ij}})$. Then we define A^* as the $n \times m$ matrix $A^* = (\overline{A})^\top = \overline{(A^\top)}$. The $n \times n$ complex matrix A is *Hermitian* if $A^* = A$. This is the complex analog of the notion of a real symmetric matrix.

Definition 8.5. A complex $n \times n$ matrix A is *Hermitian positive definite*, for short *HPD*, if it is Hermitian and if

$$z^*Az > 0 \quad \text{for all } z \in \mathbb{C}^n \text{ with } z \neq 0.$$

It is easily verified that Properties (1)-(5) hold for Hermitian positive definite matrices; replace $\top$ by $*$.

It is instructive to characterize when a 2×2 real matrix A is symmetric positive definite. Write

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

Then we have

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & c \\ c & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2cxy + by^2.$$

If the above expression is strictly positive for all nonzero vectors $\binom{x}{y}$, then for $x = 1, y = 0$ we get $a > 0$ and for $x = 0, y = 1$ we get $b > 0$. Then we can write

$$\begin{aligned} ax^2 + 2cxy + by^2 &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + by^2 - \frac{c^2}{a}y^2 \\ &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + \frac{1}{a}\left(ab - c^2\right) y^2. \end{aligned} \tag{$\dagger$}$$

Since $a > 0$, if $ab - c^2 \leq 0$, then we can choose $y > 0$ so that the second term is negative or zero, and we can set $x = -(c/a)y$ to make the first term zero, in which case $ax^2+2cxy+by^2 \leq 0$, so we must have $ab - c^2 > 0$.

Conversely, if $a > 0, b > 0$ and $ab > c^2$, then for any $(x, y) \neq (0, 0)$, if $y = 0$, then $x \neq 0$ and the first term of ($\dagger$) is positive, and if $y \neq 0$, then the second term of ($\dagger$) is positive. Therefore, the matrix A is symmetric positive definite iff

$$a > 0,\ b > 0,\ ab > c^2. \tag{$*$}$$

Note that $ab - c^2 = \det(A)$, so the third condition says that $\det(A) > 0$.

Observe that the condition $b > 0$ is redundant, since if $a > 0$ and $ab > c^2$, then we must have $b > 0$ (and similarly $b > 0$ and $ab > c^2$ implies that $a > 0$).

We can try to visualize the space of 2×2 real symmetric positive definite matrices in $\mathbb{R}^3$, by viewing (a, b, c) as the coordinates along the x, y, z axes. Then the locus determined by the strict inequalities in ($*$) corresponds to the region on the side of the cone of equation $xy = z^2$ that does not contain the origin and for which $x > 0$ and $y > 0$. For $z = \delta$ fixed, the equation $xy = \delta^2$ define a hyperbola in the plane $z = \delta$. The cone of equation $xy = z^2$ consists of the lines through the origin that touch the hyperbola $xy = 1$ in the plane $z = 1$. We only consider the branch of this hyperbola for which $x > 0$ and $y > 0$. See Figure 8.6.

It is not hard to show that the inverse of a real symmetric positive definite matrix is also real symmetric positive definite, but the product of two real symmetric positive definite matrices may *not* be symmetric positive definite, as the following example shows:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1\sqrt{2} \\ -1/\sqrt{2} & 3/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix}.$$

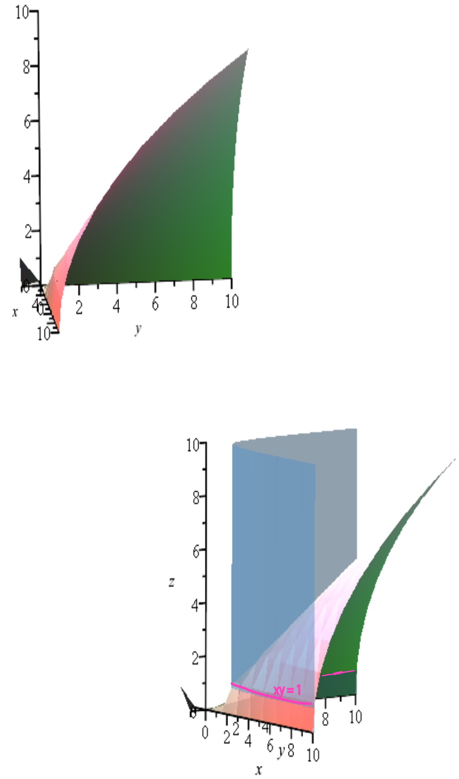


Figure 8.6: Two views of the surface $xy = z^2$ in $\mathbb{R}^3$. The intersection of the surface with a constant z plane results in a hyperbola. The region associated with the 2×2 symmetric positive definite matrices lies in "front" of the green side.

According to the above criterion, the two matrices on the left-hand side are real symmetric positive definite, but the matrix on the right-hand side is not even symmetric, and

$$\begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix} \begin{pmatrix} -6 \\ 1 \end{pmatrix} = \begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 2/\sqrt{2} \\ 11/\sqrt{2} \end{pmatrix} = -1/\sqrt{2},$$

even though its eigenvalues are both real and positive.

Next we show that a real symmetric positive definite matrix has a special LU-factorization of the form $A = BB^\top$, where B is a lower-triangular matrix whose diagonal elements are strictly positive. This is the *Cholesky factorization*.

First we note that a symmetric positive definite matrix satisfies the condition of Proposition 8.2.

Proposition 8.9. *If A is a real symmetric positive definite matrix, then $A(1 : k, 1 : k)$ is symmetric positive definite and thus invertible for $k = 1, \ldots, n$.*

Proof. Since A is symmetric, each $A(1:k,1:k)$ is also symmetric. If $w \in \mathbb{R}^k$, with $1 \leq k \leq n$, we let $x \in \mathbb{R}^n$ be the vector with $x_i = w_i$ for $i = 1, \ldots, k$ and $x_i = 0$ for $i = k+1, \ldots, n$. Now since A is symmetric positive definite, we have $x^\top Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$. This holds in particular for all vectors x obtained from nonzero vectors $w \in \mathbb{R}^k$ as defined earlier, and clearly

$$x^\top Ax = w^\top A(1:k,1:k)\, w,$$

which implies that $A(1:k,1:k)$ is symmetric positive definite. Thus, by Fact 1 above, $A(1:k,1:k)$ is also invertible. $\square$

Proposition 8.9 also holds for a complex Hermitian positive definite matrix. Proposition 8.9 can be strengthened as follows: *A real (resp. complex) matrix A is symmetric (resp. Hermitian) positive definite iff* $\det(A(1:k,1:k)) > 0$ *for* $k = 1, \ldots, n$.

The above fact is known as *Sylvester's criterion*. We will prove it after establishing the Cholesky factorization.

Let A be an $n \times n$ real symmetric positive definite matrix and write

$$A = \begin{pmatrix} a_{1\,1} & W^\top \\ W & C \end{pmatrix},$$

where C is an $(n-1) \times (n-1)$ symmetric matrix and W is an $(n-1) \times 1$ matrix. Since A is symmetric positive definite, $a_{1\,1} > 0$, and we can compute $\alpha = \sqrt{a_{1\,1}}$. The trick is that we can factor A uniquely as

$$A = \begin{pmatrix} a_{1\,1} & W^\top \\ W & C \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{1\,1} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix},$$

i.e., as $A = B_1 A_1 B_1^\top$, where B_1 is lower-triangular with positive diagonal entries. Thus, B_1 is invertible, and by Fact (3) above, A_1 is also symmetric positive definite.

Remark: The matrix $C - WW^\top/a_{1\,1}$ is known as the *Schur complement* of the 1×1 matrix (a_{11}) in A.

Theorem 8.10. *(Cholesky factorization) Let A be a real symmetric positive definite matrix. Then there is some real lower-triangular matrix B so that $A = BB^\top$. Furthermore, B can be chosen so that its diagonal elements are strictly positive, in which case B is unique.*

Proof. We proceed by induction on the dimension n of A. For $n = 1$, we must have $a_{1\,1} > 0$, and if we let $\alpha = \sqrt{a_{1\,1}}$ and $B = (\alpha)$, the theorem holds trivially. If $n \geq 2$, as we explained above, again we must have $a_{1\,1} > 0$, and we can write

$$A = \begin{pmatrix} a_{1\,1} & W^\top \\ W & C \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{1\,1} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} = B_1 A_1 B_1^\top,$$

where $\alpha = \sqrt{a_{1\,1}}$, the matrix B_1 is invertible and

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{1\,1} \end{pmatrix}$$

is symmetric positive definite. However, this implies that $C - WW^\top/a_{1\,1}$ is also symmetric positive definite (consider $x^\top A_1 x$ for every $x \in \mathbb{R}^n$ with $x \neq 0$ and $x_1 = 0$). Thus, we can apply the induction hypothesis to $C - WW^\top/a_{1\,1}$ (which is an $(n-1) \times (n-1)$ matrix), and we find a unique lower-triangular matrix L with positive diagonal entries so that

$$C - WW^\top/a_{1\,1} = LL^\top.$$

But then we get

$$\begin{aligned} A &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C - WW^\top/a_{1\,1} \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & LL^\top \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L^\top \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \alpha & 0 \\ W/\alpha & L \end{pmatrix} \begin{pmatrix} \alpha & W^\top/\alpha \\ 0 & L^\top \end{pmatrix}. \end{aligned}$$

Therefore, if we let

$$B = \begin{pmatrix} \alpha & 0 \\ W/\alpha & L \end{pmatrix},$$

we have a unique lower-triangular matrix with positive diagonal entries and $A = BB^\top$. $\square$

Remark: The uniqueness of the Cholesky decomposition can also be established using the uniqueness of an LU-decomposition. Indeed, if $A = B_1 B_1^\top = B_2 B_2^\top$ where B_1 and B_2 are lower triangular with positive diagonal entries, if we let Δ_1 (resp. Δ_2) be the diagonal matrix consisting of the diagonal entries of B_1 (resp. B_2) so that $(\Delta_k)_{ii} = (B_k)_{ii}$ for $k = 1, 2$, then we have two LU-decompositions

$$A = (B_1 \Delta_1^{-1})(\Delta_1 B_1^\top) = (B_2 \Delta_2^{-1})(\Delta_2 B_2^\top)$$

with $B_1 \Delta_1^{-1}, B_2 \Delta_2^{-1}$ unit lower triangular, and $\Delta_1 B_1^\top, \Delta_2 B_2^\top$ upper triangular. By uniquenes of LU-factorization (Theorem 8.5(1)), we have

$$B_1 \Delta_1^{-1} = B_2 \Delta_2^{-1}, \quad \Delta_1 B_1^\top = \Delta_2 B_2^\top,$$

and the second equation yields

$$B_1 \Delta_1 = B_2 \Delta_2. \tag{$*$}$$

The diagonal entries of $B_1\Delta_1$ are $(B_1)^2_{ii}$ and similarly the diagonal entries of $B_2\Delta_2$ are $(B_2)^2_{ii}$, so the above equation implies that

$$(B_1)^2_{ii} = (B_2)^2_{ii}, \quad i = 1, \ldots, n.$$

Since the diagonal entries of both B_1 and B_2 are assumed to be positive, we must have

$$(B_1)_{ii} = (B_2)_{ii}, \quad i = 1, \ldots, n;$$

that is, $\Delta_1 = \Delta_2$, and since both are invertible, we conclude from $(*)$ that $B_1 = B_2$.

Theorem 8.10 also holds for complex Hermitian positive definite matrices. In this case, we have $A = BB^*$ for some unique lower triangular matrix B with positive diagonal entries.

The proof of Theorem 8.10 immediately yields an algorithm to compute B from A by solving for a lower triangular matrix B such that $A = BB^\top$ (where both A and B are real matrices). For $j = 1, \ldots, n$,

$$b_{j\,j} = \left(a_{j\,j} - \sum_{k=1}^{j-1} b^2_{j\,k}\right)^{1/2},$$

and for $i = j+1, \ldots, n$ (and $j = 1, \ldots, n-1$)

$$b_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k} b_{j\,k}\right) / b_{j\,j}.$$

The above formulae are used to compute the jth column of B from top-down, using the first $j-1$ columns of B previously computed, and the matrix A. In the case of $n = 3$, $A = BB^\top$ yields

$$\begin{aligned}
\begin{pmatrix} a_{11} & a_{12} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} &= \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{21} & b_{31} \\ 0 & b_{22} & b_{32} \\ 0 & 0 & b_{33} \end{pmatrix} \\
&= \begin{pmatrix} b^2_{11} & b_{11}b_{21} & b_{11}b_{31} \\ b_{11}b_{21} & b^2_{21} + b^2_{22} & b_{21}b_{31} + b_{22}b_{32} \\ b_{11}b_{31} & b_{21}b_{31} + b_{22}b_{32} & b^2_{31} + b^2_{32} + b^2_{33} \end{pmatrix}.
\end{aligned}$$

We work down the first column of A, compare entries, and discover that

$$\begin{aligned}
a_{11} &= b^2_{11} & b_{11} &= \sqrt{a_{11}} \\
a_{21} &= b_{11}b_{21} & b_{21} &= \frac{a_{21}}{b_{11}} \\
a_{31} &= b_{11}b_{31} & b_{31} &= \frac{a_{31}}{b_{11}}.
\end{aligned}$$

Next we work down the second column of A using previously calculated expressions for b_{21} and b_{31} to find that

$$a_{22} = b_{21}^2 + b_{22}^2 \qquad b_{22} = \left(a_{22} - b_{21}^2\right)^{\frac{1}{2}}$$
$$a_{32} = b_{21}b_{31} + b_{22}b_{32} \qquad b_{32} = \frac{a_{32} - b_{21}b_{31}}{b_{22}}.$$

Finally, we use the third column of A and the previously calculated expressions for b_{31} and b_{32} to determine b_{33} as

$$a_{33} = b_{31}^2 + b_{32}^2 + b_{33}^2 \qquad b_{33} = \left(a_{33} - b_{31}^2 - b_{32}^2\right)^{\frac{1}{2}}.$$

For another example, if

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix},$$

we find that

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

We leave it as an exercise to find similar formulae (involving conjugation) to factor a complex Hermitian positive definite matrix A as $A = BB^*$. The following `Matlab` program implements the Cholesky factorization.

```
function B = Cholesky(A)
n = size(A,1);
B = zeros(n,n);
for j = 1:n-1;
    if j == 1
       B(1,1) = sqrt(A(1,1));
       for i = 2:n
           B(i,1) = A(i,1)/B(1,1);
       end
    else
       B(j,j) = sqrt(A(j,j) - B(j,1:j-1)*B(j,1:j-1)');
       for i = j+1:n
```

```
            B(i,j) = (A(i,j) - B(i,1:j-1)*B(j,1:j-1)')/B(j,j);
        end
    end
end
B(n,n) = sqrt(A(n,n) - B(n,1:n-1)*B(n,1:n-1)');
end
```

If we run the above algorithm on the following matrix

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix},$$

we obtain

$$B = \begin{pmatrix} 2.0000 & 0 & 0 & 0 & 0 \\ 0.5000 & 1.9365 & 0 & 0 & 0 \\ 0 & 0.5164 & 1.9322 & 0 & 0 \\ 0 & 0 & 0.5175 & 1.9319 & 0 \\ 0 & 0 & 0 & 0.5176 & 1.9319 \end{pmatrix}.$$

The Cholesky factorization can be used to solve linear systems $Ax = b$ where A is symmetric positive definite: Solve the two systems $Bw = b$ and $B^\top x = w$.

Remark: It can be shown that this method requires $n^3/6 + O(n^2)$ additions, $n^3/6 + O(n^2)$ multiplications, $n^2/2 + O(n)$ divisions, and $O(n)$ square root extractions. Thus, the Cholesky method requires half of the number of operations required by Gaussian elimination (since Gaussian elimination requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications, and $n^2/2 + O(n)$ divisions). It also requires half of the space (only B is needed, as opposed to both L and U). Furthermore, it can be shown that Cholesky's method is numerically stable (see Trefethen and Bau [176], Lecture 23). In `Matlab` the function `chol` returns the lower-triangular matrix B such that $A = BB^\top$ using the call $B = \text{chol}(A, \text{'lower'})$.

Remark: If $A = BB^\top$, where B is any invertible matrix, then A is symmetric positive definite.

Proof. Obviously, $BB^\top$ is symmetric, and since B is invertible, $B^\top$ is invertible, and from

$$x^\top A x = x^\top B B^\top x = (B^\top x)^\top B^\top x,$$

it is clear that $x^\top A x > 0$ if $x \neq 0$. $\square$

We now give three more criteria for a symmetric matrix to be positive definite.

Proposition 8.11. *Let A be any $n \times n$ real symmetric matrix. The following conditions are equivalent:*

(a) A is positive definite.

(b) All principal minors of A are positive; that is: $\det(A(1:k, 1:k)) > 0$ *for* $k = 1, \ldots, n$ *(Sylvester's criterion).*

(c) A has an LU-factorization and all pivots are positive.

(d) A has an $LDL^\top$-factorization and all pivots in D are positive.

Proof. By Proposition 8.9, if A is symmetric positive definite, then each matrix $A(1:k, 1:k)$ is symmetric positive definite for $k = 1, \ldots, n$. By the Cholesky decomposition, $A(1:k, 1:k) = Q^\top Q$ for some invertible matrix Q, so $\det(A(1:k, 1:k)) = \det(Q)^2 > 0$. This shows that (a) implies (b).

If $\det(A(1:k, 1:k)) > 0$ for $k = 1, \ldots, n$, then each $A(1:k, 1:k)$ is invertible. By Proposition 8.2, the matrix A has an LU-factorization, and since the pivots π_k are given by

$$\pi_k = \begin{cases} a_{11} = \det(A(1:1, 1:1)) & \text{if } k = 1 \\ \dfrac{\det(A(1:k, 1:k))}{\det(A(1:k-1, 1:k-1))} & \text{if } k = 2, \ldots, n, \end{cases}$$

we see that $\pi_k > 0$ for $k = 1, \ldots, n$. Thus (b) implies (c).

Assume A has an LU-factorization and that the pivots are all positive. Since A is symmetric, this implies that A has a factorization of the form

$$A = LDL^\top,$$

with L lower-triangular with 1s on its diagonal, and where D is a diagonal matrix with positive entries on the diagonal (the pivots). This shows that (c) implies (d).

Given a factorization $A = LDL^\top$ with all pivots in D positive, if we form the diagonal matrix

$$\sqrt{D} = \mathrm{diag}(\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$$

and if we let $B = L\sqrt{D}$, then we have

$$A = BB^\top,$$

with B lower-triangular and invertible. By the remark before Proposition 8.11, A is positive definite. Hence, (d) implies (a). $\square$

Criterion (c) yields a simple computational test to check whether a symmetric matrix is positive definite. There is one more criterion for a symmetric matrix to be positive definite:

its eigenvalues must be positive. We will have to learn about the spectral theorem for symmetric matrices to establish this criterion (see Proposition 22.3).

Proposition 8.11 also holds for complex Hermitian positive definite matrices, where in (d), the factorization $LDL^\top$ is replaced by LDL^*.

For more on the stability analysis and efficient implementation methods of Gaussian elimination, LU-factoring and Cholesky factoring, see Demmel [48], Trefethen and Bau [176], Ciarlet [41], Golub and Van Loan [80], Meyer [125], Strang [169, 170], and Kincaid and Cheney [102].

8.10 Reduced Row Echelon Form (RREF)

Gaussian elimination described in Section 8.2 can also be applied to rectangular matrices. This yields a method for determining whether a system $Ax = b$ is solvable and a description of all the solutions when the system is solvable, for any rectangular $m \times n$ matrix A.

It turns out that the discussion is simpler if we rescale all pivots to be 1, and for this we need a third kind of elementary matrix. For any $\lambda \neq 0$, let $E_{i,\lambda}$ be the $n \times n$ diagonal matrix

$$E_{i,\lambda} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \lambda & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix},$$

with $(E_{i,\lambda})_{ii} = \lambda$ $(1 \leq i \leq n)$. Note that $E_{i,\lambda}$ is also given by

$$E_{i,\lambda} = I + (\lambda - 1)e_{i\,i},$$

and that $E_{i,\lambda}$ is invertible with

$$E_{i,\lambda}^{-1} = E_{i,\lambda^{-1}}.$$

Now after $k-1$ elimination steps, if the bottom portion

$$(a_{kk}^{(k)}, a_{k+1k}^{(k)}, \ldots, a_{mk}^{(k)})$$

of the kth column of the current matrix A_k is nonzero so that a pivot π_k can be chosen, after a permutation of rows if necessary, we also divide row k by π_k to obtain the pivot 1, and not only do we zero all the entries $i = k+1, \ldots, m$ in column k, but also all the entries $i = 1, \ldots, k-1$, so that the only nonzero entry in column k is a 1 in row k. These row operations are achieved by multiplication on the left by elementary matrices.

If $a_{kk}^{(k)} = a_{k+1k}^{(k)} = \cdots = a_{mk}^{(k)} = 0$, we move on to column $k+1$.

When the kth column contains a pivot, the kth stage of the procedure for converting a matrix to *rref* consists of the following three steps illustrated below:

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{rescale}}{\Longrightarrow}$$

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & 1 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} 1 & \times & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 1 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{1} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \boldsymbol{\times} & \boldsymbol{\times} & \boldsymbol{\times} \\ 0 & 0 & 0 & \mathbf{0} & \boldsymbol{\times} & \boldsymbol{\times} & \boldsymbol{\times} \\ 0 & 0 & 0 & \mathbf{0} & \boldsymbol{\times} & \boldsymbol{\times} & \boldsymbol{\times} \end{pmatrix}.$$

If the kth column does not contain a pivot, we simply move on to the next column.

The result is that after performing such elimination steps, we obtain a matrix that has a special shape known as a *reduced row echelon matrix*, for short *rref*.

Here is an example illustrating this process: Starting from the matrix

$$A_1 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix},$$

we perform the following steps

$$A_1 \longrightarrow A_2 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 1 & 2 \\ 0 & 2 & 6 & 3 & 7 \end{pmatrix},$$

by subtracting row 1 from row 2 and row 3;

$$A_2 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 2 & 6 & 3 & 7 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow A_3 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & -1/2 & -3/2 \end{pmatrix},$$

after choosing the pivot 2 and permuting row 2 and row 3, dividing row 2 by 2, and subtracting row 2 from row 3;

$$A_3 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix} \longrightarrow A_4 = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 3 & 0 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix},$$

after dividing row 3 by $-1/2$, subtracting row 3 from row 1, and subtracting $(3/2) \times$ row 3 from row 2.

It is clear that columns $1, 2$ and 4 are linearly independent, that column 3 is a linear combination of columns 1 and 2, and that column 5 is a linear combination of columns $1, 2, 4$.

In general, the sequence of steps leading to a reduced echelon matrix is not unique. For example, we could have chosen 1 instead of 2 as the second pivot in matrix A_2. Nevertheless, *the reduced row echelon matrix obtained from any given matrix is unique*; that is, it does not depend on the the sequence of steps that are followed during the reduction process. This fact is not so easy to prove rigorously, but we will do it later.

If we want to solve a linear system of equations of the form $Ax = b$, we apply elementary row operations to both the matrix A and the right-hand side b. To do this conveniently, we form the *augmented matrix* (A, b), which is the $m \times (n+1)$ matrix obtained by adding b as an extra column to the matrix A. For example if

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 2 & 8 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 5 \\ 7 \\ 12 \end{pmatrix},$$

then the augmented matrix is

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}.$$

Now for any matrix M, since

$$M(A, b) = (MA, Mb),$$

performing elementary row operations on (A, b) is equivalent to simultaneously performing operations on both A and b. For example, consider the system

$$\begin{array}{ccccccccc} x_1 & & & + & 2x_3 & + & x_4 & = & 5 \\ x_1 & + & x_2 & + & 5x_3 & + & 2x_4 & = & 7 \\ x_1 & + & 2x_2 & + & 8x_3 & + & 4x_4 & = & 12. \end{array}$$

Its augmented matrix is the matrix

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}$$

considered above, so the reduction steps applied to this matrix yield the system

$$\begin{array}{ccccccc} x_1 & & & + & 2x_3 & & = & 2 \\ & & x_2 & + & 3x_3 & & = & -1 \\ & & & & & x_4 & = & 3. \end{array}$$

This reduced system has the same set of solutions as the original, and obviously x_3 can be chosen arbitrarily. Therefore, our system has infinitely many solutions given by

$$x_1 = 2 - 2x_3, \quad x_2 = -1 - 3x_3, \quad x_4 = 3,$$

where x_3 is arbitrary.

The following proposition shows that the set of solutions of a system $Ax = b$ is preserved by any sequence of row operations.

Proposition 8.12. *Given any $m \times n$ matrix A and any vector $b \in \mathbb{R}^m$, for any sequence of elementary row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $(A', b') = P(A, b)$, then the solutions of $Ax = b$ are the same as the solutions of $A'x = b'$.*

Proof. Since each elementary row operation E_i is invertible, so is P, and since $(A', b') = P(A, b)$, then $A' = PA$ and $b' = Pb$. If x is a solution of the original system $Ax = b$, then multiplying both sides by P we get $PAx = Pb$; that is, $A'x = b'$, so x is a solution of the new system. Conversely, assume that x is a solution of the new system, that is $A'x = b'$. Then because $A' = PA$, $b' = Pb$, and P is invertible, we get

$$Ax = P^{-1}A'x = P^{-1}b' = b,$$

so x is a solution of the original system $Ax = b$. $\square$

Another important fact is this:

Proposition 8.13. *Given an $m \times n$ matrix A, for any sequence of row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $B = PA$, then the subspaces spanned by the rows of A and the rows of B are identical. Therefore, A and B have the same row rank. Furthermore, the matrices A and B also have the same (column) rank.*

Proof. Since $B = PA$, from a previous observation, the rows of B are linear combinations of the rows of A, so the span of the rows of B is a subspace of the span of the rows of A. Since P is invertible, $A = P^{-1}B$, so by the same reasoning the span of the rows of A is a subspace of the span of the rows of B. Therefore, the subspaces spanned by the rows of A and the rows of B are identical, which implies that A and B have the same row rank.

Proposition 8.12 implies that the systems $Ax = 0$ and $Bx = 0$ have the same solutions. Since Ax is a linear combinations of the columns of A and Bx is a linear combinations of the columns of B, the maximum number of linearly independent columns in A is equal to the maximum number of linearly independent columns in B; that is, A and B have the same rank. $\square$

Remark: The subspaces spanned by the columns of A and B can be different! However, their dimension must be the same.

We will show in Section 8.14 that the row rank is equal to the column rank. This will also be proven in Proposition 11.15 Let us now define precisely what is a reduced row echelon matrix.

Definition 8.6. An $m \times n$ matrix A is a *reduced row echelon matrix* iff the following conditions hold:

(a) The first nonzero entry in every row is 1. This entry is called a *pivot*.

(b) The first nonzero entry of row $i + 1$ is to the right of the first nonzero entry of row i.

(c) The entries above a pivot are zero.

If a matrix satisfies the above conditions, we also say that it is in *reduced row echelon form*, for short *rref*.

Note that Condition (b) implies that the entries below a pivot are also zero. For example, the matrix

$$A = \begin{pmatrix} 1 & 6 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

is a reduced row echelon matrix. In general, a matrix in *rref* has the following shape:

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the last row consists of zeros, or

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \end{pmatrix}$$

if the last row contains a pivot.

The following proposition shows that every matrix can be converted to a reduced row echelon form using row operations.

Proposition 8.14. *Given any $m \times n$ matrix A, there is a sequence of row operations $E_1, \ldots, E_k$ such that if $P = E_k \cdots E_1$, then $U = PA$ is a reduced row echelon matrix.*

Proof. We proceed by induction on m. If $m = 1$, then either all entries on this row are zero, so $A = 0$, or if a_j is the first nonzero entry in A, let $P = (a_j^{-1})$ (a 1×1 matrix); clearly, PA is a reduced row echelon matrix.

Let us now assume that $m \geq 2$. If $A = 0$, we are done, so let us assume that $A \neq 0$. Since $A \neq 0$, there is a leftmost column j which is nonzero, so pick any pivot $\pi = a_{ij}$ in the jth column, permute row i and row 1 if necessary, multiply the new first row by π^{-1}, and clear out the other entries in column j by subtracting suitable multiples of row 1. At the end of this process, we have a matrix A_1 that has the following shape:

$$A_1 = \begin{pmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * \\ 0 & \cdots & 0 & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{pmatrix},$$

where $*$ stands for an arbitrary scalar, or more concisely

$$A_1 = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & D \end{pmatrix},$$

where D is a $(m-1) \times (n-j)$ matrix (and B is a $1 \times n - j$ matrix). If $j = n$, we are done. Otherwise, by the induction hypothesis applied to D, there is a sequence of row operations that converts D to a reduced row echelon matrix R', and these row operations do not affect the first row of A_1, which means that A_1 is reduced to a matrix of the form

$$R = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & R' \end{pmatrix}.$$

Because R' is a reduced row echelon matrix, the matrix R satisfies Conditions (a) and (b) of the reduced row echelon form. Finally, the entries above all pivots in R' can be cleared out by subtracting suitable multiples of the rows of R' containing a pivot. The resulting matrix also satisfies Condition (c), and the induction step is complete. $\square$

Remark: There is a `Matlab` function named `rref` that converts any matrix to its reduced row echelon form.

If A is any matrix and if R is a reduced row echelon form of A, the second part of Proposition 8.13 can be sharpened a little, since the structure of a reduced row echelon matrix makes it clear that its rank is equal to the number of pivots.

Proposition 8.15. *The rank of a matrix A is equal to the number of pivots in its rref R.*

8.11 RREF, Free Variables, and Homogenous Linear Systems

Given a system of the form $Ax = b$, we can apply the reduction procedure to the augmented matrix (A, b) to obtain a reduced row echelon matrix (A', b') such that the system $A'x = b'$ has the same solutions as the original system $Ax = b$. The advantage of the reduced system $A'x = b'$ is that there is a simple test to check whether this system is solvable, and to find its solutions if it is solvable.

Indeed, if any row of the matrix A' is zero and if the corresponding entry in b' is nonzero, then it is a pivot and we have the "equation"

$$0 = 1,$$

which means that the system $A'x = b'$ has no solution. On the other hand, if there is no pivot in b', then for every row i in which $b'_i \neq 0$, there is some column j in A' where the entry on row i is 1 (a pivot). Consequently, we can assign arbitrary values to the variable x_k if column k does not contain a pivot, and then solve for the pivot variables.

For example, if we consider the reduced row echelon matrix

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

there is no solution to $A'x = b'$ because the third equation is $0 = 1$. On the other hand, the reduced system

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

has solutions. We can pick the variables x_2, x_4 corresponding to nonpivot columns arbitrarily, and then solve for x_3 (using the second equation) and x_1 (using the first equation).

The above reasoning proves the following theorem:

Theorem 8.16. *Given any system $Ax = b$ where A is a $m \times n$ matrix, if the augmented matrix (A, b) is a reduced row echelon matrix, then the system $Ax = b$ has a solution iff there is no pivot in b. In that case, an arbitrary value can be assigned to the variable x_j if column j does not contain a pivot.*

Definition 8.7. Nonpivot variables are often called *free variables*.

Putting Proposition 8.14 and Theorem 8.16 together we obtain a criterion to decide whether a system $Ax = b$ has a solution: Convert the augmented system (A, b) to a row reduced echelon matrix (A', b') and check whether b' has no pivot.

Remark: When writing a program implementing row reduction, we may stop when the last column of the matrix A is reached. In this case, the test whether the system $Ax = b$ is solvable is that the row-reduced matrix A' has no zero row of index $i > r$ such that $b'_i \neq 0$ (where r is the number of pivots, and b' is the row-reduced right-hand side).

If we have a *homogeneous system* $Ax = 0$, which means that $b = 0$, of course $x = 0$ is always a solution, but Theorem 8.16 implies that if the system $Ax = 0$ has more variables than equations, then it has some nonzero solution (we call it a *nontrivial solution*).

Proposition 8.17. *Given any homogeneous system $Ax = 0$ of m equations in n variables, if $m < n$, then there is a nonzero vector $x \in \mathbb{R}^n$ such that $Ax = 0$.*

Proof. Convert the matrix A to a reduced row echelon matrix A'. We know that $Ax = 0$ iff $A'x = 0$. If r is the number of pivots of A', we must have $r \leq m$, so by Theorem 8.16 we may assign arbitrary values to $n - r > 0$ nonpivot variables and we get nontrivial solutions. $\square$

Theorem 8.16 can also be used to characterize when a square matrix is invertible. First, note the following simple but important fact:

If a square $n \times n$ matrix A is a row reduced echelon matrix, then either A is the identity or the bottom row of A is zero.

Proposition 8.18. *Let A be a square matrix of dimension n. The following conditions are equivalent:*

(a) The matrix A can be reduced to the identity by a sequence of elementary row operations.

(b) The matrix A is a product of elementary matrices.

(c) The matrix A is invertible.

(d) The system of homogeneous equations $Ax = 0$ has only the trivial solution $x = 0$.

Proof. First we prove that (a) implies (b). If (a) can be reduced to the identity by a sequence of row operations $E_1, \ldots, E_p$, this means that $E_p \cdots E_1 A = I$. Since each E_i is invertible, we get

$$A = E_1^{-1} \cdots E_p^{-1},$$

where each E_i^{-1} is also an elementary row operation, so (b) holds. Now if (b) holds, since elementary row operations are invertible, A is invertible and (c) holds. If A is invertible, we already observed that the homogeneous system $Ax = 0$ has only the trivial solution $x = 0$, because from $Ax = 0$, we get $A^{-1}Ax = A^{-1}0$; that is, $x = 0$. It remains to prove that (d) implies (a) and for this we prove the contrapositive: if (a) does not hold, then (d) does not hold.

Using our basic observation about reducing square matrices, if A does not reduce to the identity, then A reduces to a row echelon matrix A' whose bottom row is zero. Say $A' = PA$, where P is a product of elementary row operations. Because the bottom row of A' is zero, the system $A'x = 0$ has at most $n-1$ nontrivial equations, and by Proposition 8.17, this system has a nontrivial solution x. But then, $Ax = P^{-1}A'x = 0$ with $x \neq 0$, contradicting the fact that the system $Ax = 0$ is assumed to have only the trivial solution. Therefore, (d) implies (a) and the proof is complete. $\square$

Proposition 8.18 yields a method for computing the inverse of an invertible matrix A: reduce A to the identity using elementary row operations, obtaining

$$E_p \cdots E_1 A = I.$$

Multiplying both sides by A^{-1} we get

$$A^{-1} = E_p \cdots E_1.$$

From a practical point of view, we can build up the product $E_p \cdots E_1$ by reducing to row echelon form the augmented $n \times 2n$ matrix (A, I_n) obtained by adding the n columns of the identity matrix to A. This is just another way of performing the Gauss–Jordan procedure.

Here is an example: let us find the inverse of the matrix

$$A = \begin{pmatrix} 5 & 4 \\ 6 & 5 \end{pmatrix}.$$

We form the 2×4 block matrix

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix}$$

and apply elementary row operations to reduce A to the identity. For example:

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting row 1 from row 2,

$$\begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting $4 \times$ row 2 from row 1,

$$\begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 0 & 1 & -6 & 5 \end{pmatrix} = (I, A^{-1}),$$

by subtracting row 1 from row 2. Thus

$$A^{-1} = \begin{pmatrix} 5 & -4 \\ -6 & 5 \end{pmatrix}.$$

Proposition 8.18 can also be used to give an elementary proof of the fact that if a square matrix A has a left inverse B (resp. a right inverse B), so that $BA = I$ (resp. $AB = I$), then A is invertible and $A^{-1} = B$. This is an interesting exercise, try it!

8.12 Uniqueness of RREF Form

For the sake of completeness, we prove that the reduced row echelon form of a matrix is unique. The neat proof given below is borrowed and adapted from W. Kahan.

Proposition 8.19. *Let A be any $m \times n$ matrix. If U and V are two reduced row echelon matrices obtained from A by applying two sequences of elementary row operations $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, so that*

$$U = E_p \cdots E_1 A \quad \text{and} \quad V = F_q \cdots F_1 A,$$

then $U = V$. In other words, the reduced row echelon form of any matrix is unique.

Proof. Let

$$C = E_p \cdots E_1 F_1^{-1} \cdots F_q^{-1}$$

so that

$$U = CV \quad \text{and} \quad V = C^{-1} U.$$

Recall from Proposition 8.13 that U and V have the same row rank r, and since U and V are in *rref*, this is the number of nonzero rows in both U and V. We prove by induction on n that $U = V$ (and that the first r columns of C are the first r columns in I_m). If $r = 0$ then $A = U = V = 0$ and the result is trivial. We now assume that $r \geq 1$.

Let ℓ_j^n denote the jth column of the identity matrix I_n, and let $u_j = U\ell_j^n$, $v_j = V\ell_j^n$, $c_j = C\ell_j^m$, and $a_j = A\ell_j^n$, be the jth column of U, V, C, and A respectively.

First I claim that $u_j = 0$ iff $v_j = 0$ iff $a_j = 0$.

Indeed, if $v_j = 0$, then (because $U = CV$) $u_j = Cv_j = 0$, and if $u_j = 0$, then $v_j = C^{-1}u_j = 0$. Since $U = E_p \cdots E_1 A$, we also get $a_j = 0$ iff $u_j = 0$.

Therefore, we may simplify our task by striking out columns of zeros from U, V, and A, since they will have corresponding indices. We still use n to denote the number of columns of A. Observe that because U and V are reduced row echelon matrices with no zero columns, we must have $u_1 = v_1 = \ell_1^m$.

Claim. If U and V are reduced row echelon matrices without zero columns such that $U = CV$, for all $k \geq 1$, if $k \leq m$, then ℓ_k^m occurs in U iff ℓ_k^m occurs in V, and if ℓ_k^m does occur in U, then

1. ℓ_k^m occurs for the same column index j_k in both U and V;

2. the first j_k columns of U and V match;

3. the subsequent columns in U and V (of column index $> j_k$) whose coordinates of index $k+1$ through m are all equal to 0 also match. Let n_k be the rightmost index of such a column, with $n_k = j_k$ if there is none.

4. the first k columns of C match the first k columns of I_m.

We prove this claim by induction on k.

For the base case $k = 1$, we already know that $u_1 = v_1 = \ell_1^m$. We also have

$$c_1 = C\ell_1^m = Cv_1 = u_1 = \ell_1^m.$$

If $v_j = \lambda \ell_1^m$ for some $\lambda \in \mathbb{R}$, then

$$u_j = U\ell_j^n = CV\ell_j^n = Cv_j = \lambda C\ell_1^m = \lambda c_1 = \lambda \ell_1^m = v_j.$$

A similar argument using C^{-1} shows that if $u_j = \lambda \ell_1^m$, then $v_j = u_j$. Therefore, all the columns of U and V proportional to ℓ_1^m match, which establishes the base case. Observe that if ℓ_2^m appears in U, then it must appear in both U and V for the same index, and if not then $n_1 = n$ and $U = V$.

Next us now prove the induction step. If $n_k = n$, then $U = V$ and we are done. If $k = r$, then C is a block matrix of the form

$$C = \begin{pmatrix} I_r & B \\ 0_{m-r,r} & C \end{pmatrix}$$

and since the last $m - r$ rows of both U and V are zero rows, C acts as the identity on the first r rows, and so $U = V$. Otherwise $k < r$, $n_k < n$, and ℓ_{k+1}^m appears in both U and V, in which case, by (2) and (3) of the induction hypothesis, it appears in both U and V for the same index, say j_{k+1}. Thus, $u_{j_{k+1}} = v_{j_{k+1}} = \ell_{k+1}^m$. It follows that

$$c_{k+1} = C\ell_{k+1}^m = Cv_{j_{k+1}} = u_{j_{k+1}} = \ell_{k+1}^m,$$

so the first $k+1$ columns of C match the first $k+1$ columns of I_m.

Consider any subsequent column v_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish. Then v_j is a linear combination of columns of V to the left of v_j, so

$$u_j = Cv_j = v_j.$$

because the first $k+1$ columns of C match the first $k+1$ column of I_m. Similarly, any subsequent column u_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish is equal to v_j. Therefore, all the subsequent columns in U and V (of index $> j_{k+1}$) whose elements beyond the $(k+1)$th all vanish also match, which completes the induction hypothesis. $\square$

Remark: Observe that $C = E_p \cdots E_1 F_1^{-1} \cdots F_q^{-1}$ is *not* necessarily the identity matrix I_m. However, $C = I_m$ if $r = m$ (A has row rank m).

The reduction to row echelon form also provides a method to describe the set of solutions of a linear system of the form $Ax = b$.

8.13 Solving Linear Systems Using RREF

First we have the following simple result.

Proposition 8.20. *Let A be any $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. If the system $Ax = b$ has a solution, then the set Z of all solutions of this system is the set*

$$Z = x_0 + \mathrm{Ker}\,(A) = \{x_0 + x \mid Ax = 0\},$$

where $x_0 \in \mathbb{R}^n$ is any solution of the system $Ax = b$, which means that $Ax_0 = b$ (x_0 is called a special solution or a particular solution), and where $\mathrm{Ker}\,(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$, the set of solutions of the homogeneous system associated with $Ax = b$.

Proof. Assume that the system $Ax = b$ is solvable and let x_0 and x_1 be any two solutions so that $Ax_0 = b$ and $Ax_1 = b$. Subtracting the first equation from the second, we get

$$A(x_1 - x_0) = 0,$$

which means that $x_1 - x_0 \in \mathrm{Ker}\,(A)$. Therefore, $Z \subseteq x_0 + \mathrm{Ker}\,(A)$, where x_0 is a special solution of $Ax = b$. Conversely, if $Ax_0 = b$, then for any $z \in \mathrm{Ker}\,(A)$, we have $Az = 0$, and so

$$A(x_0 + z) = Ax_0 + Az = b + 0 = b,$$

which shows that $x_0 + \mathrm{Ker}\,(A) \subseteq Z$. Therefore, $Z = x_0 + \mathrm{Ker}\,(A)$. $\square$

Given a linear system $Ax = b$, reduce the augmented matrix (A, b) to its row echelon form (A', b'). As we showed before, the system $Ax = b$ has a solution iff b' contains no pivot. Assume that this is the case. Then, if (A', b') has r pivots, which means that A' has r pivots since b' has no pivot, we know that the first r columns of I_m appear in A'.

We can permute the columns of A' and renumber the variables in x correspondingly so that the first r columns of I_m match the first r columns of A', and then our reduced echelon matrix is of the form (R, b') with

$$R = \begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}$$

and

$$b' = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix},$$

where F is a $r \times (n - r)$ matrix and $d \in \mathbb{R}^r$. Note that R has $m - r$ zero rows.

Then because

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix} = b',$$

we see that

$$x_0 = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix}$$

is a special solution of $Rx = b'$, and thus to $Ax = b$. In other words, we get a special solution by assigning the first r components of b' to the pivot variables and setting the nonpivot variables (the *free variables*) to zero.

Here is an example of the preceding construction taken from Kumpel and Thorpe [107]. The linear system

$$\begin{aligned} x_1 - x_2 + x_3 + x_4 - 2x_5 &= -1 \\ -2x_1 + 2x_2 - x_3 + x_5 &= 2 \\ x_1 - x_2 + 2x_3 + 3x_4 - 5x_5 &= -1, \end{aligned}$$

is represented by the augmented matrix

$$(A, b) = \begin{pmatrix} 1 & -1 & 1 & 1 & -2 & -1 \\ -2 & 2 & -1 & 0 & 1 & 2 \\ 1 & -1 & 2 & 3 & -5 & -1 \end{pmatrix},$$

where A is a 3×5 matrix. The reader should find that the row echelon form of this system is

$$(A', b') = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 2 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The 3×5 matrix A' has rank 2. We permute the second and third columns (which is equivalent to interchanging variables x_2 and x_3) to form

$$R = \begin{pmatrix} I_2 & F \\ 0_{1,2} & 0_{1,3} \end{pmatrix}, \qquad F = \begin{pmatrix} -1 & -1 & 1 \\ 0 & 2 & -3 \end{pmatrix}.$$

Then a special solution to this linear system is given by

$$x_0 = \begin{pmatrix} d \\ 0_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0_3 \end{pmatrix}.$$

We can also find a basis of the kernel (nullspace) of A using F. If $x = (u, v)$ is in the kernel of A, with $u \in \mathbb{R}^r$ and $v \in \mathbb{R}^{n-r}$, then x is also in the kernel of R, which means that $Rx = 0$; that is,

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u + Fv \\ 0_{m-r} \end{pmatrix} = \begin{pmatrix} 0_r \\ 0_{m-r} \end{pmatrix}.$$

Therefore, $u = -Fv$, and $\mathrm{Ker}\,(A)$ consists of all vectors of the form

$$\begin{pmatrix} -Fv \\ v \end{pmatrix} = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix} v,$$

for any arbitrary $v \in \mathbb{R}^{n-r}$. It follows that the $n-r$ columns of the matrix

$$N = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix}$$

form a basis of the kernel of A. This is because N contains the identity matrix I_{n-r} as a submatrix, so the columns of N are linearly independent. In summary, if $N^1, \ldots, N^{n-r}$ are the columns of N, then the general solution of the equation $Ax = b$ is given by

$$x = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} + x_{r+1} N^1 + \cdots + x_n N^{n-r},$$

where $x_{r+1}, \ldots, x_n$ are the free variables; that is, the nonpivot variables.

Going back to our example from Kumpel and Thorpe [107], we see that

$$N = \begin{pmatrix} -F \\ I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & -2 & 3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since earlier we permuted the second and the third column, row 2 and row 3 need to be swapped so the general solution in terms of the original variables is given by

$$x = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} -1 \\ 0 \\ 3 \\ 0 \\ 1 \end{pmatrix}.$$

In the general case where the columns corresponding to pivots are mixed with the columns corresponding to free variables, we find the special solution as follows. Let $i_1 < \cdots < i_r$ be the indices of the columns corresponding to pivots. Assign b'_k to the pivot variable x_{i_k} for $k = 1, \ldots, r$, and set all other variables to 0. To find a basis of the kernel, we form the $n-r$ vectors N^k obtained as follows. Let $j_1 < \cdots < j_{n-r}$ be the indices of the columns corresponding to free variables. For every column j_k corresponding to a free variable $(1 \leq k \leq n-r)$, form the vector N^k defined so that the entries $N^k_{i_1}, \ldots, N^k_{i_r}$ are equal to the negatives of the first r entries in column j_k (flip the sign of these entries); let $N^k_{j_k} = 1$, and set all other entries to zero. Schematically, if the column of index j_k (corresponding to the free variable x_{j_k}) is

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

then the vector N^k is given by

$$\begin{array}{c} 1 \\ \vdots \\ i_1 - 1 \\ i_1 \\ i_1 + 1 \\ \vdots \\ i_r - 1 \\ i_r \\ i_r + 1 \\ \vdots \\ j_k - 1 \\ j_k \\ j_k + 1 \\ \vdots \\ n \end{array} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\alpha_1 \\ 0 \\ \vdots \\ 0 \\ -\alpha_r \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The presence of the 1 in position j_k guarantees that $N^1, \ldots, N^{n-r}$ are linearly independent.

As an illustration of the above method, consider the problem of finding a basis of the subspace V of $n \times n$ matrices $A \in \mathrm{M}_n(\mathbb{R})$ satisfying the following properties:

1. The sum of the entries in every row has the same value (say c_1);

2. The sum of the entries in every column has the same value (say c_2).

It turns out that $c_1 = c_2$ and that the $2n-2$ equations corresponding to the above conditions are linearly independent. We leave the proof of these facts as an interesting exercise. It can be shown using the duality theorem (Theorem 11.4) that the dimension of the space V of matrices satisying the above equations is $n^2 - (2n - 2)$. Let us consider the case $n = 4$. There are 6 equations, and the space V has dimension 10. The equations are

$$\begin{aligned}
a_{11} + a_{12} + a_{13} + a_{14} - a_{21} - a_{22} - a_{23} - a_{24} &= 0 \\
a_{21} + a_{22} + a_{23} + a_{24} - a_{31} - a_{32} - a_{33} - a_{34} &= 0 \\
a_{31} + a_{32} + a_{33} + a_{34} - a_{41} - a_{42} - a_{43} - a_{44} &= 0 \\
a_{11} + a_{21} + a_{31} + a_{41} - a_{12} - a_{22} - a_{32} - a_{42} &= 0 \\
a_{12} + a_{22} + a_{32} + a_{42} - a_{13} - a_{23} - a_{33} - a_{43} &= 0 \\
a_{13} + a_{23} + a_{33} + a_{43} - a_{14} - a_{24} - a_{34} - a_{44} &= 0,
\end{aligned}$$

and the corresponding matrix is

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The result of performing the reduction to row echelon form yields the following matrix in rref:

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

The list *pivlist* of indices of the pivot variables and the list *freelist* of indices of the free variables is given by

$$\begin{aligned} \mathit{pivlist} &= (1, 2, 3, 4, 5, 9), \\ \mathit{freelist} &= (\mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{10}, \mathbf{11}, \mathbf{12}, \mathbf{13}, \mathbf{14}, \mathbf{15}, \mathbf{16}). \end{aligned}$$

After applying the algorithm to find a basis of the kernel of U, we find the following 16×10 matrix

$$BK = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -2 & -1 & -1 & -1 \\ -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \end{pmatrix}.$$

The reader should check that that in each column j of BK, the lowest bold 1 belongs to the row whose index is the jth element in *freelist*, and that in each column j of BK, the signs of the entries whose indices belong to *pivlist* are the flipped signs of the 6 entries in the column U corresponding to the jth index in *freelist*. We can now read off from BK the 4×4 matrices that form a basis of V: every column of BK corresponds to a matrix whose rows have been concatenated. We get the following 10 matrices:

$$M_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_5 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_6 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} -2 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad M_8 = \begin{pmatrix} -1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad M_9 = \begin{pmatrix} -1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$M_{10} = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Recall that a *magic square* is a square matrix that satisfies the two conditions about the sum of the entries in each row and in each column to be the same number, and also the additional two constraints that the main descending and the main ascending diagonals add up to this common number. Furthermore, the entries are also required to be positive integers. For $n = 4$, the additional two equations are

$$\begin{aligned} a_{22} + a_{33} + a_{44} - a_{12} - a_{13} - a_{14} &= 0 \\ a_{41} + a_{32} + a_{23} - a_{11} - a_{12} - a_{13} &= 0, \end{aligned}$$

and the 8 equations stating that a matrix is a magic square are linearly independent. Again, by running row elimination, we get a basis of the "generalized magic squares" whose entries are not restricted to be positive integers. We find a basis of 8 matrices. For $n = 3$, we find a basis of 3 matrices.

A magic square is said to be *normal* if its entries are precisely the integers $1, 2 \ldots, n^2$. Then since the sum of these entries is

$$1 + 2 + 3 + \cdots + n^2 = \frac{n^2(n^2+1)}{2},$$

and since each row (and column) sums to the same number, this common value (the *magic sum*) is

$$\frac{n(n^2+1)}{2}.$$

It is easy to see that there are no normal magic squares for $n = 2$. For $n = 3$, the magic sum is 15, for $n = 4$, it is 34, and for $n = 5$, it is 65.

In the case $n = 3$, we have the additional condition that the rows and columns add up to 15, so we end up with a solution parametrized by two numbers x_1, x_2; namely,

$$\begin{pmatrix} x_1 + x_2 - 5 & 10 - x_2 & 10 - x_1 \\ 20 - 2x_1 - x_2 & 5 & 2x_1 + x_2 - 10 \\ x_1 & x_2 & 15 - x_1 - x_2 \end{pmatrix}.$$

Thus, in order to find a normal magic square, we have the additional inequality constraints

$$\begin{aligned} x_1 + x_2 &> 5 \\ x_1 &< 10 \\ x_2 &< 10 \\ 2x_1 + x_2 &< 20 \\ 2x_1 + x_2 &> 10 \\ x_1 &> 0 \\ x_2 &> 0 \\ x_1 + x_2 &< 15, \end{aligned}$$

and all 9 entries in the matrix must be distinct. After a tedious case analysis, we discover the remarkable fact that there is a unique normal magic square (up to rotations and reflections):

$$\begin{pmatrix} 2 & 7 & 6 \\ 9 & 5 & 1 \\ 4 & 3 & 8 \end{pmatrix}.$$

It turns out that there are 880 different normal magic squares for $n = 4$, and $275,305,224$ normal magic squares for $n = 5$ (up to rotations and reflections). Even for $n = 4$, it takes a fair amount of work to enumerate them all! Finding the number of magic squares for $n > 5$ is an open problem!

8.14 Elementary Matrices and Columns Operations

Instead of performing elementary row operations on a matrix A, we can perform elementary columns operations, which means that we multiply A by elementary matrices on the *right*. As elementary row and column operations, $P(i, k)$, $E_{i,j;\beta}$, $E_{i,\lambda}$ perform the following actions:

1. As a row operation, $P(i,k)$ permutes row i and row k.

2. As a column operation, $P(i,k)$ permutes column i and column k.

3. The inverse of $P(i,k)$ is $P(i,k)$ itself.

4. As a row operation, $E_{i,j;\beta}$ adds β times row j to row i.

5. As a column operation, $E_{i,j;\beta}$ adds β times column i to column j (note the switch in the indices).

6. The inverse of $E_{i,j;\beta}$ is $E_{i,j;-\beta}$.

7. As a row operation, $E_{i,\lambda}$ multiplies row i by λ.

8. As a column operation, $E_{i,\lambda}$ multiplies column i by λ.

9. The inverse of $E_{i,\lambda}$ is $E_{i,\lambda^{-1}}$.

We can define the notion of a reduced column echelon matrix and show that every matrix can be reduced to a unique reduced column echelon form. Now given any $m \times n$ matrix A, if we first convert A to its reduced row echelon form R, it is easy to see that we can apply elementary column operations that will reduce R to a matrix of the form

$$\begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

where r is the number of pivots (obtained during the row reduction). Therefore, for every $m \times n$ matrix A, there exist two sequences of elementary matrices $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, such that

$$E_p \cdots E_1 A F_1 \cdots F_q = \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}.$$

The matrix on the right-hand side is called the *rank normal form* of A. Clearly, r is the rank of A. As a corollary we obtain the following important result whose proof is immediate.

Proposition 8.21. *A matrix A and its transpose $A^\top$ have the same rank.*

8.15 Transvections and Dilatations ⊛

In this section we characterize the linear isomorphisms of a vector space E that leave every vector in some hyperplane fixed. These maps turn out to be the linear maps that are represented in some suitable basis by elementary matrices of the form $E_{i,j;\beta}$ (transvections) or $E_{i,\lambda}$ (dilatations). Furthermore, the transvections generate the group $\mathbf{SL}(E)$, and the dilatations generate the group $\mathbf{GL}(E)$.

Let H be any hyperplane in E, and pick some (nonzero) vector $v \in E$ such that $v \notin H$, so that

$$E = H \oplus Kv.$$

Assume that $f \colon E \to E$ is a linear isomorphism such that $f(u) = u$ for all $u \in H$, and that f is not the identity. We have

$$f(v) = h + \alpha v, \quad \text{for some } h \in H \text{ and some } \alpha \in K,$$

with $\alpha \neq 0$, because otherwise we would have $f(v) = h = f(h)$ since $h \in H$, contradicting the injectivity of f ($v \neq h$ since $v \notin H$). For any $x \in E$, if we write

$$x = y + tv, \quad \text{for some } y \in H \text{ and some } t \in K,$$

then

$$f(x) = f(y) + f(tv) = y + tf(v) = y + th + t\alpha v,$$

and since $\alpha x = \alpha y + t\alpha v$, we get

$$\begin{aligned} f(x) - \alpha x &= (1 - \alpha)y + th \\ f(x) - x &= t(h + (\alpha - 1)v). \end{aligned}$$

Observe that if E is finite-dimensional, by picking a basis of E consisting of v and basis vectors of H, then the matrix of f is a lower triangular matrix whose diagonal entries are all 1 except the first entry which is equal to α. Therefore, $\det(f) = \alpha$.

Case 1. $\alpha \neq 1$.

We have $f(x) = \alpha x$ iff $(1 - \alpha)y + th = 0$ iff

$$y = \frac{t}{\alpha - 1} h.$$

Then if we let $w = h + (\alpha - 1)v$, for $y = (t/(\alpha - 1))h$, we have

$$x = y + tv = \frac{t}{\alpha - 1} h + tv = \frac{t}{\alpha - 1}(h + (\alpha - 1)v) = \frac{t}{\alpha - 1} w,$$

which shows that $f(x) = \alpha x$ iff $x \in Kw$. Note that $w \notin H$, since $\alpha \neq 1$ and $v \notin H$. Therefore,

$$E = H \oplus Kw,$$

and f is the identity on H and a magnification by α on the line $D = Kw$.

Definition 8.8. Given a vector space E, for any hyperplane H in E, any nonzero vector $u \in E$ such that $u \notin H$, and any scalar $\alpha \neq 0, 1$, a linear map f such that $f(x) = x$ for all $x \in H$ and $f(x) = \alpha x$ for every $x \in D = Ku$ is called a *dilatation of hyperplane H, direction D, and scale factor α.*

If π_H and π_D are the projections of E onto H and D, then we have

$$f(x) = \pi_H(x) + \alpha\pi_D(x).$$

The inverse of f is given by

$$f^{-1}(x) = \pi_H(x) + \alpha^{-1}\pi_D(x).$$

When $\alpha = -1$, we have $f^2 = \mathrm{id}$, and f is a symmetry about the hyperplane H in the direction D. This situation includes orthogonal reflections about H.

Case 2. $\alpha = 1$.

In this case,

$$f(x) - x = th,$$

that is, $f(x) - x \in Kh$ for all $x \in E$. Assume that the hyperplane H is given as the kernel of some linear form φ, and let $a = \varphi(v)$. We have $a \neq 0$, since $v \notin H$. For any $x \in E$, we have

$$\varphi(x - a^{-1}\varphi(x)v) = \varphi(x) - a^{-1}\varphi(x)\varphi(v) = \varphi(x) - \varphi(x) = 0,$$

which shows that $x - a^{-1}\varphi(x)v \in H$ for all $x \in E$. Since every vector in H is fixed by f, we get

$$\begin{aligned} x - a^{-1}\varphi(x)v &= f(x - a^{-1}\varphi(x)v) \\ &= f(x) - a^{-1}\varphi(x)f(v), \end{aligned}$$

so

$$f(x) = x + \varphi(x)(f(a^{-1}v) - a^{-1}v).$$

Since $f(z) - z \in Kh$ for all $z \in E$, we conclude that $u = f(a^{-1}v) - a^{-1}v = \beta h$ for some $\beta \in K$, so $\varphi(u) = 0$, and we have

$$f(x) = x + \varphi(x)u, \quad \varphi(u) = 0. \tag{$*$}$$

A linear map defined as above is denoted by $\tau_{\varphi,u}$.

Conversely for any linear map $f = \tau_{\varphi,u}$ given by Equation ($*$), where φ is a nonzero linear form and u is some vector $u \in E$ such that $\varphi(u) = 0$, if $u = 0$, then f is the identity, so assume that $u \neq 0$. If so, we have $f(x) = x$ iff $\varphi(x) = 0$, that is, iff $x \in H$. We also claim that the inverse of f is obtained by changing u to $-u$. Actually, we check the slightly more general fact that

$$\tau_{\varphi,u} \circ \tau_{\varphi,w} = \tau_{\varphi,u+w}.$$

Indeed, using the fact that $\varphi(w) = 0$, we have

$$\begin{aligned} \tau_{\varphi,u}(\tau_{\varphi,w}(x)) &= \tau_{\varphi,w}(x) + \varphi(\tau_{\varphi,w}(x))u \\ &= \tau_{\varphi,w}(x) + (\varphi(x) + \varphi(x)\varphi(w))u \\ &= \tau_{\varphi,w}(x) + \varphi(x)u \\ &= x + \varphi(x)w + \varphi(x)u \\ &= x + \varphi(x)(u + w). \end{aligned}$$

For $v = -u$, we have $\tau_{\varphi,u+v} = \varphi_{\varphi,0} = \mathrm{id}$, so $\tau_{\varphi,u}^{-1} = \tau_{\varphi,-u}$, as claimed.

Therefore, we proved that every linear isomorphism of E that leaves every vector in some hyperplane H fixed and has the property that $f(x) - x \in H$ for all $x \in E$ is given by a map $\tau_{\varphi,u}$ as defined by Equation $(*)$, where φ is some nonzero linear form defining H and u is some vector in H. We have $\tau_{\varphi,u} = \mathrm{id}$ iff $u = 0$.

Definition 8.9. Given any hyperplane H in E, for any nonzero nonlinear form $\varphi \in E^*$ defining H (which means that $H = \mathrm{Ker}\,(\varphi)$) and any nonzero vector $u \in H$, the linear map $f = \tau_{\varphi,u}$ given by

$$\tau_{\varphi,u}(x) = x + \varphi(x)u, \quad \varphi(u) = 0,$$

for all $x \in E$ is called a *transvection of hyperplane H and direction u*. The map $f = \tau_{\varphi,u}$ leaves every vector in H fixed, and $f(x) - x \in Ku$ for all $x \in E$.

The above arguments show the following result.

Proposition 8.22. *Let $f\colon E \to E$ be a bijective linear map and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If there is some nonzero vector $u \in E$ such that $u \notin H$ and $f(u) - u \in H$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Proof. Using the notation as above, for some $v \notin H$, we have $f(v) = h + \alpha v$ with $\alpha \neq 0$, and write $u = y + tv$ with $y \in H$ and $t \neq 0$ since $u \notin H$. If $f(u) - u \in H$, from

$$f(u) - u = t(h + (\alpha - 1)v),$$

we get $(\alpha - 1)v \in H$, and since $v \notin H$, we must have $\alpha = 1$, and we proved that f is a transvection. Otherwise, $\alpha \neq 0, 1$, and we proved that f is a dilatation. $\square$

If E is finite-dimensional, then $\alpha = \det(f)$, so we also have the following result.

Proposition 8.23. *Let $f\colon E \to E$ be a bijective linear map of a finite-dimensional vector space E and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If $\det(f) = 1$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Suppose that f is a dilatation of hyperplane H and direction u, and say $\det(f) = \alpha \neq 0, 1$. Pick a basis $(u, e_2, \ldots, e_n)$ of E where $(e_2, \ldots, e_n)$ is a basis of H. Then the matrix of f is of the form

$$\begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{1,\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,\alpha}$ with $\alpha \neq 0, 1$ is a dilatation.

Now, assume that f is a transvection of hyperplane H and direction $u \in H$. Pick some $v \notin H$, and pick some basis $(u, e_3, \ldots, e_n)$ of H, so that $(v, u, e_3, \ldots, e_n)$ is a basis of E. Since $f(v) - v \in Ku$, the matrix of f is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{2,1;\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,j;\alpha}$ $(\alpha \neq 0)$ is a transvection.

The following proposition is an interesting exercise that requires good mastery of the elementary row operations $E_{i,j;\beta}$; see Problems 8.10 and 8.11.

Proposition 8.24. *Given any invertible $n \times n$ matrix A, there is a matrix S such that*

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$; that is, S is a composition of transvections.

Surprisingly, every transvection is the composition of two dilatations!

Proposition 8.25. *If the field K is not of characteristic 2, then every transvection f of hyperplane H can be written as $f = d_2 \circ d_1$, where d_1, d_2 are dilatations of hyperplane H, where the direction of d_1 can be chosen arbitrarily.*

Proof. Pick some dilatation d_1 of hyperplane H and scale factor $\alpha \neq 0, 1$. Then, $d_2 = f \circ d_1^{-1}$ leaves every vector in H fixed, and $\det(d_2) = \alpha^{-1} \neq 1$. By Proposition 8.23, the linear map d_2 is a dilatation of hyperplane H, and we have $f = d_2 \circ d_1$, as claimed. ☐

Observe that in Proposition 8.25, we can pick $\alpha = -1$; that is, every transvection of hyperplane H is the compositions of two symmetries about the hyperplane H, one of which can be picked arbitrarily.

Remark: Proposition 8.25 holds as long as $K \neq \{0, 1\}$.

The following important result is now obtained.

Theorem 8.26. *Let E be any finite-dimensional vector space over a field K of characteristic not equal to 2. Then the group $\mathbf{SL}(E)$ is generated by the transvections, and the group $\mathbf{GL}(E)$ is generated by the dilatations.*

Proof. Consider any $f \in \mathbf{SL}(E)$, and let A be its matrix in any basis. By Proposition 8.24, there is a matrix S such that

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$. Since $\det(A) = 1$, we have $\alpha = 1$, and the result is proven. Otherwise, if f is invertible but $f \notin \mathbf{SL}(E)$, the above equation shows $E_{n,\alpha}$ is a dilatation, S is a product of transvections, and by Proposition 8.25, every transvection is the composition of two dilatations. Thus, the second result is also proven. $\square$

We conclude this section by proving that any two transvections are conjugate in $\mathbf{GL}(E)$. Let $\tau_{\varphi,u}$ ($u \neq 0$) be a transvection and let $g \in \mathbf{GL}(E)$ be any invertible linear map. We have

$$\begin{aligned}(g \circ \tau_{\varphi,u} \circ g^{-1})(x) &= g(g^{-1}(x) + \varphi(g^{-1}(x))u) \\ &= x + \varphi(g^{-1}(x))g(u).\end{aligned}$$

Let us find the hyperplane determined by the linear form $x \mapsto \varphi(g^{-1}(x))$. This is the set of vectors $x \in E$ such that $\varphi(g^{-1}(x)) = 0$, which holds iff $g^{-1}(x) \in H$ iff $x \in g(H)$. Therefore, $\mathrm{Ker}\,(\varphi \circ g^{-1}) = g(H) = H'$, and we have $g(u) \in g(H) = H'$, so $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $H' = g(H)$ and direction $u' = g(u)$ (with $u' \in H'$).

Conversely, let $\tau_{\psi,u'}$ be some transvection ($u' \neq 0$). Pick some vectors v, v' such that $\varphi(v) = \psi(v') = 1$, so that

$$E = H \oplus Kv = H' \oplus Kv'.$$

There is a linear map $g \in \mathbf{GL}(E)$ such that $g(u) = u'$, $g(v) = v'$, and $g(H) = H'$. To define g, pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H and pick a basis $(v', u', e'_2, \ldots, e'_{n-1})$ where $(u', e'_2, \ldots, e'_{n-1})$ is a basis of H'; then g is defined so that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = g(e'_i)$, for $i = 2, \ldots, n-1$. If $n = 2$, then e_i and e'_i are missing. Then, we have

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \varphi(g^{-1}(x))u'.$$

Now $\varphi \circ g^{-1}$ also determines the hyperplane $H' = g(H)$, so we have $\varphi \circ g^{-1} = \lambda\psi$ for some nonzero λ in K. Since $v' = g(v)$, we get

$$\varphi(v) = \varphi \circ g^{-1}(v') = \lambda\psi(v'),$$

and since $\varphi(v) = \psi(v') = 1$, we must have $\lambda = 1$. It follows that

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \psi(x)u' = \tau_{\psi,u'}(x).$$

In summary, we proved almost all parts the following result.

Proposition 8.27. *Let E be any finite-dimensional vector space. For every transvection $\tau_{\varphi,u}$ ($u \neq 0$) and every linear map $g \in \mathbf{GL}(E)$, the map $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $g(H)$ and direction $g(u)$ (that is, $g \circ \tau_{\varphi,u} \circ g^{-1} = \tau_{\varphi \circ g^{-1}, g(u)}$). For every other transvection $\tau_{\psi,u'}$ ($u' \neq 0$), there is some $g \in \mathbf{GL}(E)$ such $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$; in other words any two transvections ($\neq \mathrm{id}$) are conjugate in $\mathbf{GL}(E)$. Moreover, if $n \geq 3$, then the linear isomorphism g as above can be chosen so that $g \in \mathbf{SL}(E)$.*

Proof. We just need to prove that if $n \geq 3$, then for any two transvections $\tau_{\varphi,u}$ and $\tau_{\psi,u'}$ ($u, u' \neq 0$), there is some $g \in \mathbf{SL}(E)$ such that $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$. As before, we pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H, we pick a basis $(v', u', e_2', \ldots, e_{n-1}')$ where $(u', e_2', \ldots, e_{n-1}')$ is a basis of H', and we define g as the unique linear map such that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = e_i'$, for $i = 1, \ldots, n-1$. But in this case, both H and $H' = g(H)$ have dimension at least 2, so in any basis of H' including u', there is some basis vector e_2' independent of u', and we can rescale e_2' in such a way that the matrix of g over the two bases has determinant $+1$. $\square$

8.16 Summary

The main concepts and results of this chapter are listed below:

- One does not solve (large) linear systems by computing determinants.
- *Upper-triangular* (*lower-triangular*) matrices.
- Solving by *back-substitution* (*forward-substitution*).
- *Gaussian elimination*.
- Permuting rows.
- The *pivot* of an elimination step; *pivoting*.
- *Transposition matrix*; *elementary matrix*.
- The *Gaussian elimination theorem* (Theorem 8.1).
- *Gauss-Jordan factorization*.
- *LU-factorization*; Necessary and sufficient condition for the existence of an LU-factorization (Proposition 8.2).
- *LDU-factorization*.
- "$PA = LU$ theorem" (Theorem 8.5).
- $LDL^\top$-*factorization* of a symmetric matrix.

- Avoiding small pivots: *partial pivoting*; *complete pivoting*.
- Gaussian elimination of tridiagonal matrices.
- LU-factorization of tridiagonal matrices.
- *Symmetric positive definite* matrices (SPD matrices).
- *Cholesky factorization* (Theorem 8.10).
- Criteria for a symmetric matrix to be positive definite; *Sylvester's criterion*.
- *Reduced row echelon form*.
- Reduction of a rectangular matrix to its row echelon form.
- Using the reduction to row echelon form to decide whether a system $Ax = b$ is solvable, and to find its solutions, using a *special* solution and a basis of the *homogeneous system* $Ax = 0$.
- *Magic squares*.
- *Transvections and dilatations*.

8.17 Problems

Problem 8.1. Solve the following linear systems by Gaussian elimination:

$$\begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & -1 \\ -3 & -5 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ -7 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \\ 14 \end{pmatrix}.$$

Problem 8.2. Solve the following linear system by Gaussian elimination:

$$\begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 4 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 14 \\ -1 \\ 2 \end{pmatrix}.$$

Problem 8.3. Consider the matrix

$$A = \begin{pmatrix} 1 & c & 0 \\ 2 & 4 & 1 \\ 3 & 5 & 1 \end{pmatrix}.$$

When applying Gaussian elimination, which value of c yields zero in the second pivot position? Which value of c yields zero in the third pivot position? In this case, what can you say about the matrix A?

Problem 8.4. Solve the system

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

using the LU-factorization of Example 8.1.

Problem 8.5. Apply **rref** to the matrix

$$A_2 = \begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 3 & 0 \end{pmatrix}.$$

Problem 8.6. Apply **rref** to the matrix

$$\begin{pmatrix} 1 & 4 & 9 & 16 \\ 4 & 9 & 16 & 25 \\ 9 & 16 & 25 & 36 \\ 16 & 25 & 36 & 49 \end{pmatrix}.$$

Problem 8.7. (1) Prove that the dimension of the subspace of 2×2 matrices A, such that the sum of the entries of every row is the same (say c_1) and the sum of entries of every column is the same (say c_2) is 2.

(2) Prove that the dimension of the subspace of 2×2 matrices A, such that the sum of the entries of every row is the same (say c_1), the sum of entries of every column is the same (say c_2), and $c_1 = c_2$ is also 2. Prove that every such matrix is of the form

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

and give a basis for this subspace.

(3) Prove that the dimension of the subspace of 3×3 matrices A, such that the sum of the entries of every row is the same (say c_1), the sum of entries of every column is the same (say c_2), and $c_1 = c_2$ is 5. Begin by showing that the above constraints are given by the set of equations

$$\begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{31} \\ a_{32} \\ a_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Prove that every matrix satisfying the above constraints is of the form

$$\begin{pmatrix} a+b-c & -a+c+e & -b+c+d \\ -a-b+c+d+e & a & b \\ c & d & e \end{pmatrix},$$

with $a, b, c, d, e \in \mathbb{R}$. Find a basis for this subspace. (Use the method to find a basis for the kernel of a matrix).

Problem 8.8. If A is an $n \times n$ matrix and B is any $n \times n$ invertible matrix, prove that A is symmetric positive definite iff $B^\top AB$ is symmetric positive definite.

Problem 8.9. (1) Consider the matrix

$$A_4 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Find three matrices of the form $E_{2,1;\beta_1}, E_{3,2;\beta_2}, E_{4,3;\beta_3}$, such that

$$E_{4,3;\beta_3} E_{3,2;\beta_2} E_{2,1;\beta_1} A_4 = U_4$$

where U_4 is an upper triangular matrix. Compute

$$M = E_{4,3;\beta_3} E_{3,2;\beta_2} E_{2,1;\beta_1}$$

and check that

$$MA_4 = U_4 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & 3/2 & -1 & 0 \\ 0 & 0 & 4/3 & -1 \\ 0 & 0 & 0 & 5/4 \end{pmatrix}.$$

(2) Now consider the matrix

$$A_5 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

Find four matrices of the form $E_{2,1;\beta_1}, E_{3,2;\beta_2}, E_{4,3;\beta_3}, E_{5,4;\beta_4}$, such that

$$E_{5,4;\beta_4} E_{4,3;\beta_3} E_{3,2;\beta_2} E_{2,1;\beta_1} A_5 = U_5$$

where U_5 is an upper triangular matrix. Compute

$$M = E_{5,4;\beta_4} E_{4,3;\beta_3} E_{3,2;\beta_2} E_{2,1;\beta_1}$$

and check that

$$MA_5 = U_5 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 \\ 0 & 0 & 0 & 5/4 & -1 \\ 0 & 0 & 0 & 0 & 6/5 \end{pmatrix}.$$

(3) Write a `Matlab` program defining the function Ematrix(n, i, j, b) which is the $n \times n$ matrix that adds b times row j to row i. Also write some `Matlab` code that produces an $n \times n$ matrix A_n generalizing the matrices A_4 and A_5.

Use your program to figure out which five matrices $E_{i,j;\beta}$ reduce A_6 to the upper triangular matrix

$$U_6 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & -1 \\ 0 & 0 & 0 & 0 & 0 & 7/6 \end{pmatrix}.$$

Also use your program to figure out which six matrices $E_{i,j;\beta}$ reduce A_7 to the upper triangular matrix

$$U_7 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7/6 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8/7 \end{pmatrix}.$$

(4) Find the lower triangular matrices L_6 and L_7 such that

$$L_6 U_6 = A_6$$

and

$$L_7 U_7 = A_7.$$

(5) It is natural to conjecture that there are $n-1$ matrices of the form $E_{i,j;\beta}$ that reduce A_n to the upper triangular matrix

$$U_n = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & (n+1)/n \end{pmatrix},$$

namely,

$$E_{2,1;1/2}, E_{3,2;2/3}, E_{4,3;3/4}, \cdots, E_{n,n-1;(n-1)/n}.$$

It is also natural to conjecture that the lower triangular matrix L_n such that

$$L_n U_n = A_n$$

is given by

$$L_n = E_{2,1;-1/2} E_{3,2;-2/3} E_{4,3;-3/4} \cdots E_{n,n-1;-(n-1)/n},$$

that is,

$$L_n = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2/3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3/4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -4/5 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & -(n-1)/n & 1 \end{pmatrix}.$$

Prove the above conjectures.

(6) Prove that the last column of A_n^{-1} is

$$\begin{pmatrix} 1/(n+1) \\ 2/(n+1) \\ \vdots \\ n/(n+1) \end{pmatrix}.$$

Problem 8.10. (1) Let A be any invertible 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Prove that there is an invertible matrix S such that

$$SA = \begin{pmatrix} 1 & 0 \\ 0 & ad - bc \end{pmatrix},$$

where S is the product of at most four elementary matrices of the form $E_{i,j;\beta}$.

Conclude that every matrix A in $\mathbf{SL}(2)$ (the group of invertible 2×2 matrices A with $\det(A) = +1$) is the product of at most four elementary matrices of the form $E_{i,j;\beta}$.

For any $a \neq 0, 1$, give an explicit factorization as above for

$$A = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}.$$

What is this decomposition for $a = -1$?

(2) Recall that a rotation matrix R (a member of the group $\mathbf{SO}(2)$) is a matrix of the form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Prove that if $\theta \neq k\pi$ (with $k \in \mathbb{Z}$), any rotation matrix can be written as a product

$$R = ULU,$$

where U is upper triangular and L is lower triangular of the form

$$U = \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix}.$$

Therefore, every plane rotation (except a flip about the origin when $\theta = \pi$) can be written as the composition of three shear transformations!

Problem 8.11. (1) Recall that $E_{i,d}$ is the diagonal matrix

$$E_{i,d} = \mathrm{diag}(1, \ldots, 1, d, 1, \ldots, 1),$$

whose diagonal entries are all $+1$, except the (i,i)th entry which is equal to d.

Given any $n \times n$ matrix A, for any pair (i,j) of distinct row indices ($1 \leq i, j \leq n$), prove that there exist two elementary matrices $E_1(i,j)$ and $E_2(i,j)$ of the form $E_{k,\ell;\beta}$, such that

$$E_{j,-1}E_1(i,j)E_2(i,j)E_1(i,j)A = P(i,j)A,$$

the matrix obtained from the matrix A by permuting row i and row j. Equivalently, we have

$$E_1(i,j)E_2(i,j)E_1(i,j)A = E_{j,-1}P(i,j)A,$$

the matrix obtained from A by permuting row i and row j and multiplying row j by -1.

Prove that for every $i = 2, \ldots, n$, there exist four elementary matrices $E_3(i,d), E_4(i,d), E_5(i,d), E_6(i,d)$ of the form $E_{k,\ell;\beta}$, such that

$$E_6(i,d)E_5(i,d)E_4(i,d)E_3(i,d)E_{n,d} = E_{i,d}.$$

What happens when $d = -1$, that is, what kind of simplifications occur?

Prove that all permutation matrices can be written as products of elementary operations of the form $E_{k,\ell;\beta}$ and the operation $E_{n,-1}$.

(2) Prove that for every invertible $n \times n$ matrix A, there is a matrix S such that

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & d \end{pmatrix} = E_{n,d},$$

with $d = \det(A)$, and where S is a product of elementary matrices of the form $E_{k,\ell;\beta}$.

In particular, every matrix in $\mathbf{SL}(n)$ (the group of invertible $n \times n$ matrices A with $\det(A) = +1$) can be written as a product of elementary matrices of the form $E_{k,\ell;\beta}$. Prove that at most $n(n+1) - 2$ such transformations are needed.

(3) Prove that every matrix in $\mathbf{SL}(n)$ can be written as a product of at most $(n-1)(\max\{n, 3\} + 1)$ elementary matrices of the form $E_{k,\ell;\beta}$.

Problem 8.12. A matrix A is called *strictly column diagonally dominant* iff

$$|a_{j\,j}| > \sum_{i=1,\, i\neq j}^{n} |a_{i\,j}|, \quad \text{for } j = 1, \ldots, n.$$

Prove that if A is strictly column diagonally dominant, then Gaussian elimination with partial pivoting does not require pivoting, and A is invertible.

Problem 8.13. (1) Find a lower triangular matrix E such that

$$E \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 2 & 1 \end{pmatrix}.$$

(2) What is the effect of the product (on the left) with

$$E_{4,3;-1}E_{3,2;-1}E_{4,3;-1}E_{2,1;-1}E_{3,2;-1}E_{4,3;-1}$$

on the matrix

$$Pa_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{pmatrix}.$$

(3) Find the inverse of the matrix Pa_3.

(4) Consider the $(n+1) \times (n+1)$ Pascal matrix Pa_n whose ith row is given by the binomial coefficients

$$\binom{i-1}{j-1},$$

with $1 \leq i \leq n+1$, $1 \leq j \leq n+1$, and with the usual convention that

$$\binom{0}{0} = 1, \quad \binom{i}{j} = 0 \quad \text{if} \quad j > i.$$

The matrix Pa_3 is shown in Part (3) and Pa_4 is shown below:

$$Pa_4 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 3 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \end{pmatrix}.$$

Find n elementary matrices $E_{i_k,j_k;\beta_k}$ such that

$$E_{i_n,j_n;\beta_n} \cdots E_{i_1,j_1;\beta_1} Pa_n = \begin{pmatrix} 1 & 0 \\ 0 & Pa_{n-1} \end{pmatrix}.$$

Use the above to prove that the inverse of Pa_n is the lower triangular matrix whose ith row is given by the signed binomial coefficients

$$(-1)^{i+j-2}\binom{i-1}{j-1} = (-1)^{i+j}\binom{i-1}{j-1},$$

with $1 \leq i \leq n+1$, $1 \leq j \leq n+1$. For example,

$$Pa_4^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ -1 & 3 & -3 & 1 & 0 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix}.$$

Hint. Given any $n \times n$ matrix A, multiplying A by the elementary matrix $E_{i,j;\beta}$ *on the right* yields the matrix $AE_{i,j;\beta}$ in which β times the ith column is added to the jth column.

Problem 8.14. (1) Implement the method for converting a rectangular matrix to reduced row echelon form in `Matlab`.

(2) Use the above method to find the inverse of an invertible $n \times n$ matrix A by applying it to the the $n \times 2n$ matrix $[A\, I]$ obtained by adding the n columns of the identity matrix to A.

(3) Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 2 & 3 & 4 & 5 & \cdots & n+1 \\ 3 & 4 & 5 & 6 & \cdots & n+2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n & n+1 & n+2 & n+3 & \cdots & 2n-1 \end{pmatrix}.$$

Using your program, find the row reduced echelon form of A for $n = 4, \ldots, 20$.

Also run the `Matlab` `rref` function and compare results.

Your program probably disagrees with `rref` even for small values of n. The problem is that some pivots are very small and the normalization step (to make the pivot 1) causes roundoff errors. Use a tolerance parameter to fix this problem.

What can you conjecture about the rank of A?

(4) Prove that the matrix A has the following row reduced form:

$$R = \begin{pmatrix} 1 & 0 & -1 & -2 & \cdots & -(n-2) \\ 0 & 1 & 2 & 3 & \cdots & n-1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Deduce from the above that A has rank 2.

Hint. Some well chosen sequence of row operations.

(5) Use your program to show that if you add any number greater than or equal to $(2/25)n^2$ to every diagonal entry of A you get an invertible matrix! In fact, running the `Matlab` function `chol` should tell you that these matrices are SPD (symmetric, positive definite).

Problem 8.15. Let A be an $n \times n$ complex Hermitian positive definite matrix. Prove that the lower-triangular matrix B with positive diagonal entries such that $A = BB^*$ is given by the following formulae: For $j = 1, \ldots, n$,

$$b_{j\,j} = \left(a_{j\,j} - \sum_{k=1}^{j-1} |b_{j\,k}|^2 \right)^{1/2},$$

and for $i = j+1, \ldots, n$ (and $j = 1, \ldots, n-1$)

$$\overline{b}_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k}\overline{b}_{j\,k} \right) / b_{j\,j}.$$

Problem 8.16. (Permutations and permutation matrices) A permutation can be viewed as an operation permuting the rows of a matrix. For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

corresponds to the matrix

$$P_\pi = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Observe that the matrix P_π has a single 1 on every row and every column, all other entries being zero, and that if we multiply any 4×4 matrix A by P_π on the left, then the rows of A are permuted according to the permutation π; that is, the $\pi(i)$th row of $P_\pi A$ is the ith row of A. For example,

$$P_\pi A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix}.$$

Equivalently, the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. In order for the matrix P_π to move the ith row of A to the $\pi(i)$th row, the $\pi(i)$th row of P_π must have a 1 in column i and zeros everywhere else; this means that the ith column of P_π contains the basis vector $e_{\pi(i)}$, the vector that has a 1 in position $\pi(i)$ and zeros everywhere else.

This is the general situation and it leads to the following definition.

Definition 8.10. Given any permutation $\pi\colon [n] \to [n]$, the *permutation matrix* $P_\pi = (p_{ij})$ representing π is the matrix given by

$$p_{ij} = \begin{cases} 1 & \text{if } i = \pi(j) \\ 0 & \text{if } i \neq \pi(j); \end{cases}$$

equivalently, the jth column of P_π is the basis vector $e_{\pi(j)}$. A *permutation matrix* P is any matrix of the form P_π (where P is an $n \times n$ matrix, and $\pi\colon [n] \to [n]$ is a permutation, for some $n \geq 1$).

Remark: There is a confusing point about the notation for permutation matrices. A permutation matrix P acts on a matrix A by multiplication on the left by permuting the rows of A. As we said before, this means that the $\pi(i)$th row of $P_\pi A$ is the ith row of A, or equivalently that the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. But then observe that the row index of the entries of the ith row of PA is $\pi^{-1}(i)$, and not $\pi(i)$! See the following example:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix},$$

where

$$\begin{aligned} \pi^{-1}(1) &= 4 \\ \pi^{-1}(2) &= 3 \\ \pi^{-1}(3) &= 1 \\ \pi^{-1}(4) &= 2. \end{aligned}$$

Prove the following results

(1) Given any two permutations $\pi_1, \pi_2 \colon [n] \to [n]$, the permutation matrix $P_{\pi_2 \circ \pi_1}$ representing the composition of π_1 and π_2 is equal to the product $P_{\pi_2} P_{\pi_1}$ of the permutation matrices P_{π_1} and P_{π_2} representing π_1 and π_2; that is,

$$P_{\pi_2 \circ \pi_1} = P_{\pi_2} P_{\pi_1}.$$

(2) The matrix $P_{\pi_1^{-1}}$ representing the inverse of the permutation π_1 is the inverse $P_{\pi_1}^{-1}$ of the matrix P_{π_1} representing the permutation π_1; that is,

$$P_{\pi_1^{-1}} = P_{\pi_1}^{-1}.$$

Furthermore,

$$P_{\pi_1}^{-1} = (P_{\pi_1})^\top.$$

(3) Prove that if P is the matrix associated with a transposition, then $\det(P) = -1$.

(4) Prove that if P is a permutation matrix, then $\det(P) = \pm 1$.

(5) Use permutation matrices to give another proof of the fact that the parity of the number of transpositions used to express a permutation π depends only on π.

Chapter 9

Vector Norms and Matrix Norms

9.1 Normed Vector Spaces

In order to define how close two vectors or two matrices are, and in order to define the convergence of sequences of vectors or matrices, we can use the notion of a norm. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$. Also recall that if $z = a + ib \in \mathbb{C}$ is a complex number, with $a, b \in \mathbb{R}$, then $\overline{z} = a - ib$ and $|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}$ ($|z|$ is the *modulus* of z).

Definition 9.1. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm* on E is a function $\| \ \| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y \in E$ and $\lambda \in K$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (homogeneity (or scaling))

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \ \|$ is called a *normed vector space*.

By (N2), setting $\lambda = -1$, we obtain

$$\|-x\| = \|(-1)x\| = |-1| \, \|x\| = \|x\| \, ;$$

that is, $\|-x\| = \|x\|$. From (N3), we have

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\| \, ,$$

which implies that

$$\|x\| - \|y\| \leq \|x - y\| \, .$$

By exchanging x and y and using the fact that by (N2),

$$\|y - x\| = \|-(x - y)\| = \|x - y\| \, ,$$

we also have

$$\|y\| - \|x\| \leq \|x - y\|.$$

Therefore,

$$|\|x\| - \|y\|| \leq \|x - y\|, \quad \text{for all } x, y \in E. \tag{$*$}$$

Observe that setting $\lambda = 0$ in (N2), we deduce that $\|0\| = 0$ without assuming (N1). Then by setting $y = 0$ in $(*)$, we obtain

$$|\|x\|| \leq \|x\|, \quad \text{for all } x \in E.$$

Therefore, the condition $\|x\| \geq 0$ in (N1) follows from (N2) and (N3), and (N1) can be replaced by the weaker condition

(N1') For all $x \in E$, if $\|x\| = 0$, then $x = 0$,

A function $\| \, \| : E \to \mathbb{R}$ satisfying Axioms (N2) and (N3) is called a *seminorm*. From the above discussion, a seminorm also has the properties

$\|x\| \geq 0$ for all $x \in E$, and $\|0\| = 0$.

However, there may be nonzero vectors $x \in E$ such that $\|x\| = 0$.

Let us give some examples of normed vector spaces.

Example 9.1.

1. Let $E = \mathbb{R}$, and $\|x\| = |x|$, the absolute value of x.

2. Let $E = \mathbb{C}$, and $\|z\| = |z|$, the modulus of z.

3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the ℓ^p-*norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

See Figures 9.1 through 9.4.

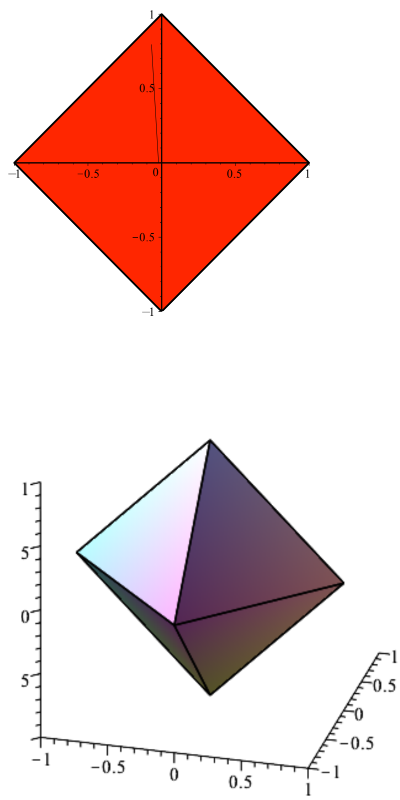


Figure 9.1: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_1 \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_1 \leq 1\}$.

There are other norms besides the ℓ^p-norms. Here are some examples.

1. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = |u_1| + 2|u_2|.$$
See Figure 9.5.

2. For $E = \mathbb{R}^2$,
$$\|(u_1, u_2)\| = \left((u_1 + u_2)^2 + u_1^2\right)^{1/2}.$$
See Figure 9.6.

3. For $E = \mathbb{C}^2$,
$$\|(u_1, u_2)\| = |u_1 + iu_2| + |u_1 - iu_2|.$$

The reader should check that they satisfy all the axioms of a norm.

Some work is required to show the triangle inequality for the ℓ^p-norm.

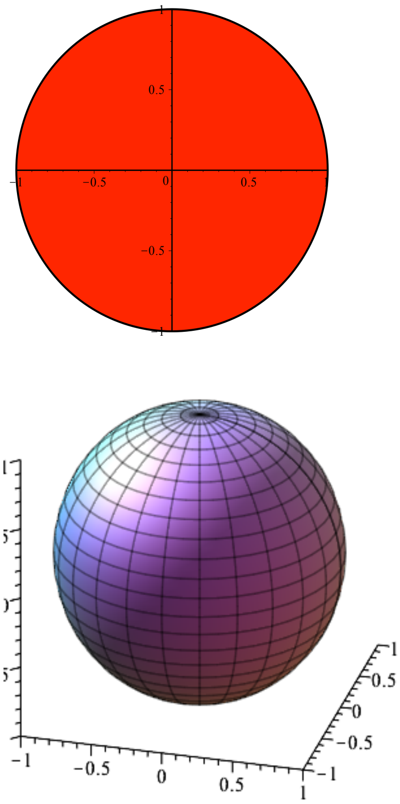


Figure 9.2: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_2 \leq 1\}$.

Proposition 9.1. *If $E = \mathbb{C}^n$ or $E = \mathbb{R}^n$, for every real number $p \geq 1$, the ℓ^p-norm is indeed a norm.*

Proof. The cases $p = 1$ and $p = \infty$ are easy and left to the reader. If $p > 1$, then let $q > 1$ such that

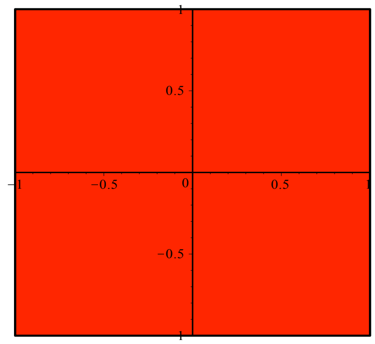
$$\frac{1}{p} + \frac{1}{q} = 1.$$

We will make use of the following fact: for all $\alpha, \beta \in \mathbb{R}$, if $\alpha, \beta \geq 0$, then

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}. \tag{$*$}$$

To prove the above inequality, we use the fact that the exponential function $t \mapsto e^t$ satisfies the following convexity inequality:

$$e^{\theta x + (1-\theta)y} \leq \theta e^x + (1-\theta)e^y,$$



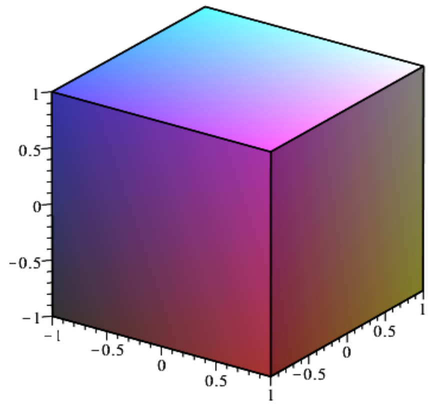


Figure 9.3: The top figure is $\{x \in \mathbb{R}^2 \mid \|x\|_\infty \leq 1\}$, while the bottom figure is $\{x \in \mathbb{R}^3 \mid \|x\|_\infty \leq 1\}$.

for all $x, y \in \mathbb{R}$ and all θ with $0 \leq \theta \leq 1$.

Since the case $\alpha\beta = 0$ is trivial, let us assume that $\alpha > 0$ and $\beta > 0$. If we replace θ by $1/p$, x by $p \log \alpha$ and y by $q \log \beta$, then we get

$$e^{\frac{1}{p} p \log \alpha + \frac{1}{q} q \log \beta} \leq \frac{1}{p} e^{p \log \alpha} + \frac{1}{q} e^{q \log \beta},$$

which simplifies to

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q},$$

as claimed.

We will now prove that for any two vectors $u, v \in E$, (where E is of dimension n), we have

$$\sum_{i=1}^{n} |u_i v_i| \leq \|u\|_p \, \|v\|_q \,. \tag{$**$}$$

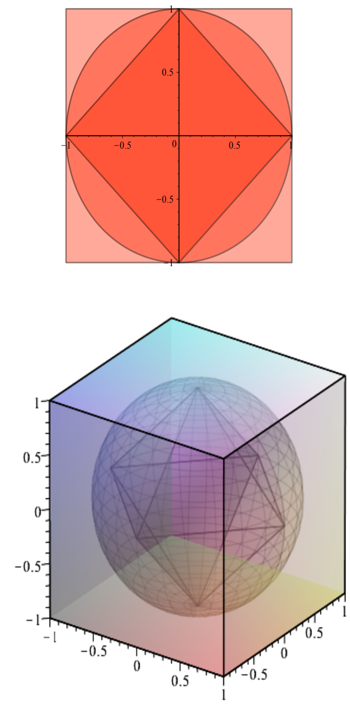


Figure 9.4: The relationships between the closed unit balls from the ℓ^1-norm, the Euclidean norm, and the sup-norm.

Since the above is trivial if $u = 0$ or $v = 0$, let us assume that $u \neq 0$ and $v \neq 0$. Then Inequality $(*)$ with $\alpha = |u_i| / \left\|u\right\|_p$ and $\beta = |v_i| / \left\|v\right\|_q$ yields

$$\frac{|u_i v_i|}{\left\|u\right\|_p \left\|v\right\|_q} \leq \frac{|u_i|^p}{p \left\|u\right\|_p^p} + \frac{|v_i|^q}{q \left\|v\right\|_q^q},$$

for $i = 1, \ldots, n$, and by summing up these inequalities, we get

$$\sum_{i=1}^{n} |u_i v_i| \leq \left\|u\right\|_p \left\|v\right\|_q,$$

as claimed. To finish the proof, we simply have to prove that property (N3) holds, since (N1) and (N2) are clear. For $i = 1, \ldots, n$, we can write

$$(|u_i| + |v_i|)^p = |u_i|(|u_i| + |v_i|)^{p-1} + |v_i|(|u_i| + |v_i|)^{p-1},$$

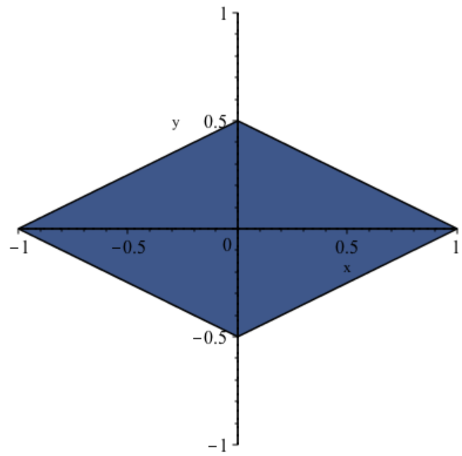


Figure 9.5: The unit closed unit ball $\{(u_1, u_2) \in \mathbb{R}^2 \mid \|(u_1, u_2)\| \leq 1\}$, where $\|(u_1, u_2)\| = |u_1| + 2|u_2|$.

so that by summing up these equations we get

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p = \sum_{i=1}^{n} |u_i|(|u_i| + |v_i|)^{p-1} + \sum_{i=1}^{n} |v_i|(|u_i| + |v_i|)^{p-1},$$

and using Inequality $(**)$, with $V \in E$ where $V_i = (|u_i| + |v_i|)^{p-1}$, we get

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p \leq \|u\|_p \|V\|_q + \|v\|_p \|V\|_q = (\|u\|_p + \|v\|_p)\left(\sum_{i=1}^{n}(|u_i| + |v_i|)^{(p-1)q}\right)^{1/q}.$$

However, $1/p + 1/q = 1$ implies $pq = p + q$, that is, $(p-1)q = p$, so we have

$$\sum_{i=1}^{n}(|u_i| + |v_i|)^p \leq (\|u\|_p + \|v\|_p)\left(\sum_{i=1}^{n}(|u_i| + |v_i|)^p\right)^{1/q},$$

which yields

$$\left(\sum_{i=1}^{n}(|u_i| + |v_i|)^p\right)^{1-1/q} = \left(\sum_{i=1}^{n}(|u_i| + |v_i|)^p\right)^{1/p} \leq \|u\|_p + \|v\|_p.$$

Since $|u_i + v_i| \leq |u_i| + |v_i|$, the above implies the triangle inequality $\|u + v\|_p \leq \|u\|_p + \|v\|_p$, as claimed. $\square$

For $p > 1$ and $1/p + 1/q = 1$, the inequality

$$\sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} |u_i|^p\right)^{1/p}\left(\sum_{i=1}^{n} |v_i|^q\right)^{1/q}$$

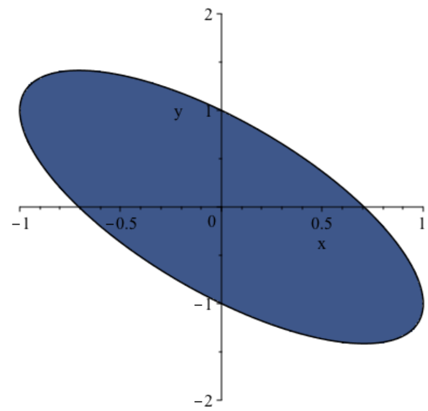


Figure 9.6: The unit closed unit ball $\{(u_1, u_2) \in \mathbb{R}^2 \mid \|(u_1, u_2)\| \leq 1\}$, where $\|(u_1, u_2)\| = \big((u_1 + u_2)^2 + u_1^2\big)^{1/2}$.

is known as *Hölder's inequality*. For $p = 2$, it is the *Cauchy–Schwarz inequality*.

Actually, if we define the *Hermitian inner product* $\langle -, - \rangle$ on $\mathbb{C}^n$ by

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i \overline{v}_i,$$

where $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$, then

$$|\langle u, v \rangle| \leq \sum_{i=1}^{n} |u_i \overline{v}_i| = \sum_{i=1}^{n} |u_i v_i|,$$

so Hölder's inequality implies the following inequalities.

Corollary 9.2. *(Hölder's inequalities) For any real numbers p, q, such that $p, q \geq 1$ and*

$$\frac{1}{p} + \frac{1}{q} = 1,$$

(with $q = +\infty$ if $p = 1$ and $p = +\infty$ if $q = 1$), we have the inequalities

$$\sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} |u_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |v_i|^q\right)^{1/q}$$

and

$$|\langle u, v \rangle| \leq \|u\|_p \, \|v\|_q \,, \qquad u, v \in \mathbb{C}^n.$$

For $p = 2$, this is the standard Cauchy–Schwarz inequality. The triangle inequality for the ℓ^p-norm,

$$\left(\sum_{i=1}^{n}(|u_i + v_i|)^p\right)^{1/p} \leq \left(\sum_{i=1}^{n}|u_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n}|v_i|^p\right)^{1/p},$$

is known as *Minkowski's inequality*.

When we restrict the Hermitian inner product to real vectors, $u, v \in \mathbb{R}^n$, we get the *Euclidean inner product*

$$\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i.$$

It is very useful to observe that if we represent (as usual) $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ (in $\mathbb{R}^n$) by column vectors, then their Euclidean inner product is given by

$$\langle u, v\rangle = u^\top v = v^\top u,$$

and when $u, v \in \mathbb{C}^n$, their Hermitian inner product is given by

$$\langle u, v\rangle = v^* u = \overline{u^* v}.$$

In particular, when $u = v$, in the complex case we get

$$\|u\|_2^2 = u^* u,$$

and in the real case this becomes

$$\|u\|_2^2 = u^\top u.$$

As convenient as these notations are, we still recommend that you do not abuse them; the notation $\langle u, v\rangle$ is more intrinsic and still "works" when our vector space is infinite dimensional.

Remark: If $0 < p < 1$, then $x \mapsto \|x\|_p$ is not a norm because the triangle inequality *fails*. For example, consider $x = (2, 0)$ and $y = (0, 2)$. Then $x + y = (2, 2)$, and we have $\|x\|_p = (2^p + 0^p)^{1/p} = 2$, $\|y\|_p = (0^p + 2^p)^{1/p} = 2$, and $\|x + y\|_p = (2^p + 2^p)^{1/p} = 2^{(p+1)/p}$. Thus

$$\|x + y\|_p = 2^{(p+1)/p}, \quad \|x\|_p + \|y\|_p = 4 = 2^2.$$

Since $0 < p < 1$, we have $2p < p + 1$, that is, $(p + 1)/p > 2$, so $2^{(p+1)/p} > 2^2 = 4$, and the triangle inequality $\|x + y\|_p \leq \|x\|_p + \|y\|_p$ fails.

Observe that

$$\|(1/2)x\|_p = (1/2)\,\|x\|_p = \|(1/2)y\|_p = (1/2)\,\|y\|_p = 1, \quad \|(1/2)(x + y)\|_p = 2^{1/p},$$

and since $p < 1$, we have $2^{1/p} > 2$, so

$$\|(1/2)(x + y)\|_p = 2^{1/p} > 2 = (1/2)\,\|x\|_p + (1/2)\,\|y\|_p,$$

and the map $x \mapsto \|x\|_p$ is not convex.

For $p = 0$, for any $x \in \mathbb{R}^n$, we have

$$\|x\|_0 = |\{i \in \{1, \ldots, n\} \mid x_i \neq 0\}|,$$

the number of nonzero components of x. The map $x \mapsto \|x\|_0$ is not a norm this time because Axiom (N2) fails. For example,

$$\|(1,0)\|_0 = \|(10,0)\|_0 = 1 \neq 10 = 10\,\|(1,0)\|_0\,.$$

The map $x \mapsto \|x\|_0$ is also not convex. For example,

$$\|(1/2)(2,2)\|_0 = \|(1,1)\|_0 = 2,$$

and

$$\|(2,0)\|_0 = \|(0,2)\|_0 = 1,$$

but

$$\|(1/2)(2,2)\|_0 = 2 > 1 = (1/2)\,\|(2,0)\|_0 + (1/2)\,\|(0,2)\|_0\,.$$

Nevertheless, the "zero-norm" $x \mapsto \|x\|_0$ is used in machine learning as a regularizing term which encourages sparsity, namely increases the number of zero components of the vector x.

The following proposition is easy to show.

Proposition 9.3. *The following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):*

$$\begin{aligned}
\|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty,\\
\|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty,\\
\|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\|x\|_2.
\end{aligned}$$

Proposition 9.3 is actually a special case of a very important result: *in a finite-dimensional vector space, any two norms are equivalent.*

Definition 9.2. Given any (real or complex) vector space E, two norms $\| \|_a$ and $\| \|_b$ are *equivalent* iff there exists some positive reals $C_1, C_2 > 0$, such that

$$\|u\|_a \leq C_1\,\|u\|_b \quad \text{and} \quad \|u\|_b \leq C_2\,\|u\|_a\,, \text{ for all } u \in E.$$

There is an illuminating interpretation of Definition 9.2 in terms of open balls. For any radius $\rho > 0$ and any $x \in E$, consider the open a-ball of center x and radius ρ (with respect the norm $\| \|_a$),

$$B_a(x, \rho) = \{z \in E \mid \|z - x\|_a < \rho\}.$$

We claim that there is some open b-ball $B_b(x, r)$ of radius $r > 0$ and center x,

$$B_b(x, r) = \{z \in E \mid \|z - x\|_n < r\},$$

such that

$$B_b(x, r) \subseteq B_a(x, \rho).$$

Indeed, if we pick $r = \rho/C_1$, for any $z \in E$, if $\|z - x\|_b < \rho/C_1$, then

$$\|z - x\|_a \leq C_1 \|z - x\|_b < C_1(\rho/C_1) = \rho,$$

which means that

$$B_b(x, \rho/C_1) \subseteq B_a(x, \rho).$$

Similarly, for any radius $\rho > 0$ and any $x \in E$, we have

$$B_a(x, \rho/C_2) \subseteq B_b(x, \rho).$$

Now given a normed vector space $(E, \| \|)$, a subset U of E is said to be *open* (with respect to the norm $\| \|$) if either $U = \emptyset$ or if for every $x \in U$, there is some open ball $B(x, \rho)$ (for some $\rho > 0$) such that $B(x, \rho) \subseteq U$.

The collection $\mathcal{U}$ of open sets defined by the norm $\| \|$ is called the *topology on E induced by the norm $\| \|$*. What we showed above regarding the containments of open a-balls and open b-balls immediately implies that *two equivalent norms induce the same topology on E*. This is the reason why the notion of equivalent norms is important.

Given any norm $\| \|$ on a vector space of dimension n, for any basis $(e_1, \ldots, e_n)$ of E, observe that for any vector $x = x_1e_1 + \cdots + x_ne_n$, we have

$$\|x\| = \|x_1e_1 + \cdots + x_ne_n\| \leq |x_1| \|e_1\| + \cdots + |x_n| \|e_n\| \leq C(|x_1| + \cdots + |x_n|) = C \|x\|_1,$$

with $C = \max_{1 \leq i \leq n} \|e_i\|$ and with the norm $\|x\|_1$ defined as

$$\|x\|_1 = \|x_1e_1 + \cdots + x_ne_n\| = |x_1| + \cdots + |x_n|.$$

The above implies that

$$| \, \|u\| - \|v\| \, | \leq \|u - v\| \leq C \|u - v\|_1,$$

and this implies the following corollary.

Corollary 9.4. *For any norm $u \mapsto \|u\|$ on a finite-dimensional (complex or real) vector space E, the map $u \mapsto \|u\|$ is continuous with respect to the norm $\| \|_1$.*

Let S_1^{n-1} be the unit sphere with respect to the norm $\| \|_1$, namely

$$S_1^{n-1} = \{x \in E \mid \|x\|_1 = 1\}.$$

Now S_1^{n-1} is a closed and bounded subset of a finite-dimensional vector space, so by Heine–Borel (or equivalently, by Bolzano–Weiertrass), S_1^{n-1} is compact. On the other hand, it is a well known result of analysis that any continuous real-valued function on a nonempty compact set has a minimum and a maximum, and that they are achieved. Using these facts, we can prove the following important theorem:

Theorem 9.5. *If E is any real or complex vector space of finite dimension, then any two norms on E are equivalent.*

Proof. It is enough to prove that any norm $\| \|$ is equivalent to the 1-norm. We already proved that the function $x \mapsto \|x\|$ is continuous with respect to the norm $\| \|_1$, and we observed that the unit sphere S_1^{n-1} is compact. Now we just recalled that because the function $f\colon x \mapsto \|x\|$ is continuous and because S_1^{n-1} is compact, the function f has a minimum m and a maximum M, and because $\|x\|$ is never zero on S_1^{n-1}, we must have $m > 0$. Consequently, we just proved that if $\|x\|_1 = 1$, then

$$0 < m \leq \|x\| \leq M,$$

so for any $x \in E$ with $x \neq 0$, we get

$$m \leq \|x/\|x\|_1\| \leq M,$$

which implies

$$m\,\|x\|_1 \leq \|x\| \leq M\,\|x\|_1\,.$$

Since the above inequality holds trivially if $x = 0$, we just proved that $\| \|$ and $\| \|_1$ are equivalent, as claimed. ☐

Remark: Let P be a $n \times n$ symmetric positive definite matrix. It is immediately verified that the map $x \mapsto \|x\|_P$ given by

$$\|x\|_P = (x^\top P x)^{1/2}$$

is a norm on $\mathbb{R}^n$ called a *quadratic norm*. Using some convex analysis (the Löwner–John ellipsoid), it can be shown that *any* norm $\| \|$ on $\mathbb{R}^n$ can be approximated by a quadratic norm in the sense that there is a quadratic norm $\| \|_P$ such that

$$\|x\|_P \leq \|x\| \leq \sqrt{n}\,\|x\|_P \qquad \text{for all } x \in \mathbb{R}^n;$$

see Boyd and Vandenberghe [29], Section 8.4.1.

Next we will consider norms on matrices.

9.2 Matrix Norms

For simplicity of exposition, we will consider the vector spaces $\mathrm{M}_n(\mathbb{R})$ and $\mathrm{M}_n(\mathbb{C})$ of square $n \times n$ matrices. Most results also hold for the spaces $\mathrm{M}_{m,n}(\mathbb{R})$ and $\mathrm{M}_{m,n}(\mathbb{C})$ of rectangular $m \times n$ matrices. Since $n \times n$ matrices can be multiplied, the idea behind matrix norms is that they should behave "well" with respect to matrix multiplication.

Definition 9.3. A *matrix norm* $\| \, \|$ on the space of square $n \times n$ matrices in $\mathrm{M}_n(K)$, with $K = \mathbb{R}$ or $K = \mathbb{C}$, is a norm on the vector space $\mathrm{M}_n(K)$, with the additional property called *submultiplicativity* that

$$\|AB\| \leq \|A\| \, \|B\| ,$$

for all $A, B \in \mathrm{M}_n(K)$. A norm on matrices satisfying the above property is often called a *submultiplicative* matrix norm.

Since $I^2 = I$, from $\|I\| = \|I^2\| \leq \|I\|^2$, we get $\|I\| \geq 1$, for every matrix norm.

Before giving examples of matrix norms, we need to review some basic definitions about matrices. Given any matrix $A = (a_{ij}) \in \mathrm{M}_{m,n}(\mathbb{C})$, the *conjugate* $\overline{A}$ of A is the matrix such that

$$\overline{A}_{ij} = \overline{a_{ij}}, \quad 1 \leq i \leq m, \; 1 \leq j \leq n.$$

The *transpose* of A is the $n \times m$ matrix $A^\top$ such that

$$A^\top_{ij} = a_{ji}, \quad 1 \leq i \leq m, \; 1 \leq j \leq n.$$

The *adjoint* of A is the $n \times m$ matrix A^* such that

$$A^* = \overline{(A^\top)} = (\overline{A})^\top.$$

When A is a real matrix, $A^* = A^\top$. A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *Hermitian* if

$$A^* = A.$$

If A is a real matrix ($A \in \mathrm{M}_n(\mathbb{R})$), we say that A is *symmetric* if

$$A^\top = A.$$

A matrix $A \in \mathrm{M}_n(\mathbb{C})$ is *normal* if

$$AA^* = A^*A,$$

and if A is a real matrix, it is *normal* if

$$AA^\top = A^\top A.$$

A matrix $U \in \mathrm{M}_n(\mathbb{C})$ is *unitary* if

$$UU^* = U^*U = I.$$

A real matrix $Q \in \mathrm{M}_n(\mathbb{R})$ is *orthogonal* if

$$QQ^\top = Q^\top Q = I.$$

Given any matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, the *trace* $\mathrm{tr}(A)$ of A is the sum of its diagonal elements

$$\mathrm{tr}(A) = a_{11} + \cdots + a_{nn}.$$

It is easy to show that the trace is a linear map, so that

$$\mathrm{tr}(\lambda A) = \lambda \mathrm{tr}(A)$$

and

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$

Moreover, if A is an $m \times n$ matrix and B is an $n \times m$ matrix, it is not hard to show that

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

We also review eigenvalues and eigenvectors. We content ourselves with definition involving matrices. A more general treatment will be given later on (see Chapter 15).

Definition 9.4. Given any square matrix $A \in \mathrm{M}_n(\mathbb{C})$, a complex number $\lambda \in \mathbb{C}$ is an *eigenvalue* of A if there is some *nonzero* vector $u \in \mathbb{C}^n$, such that

$$Au = \lambda u.$$

If λ is an eigenvalue of A, then the *nonzero* vectors $u \in \mathbb{C}^n$ such that $Au = \lambda u$ are called *eigenvectors of A associated with λ*; together with the zero vector, these eigenvectors form a subspace of $\mathbb{C}^n$ denoted by $E_\lambda(A)$, and called the *eigenspace associated with λ*.

Remark: Note that Definition 9.4 *requires an eigenvector to be nonzero*. A somewhat unfortunate consequence of this requirement is that the set of eigenvectors is *not* a subspace, since the zero vector is missing! On the positive side, whenever eigenvectors are involved, there is no need to say that they are nonzero. In contrast, even if we allow 0 to be an eigenvector, in order for a scalar λ to be an eigenvalue, there must be a *nonzero vector u* such that $Au = \lambda u$. Without this restriction, since $A0 = \lambda 0 = 0$ for all λ, every scalar would be an eigenvector, which would make the definition of an eigenvalue trivial and useless. The fact that eigenvectors are nonzero is implicitly used in all the arguments involving them, so it seems preferable (but perhaps not as elegant) to stipulate that eigenvectors should be nonzero.

If A is a square real matrix $A \in \mathrm{M}_n(\mathbb{R})$, then we restrict Definition 9.4 to real eigenvalues $\lambda \in \mathbb{R}$ and real eigenvectors. However, it should be noted that although every complex matrix always has at least some complex eigenvalue, a real matrix may not have any real eigenvalues. For example, the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has the complex eigenvalues i and $-i$, but no real eigenvalues. Thus, typically even for real matrices, we consider complex eigenvalues.

Observe that $\lambda \in \mathbb{C}$ is an eigenvalue of A

- iff $Au = \lambda u$ for some nonzero vector $u \in \mathbb{C}^n$
- iff $(\lambda I - A)u = 0$
- iff the matrix $\lambda I - A$ defines a linear map which has a nonzero kernel, that is,
- iff $\lambda I - A$ not invertible.

However, from Proposition 7.10, $\lambda I - A$ is not invertible iff

$$\det(\lambda I - A) = 0.$$

Now $\det(\lambda I - A)$ is a polynomial of degree n in the indeterminate λ, in fact, of the form

$$\lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A).$$

Thus we see that the eigenvalues of A are the zeros (also called *roots*) of the above polynomial. Since every complex polynomial of degree n has exactly n roots, counted with their multiplicity, we have the following definition:

Definition 9.5. Given any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, the polynomial

$$\det(\lambda I - A) = \lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A)$$

is called the *characteristic polynomial* of A. The n (not necessarily distinct) roots $\lambda_1, \ldots, \lambda_n$ of the characteristic polynomial are all the *eigenvalues* of A and constitute the *spectrum* of A. We let

$$\rho(A) = \max_{1 \le i \le n} |\lambda_i|$$

be the largest modulus of the eigenvalues of A, called the *spectral radius* of A.

Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are the zeros of the polynomial

$$\det(\lambda I - A) = \lambda^n - \mathrm{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A),$$

we deduce (see Section 15.1 for details) that

$$\begin{aligned} \mathrm{tr}(A) &= \lambda_1 + \cdots + \lambda_n \\ \det(A) &= \lambda_1 \cdots \lambda_n. \end{aligned}$$

Proposition 9.6. *For any matrix norm $\| \ \|$ on $\mathrm{M}_n(\mathbb{C})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\rho(A) \leq \|A\| .$$

Proof. Let λ be some eigenvalue of A for which $|\lambda|$ is maximum, that is, such that $|\lambda| = \rho(A)$. If u ($\neq 0$) is any eigenvector associated with λ and if U is the $n \times n$ matrix whose columns are all u, then $Au = \lambda u$ implies

$$AU = \lambda U,$$

and since

$$|\lambda| \, \|U\| = \|\lambda U\| = \|AU\| \leq \|A\| \, \|U\|$$

and $U \neq 0$, we have $\|U\| \neq 0$, and get

$$\rho(A) = |\lambda| \leq \|A\| ,$$

as claimed. □

Proposition 9.6 also holds for any real matrix norm $\| \ \|$ on $\mathrm{M}_n(\mathbb{R})$ but the proof is more subtle and requires the notion of induced norm. We prove it after giving Definition 9.7.

It turns out that if A is a real $n \times n$ symmetric matrix, then the eigenvalues of A are all real and there is some orthogonal matrix Q such that

$$A = Q \mathrm{diag}(\lambda_1, \ldots, \lambda_n) Q^\top,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A. Similarly, if A is a complex $n \times n$ Hermitian matrix, then the eigenvalues of A are all real and there is some unitary matrix U such that

$$A = U \mathrm{diag}(\lambda_1, \ldots, \lambda_n) U^*,$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of A. See Chapter 17 for the proof of these results.

We now return to matrix norms. We begin with the so-called *Frobenius norm*, which is just the norm $\| \ \|_2$ on $\mathbb{C}^{n^2}$, where the $n \times n$ matrix A is viewed as the vector obtained by concatenating together the rows (or the columns) of A. The reader should check that for any $n \times n$ complex matrix $A = (a_{ij})$,

$$\left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\mathrm{tr}(AA^*)}.$$

Definition 9.6. The *Frobenius norm* $\| \ \|_F$ is defined so that for every square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$,

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}(AA^*)} = \sqrt{\mathrm{tr}(A^*A)}.$$

The following proposition show that the Frobenius norm is a matrix norm satisfying other nice properties.

Proposition 9.7. *The Frobenius norm* $\| \ \|_F$ *on* $\mathrm{M}_n(\mathbb{C})$ *satisfies the following properties:*

(1) It is a matrix norm; that is, $\|AB\|_F \leq \|A\|_F \|B\|_F$*, for all* $A, B \in \mathrm{M}_n(\mathbb{C})$*.*

(2) It is unitarily invariant, which means that for all unitary matrices U, V*, we have*

$$\|A\|_F = \|UA\|_F = \|AV\|_F = \|UAV\|_F .$$

(3) $\sqrt{\rho(A^*A)} \leq \|A\|_F \leq \sqrt{n}\sqrt{\rho(A^*A)}$*, for all* $A \in \mathrm{M}_n(\mathbb{C})$*.*

Proof. (1) The only property that requires a proof is the fact $\|AB\|_F \leq \|A\|_F \|B\|_F$. This follows from the Cauchy–Schwarz inequality:

$$\begin{aligned} \|AB\|_F^2 &= \sum_{i,j=1}^n \left| \sum_{k=1}^n a_{ik} b_{kj} \right|^2 \\ &\leq \sum_{i,j=1}^n \left(\sum_{h=1}^n |a_{ih}|^2 \right) \left(\sum_{k=1}^n |b_{kj}|^2 \right) \\ &= \left(\sum_{i,h=1}^n |a_{ih}|^2 \right) \left(\sum_{k,j=1}^n |b_{kj}|^2 \right) = \|A\|_F^2 \|B\|_F^2 . \end{aligned}$$

(2) We have

$$\|A\|_F^2 = \mathrm{tr}(AA^*) = \mathrm{tr}(AVV^*A^*) = \mathrm{tr}(AV(AV)^*) = \|AV\|_F^2 ,$$

and

$$\|A\|_F^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(A^*U^*UA) = \|UA\|_F^2 .$$

The identity

$$\|A\|_F = \|UAV\|_F$$

follows from the previous two.

(3) It is shown in Section 15.1 that the trace of a matrix is equal to the sum of its eigenvalues. Furthermore, A^*A is symmetric positive semidefinite (which means that its eigenvalues are nonnegative), so $\rho(A^*A)$ is the largest eigenvalue of A^*A and

$$\rho(A^*A) \leq \mathrm{tr}(A^*A) \leq n\rho(A^*A),$$

which yields (3) by taking square roots. ∎

Remark: The Frobenius norm is also known as the *Hilbert-Schmidt norm* or the *Schur norm*. So many famous names associated with such a simple thing!

9.3 Subordinate Norms

We now give another method for obtaining matrix norms using subordinate norms. First we need a proposition that shows that in a finite-dimensional space, the linear map induced by a matrix is bounded, and thus continuous.

Proposition 9.8. *For every norm $\| \|$ on $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$), there is a real constant $C_A \geq 0$, such that*

$$\|Au\| \leq C_A \|u\|,$$

for every vector $u \in \mathbb{C}^n$ (or $u \in \mathbb{R}^n$ if A is real).

Proof. For every basis $(e_1, \ldots, e_n)$ of $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every vector $u = u_1e_1 + \cdots + u_ne_n$, we have

$$\begin{aligned} \|Au\| &= \|u_1A(e_1) + \cdots + u_nA(e_n)\| \\ &\leq |u_1| \|A(e_1)\| + \cdots + |u_n| \|A(e_n)\| \\ &\leq C_1(|u_1| + \cdots + |u_n|) = C_1 \|u\|_1, \end{aligned}$$

where $C_1 = \max_{1 \leq i \leq n} \|A(e_i)\|$. By Theorem 9.5, the norms $\| \|$ and $\| \|_1$ are equivalent, so there is some constant $C_2 > 0$ so that $\|u\|_1 \leq C_2 \|u\|$ for all u, which implies that

$$\|Au\| \leq C_A \|u\|,$$

where $C_A = C_1C_2$. ∎

Proposition 9.8 says that every linear map on a finite-dimensional space is *bounded*. This implies that every linear map on a finite-dimensional space is continuous. Actually, it is not hard to show that a linear map on a normed vector space E is bounded iff it is continuous, regardless of the dimension of E.

Proposition 9.8 implies that for every matrix $A \in \mathrm{M}_n(\mathbb{C})$ (or $A \in \mathrm{M}_n(\mathbb{R})$),

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} \leq C_A.$$

Since $\|\lambda u\| = |\lambda| \|u\|$, for every nonzero vector x, we have

$$\frac{\|Ax\|}{\|x\|} = \frac{\|x\| \|A(x/\|x\|)\|}{\|x\|} = \|A(x/\|x\|)\|,$$

which implies that

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\| = 1}} \|Ax\|.$$

Similarly

$$\sup_{\substack{x\in\mathbb{R}^n\\ x\neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x\in\mathbb{R}^n\\ \|x\|=1}} \|Ax\|.$$

The above considerations justify the following definition.

Definition 9.7. If $\| \|$ is any norm on $\mathbb{C}^n$, we define the function $\| \|_{\mathrm{op}}$ on $\mathrm{M}_n(\mathbb{C})$ by

$$\|A\|_{\mathrm{op}} = \sup_{\substack{x\in\mathbb{C}^n\\ x\neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x\in\mathbb{C}^n\\ \|x\|=1}} \|Ax\|.$$

The function $A \mapsto \|A\|_{\mathrm{op}}$ is called the *subordinate matrix norm* or *operator norm* induced by the norm $\| \|$.

Another notation for the operator norm of a matrix A (in particular, used by Horn and Johnson [95]), is $|||A|||$.

It is easy to check that the function $A \mapsto \|A\|_{\mathrm{op}}$ is indeed a norm, and by definition, it satisfies the property

$$\|Ax\| \leq \|A\|_{\mathrm{op}} \|x\|, \quad \text{for all } x \in \mathbb{C}^n.$$

A norm $\| \|_{\mathrm{op}}$ on $\mathrm{M}_n(\mathbb{C})$ satisfying the above property is said to be *subordinate* to the vector norm $\| \|$ on $\mathbb{C}^n$. As a consequence of the above inequality, we have

$$\|ABx\| \leq \|A\|_{\mathrm{op}} \|Bx\| \leq \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \|x\|,$$

for all $x \in \mathbb{C}^n$, which implies that

$$\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \quad \text{for all } A, B \in \mathrm{M}_n(\mathbb{C}),$$

showing that $A \mapsto \|A\|_{\mathrm{op}}$ is a matrix norm (it is submultiplicative).

Observe that the operator norm is also defined by

$$\|A\|_{\mathrm{op}} = \inf\{\lambda \in \mathbb{R} \mid \|Ax\| \leq \lambda \|x\|, \text{ for all } x \in \mathbb{C}^n\}.$$

Since the function $x \mapsto \|Ax\|$ is continuous (because $|\,\|Ay\| - \|Ax\|\,| \leq \|Ay - Ax\| \leq C_A \|x - y\|$) and the unit sphere $S^{n-1} = \{x \in \mathbb{C}^n \mid \|x\| = 1\}$ is compact, there is some $x \in \mathbb{C}^n$ such that $\|x\| = 1$ and

$$\|Ax\| = \|A\|_{\mathrm{op}}.$$

Equivalently, there is some $x \in \mathbb{C}^n$ such that $x \neq 0$ and

$$\|Ax\| = \|A\|_{\mathrm{op}} \|x\|.$$

Consequently we can replace sup by max in the definition of $\|A\|_{\mathrm{op}}$ (and inf by min), namely

$$\|A\|_{\mathrm{op}} = \max_{\substack{x\in\mathbb{C}^n\\ \|x\|=1}} \|Ax\|.$$

The definition of an operator norm also implies that

$$\|I\|_{\mathrm{op}} = 1.$$

The above shows that the Frobenius norm is not a subordinate matrix norm for $n \geq 2$ (why?).

If $\| \ \|$ is a vector norm on $\mathbb{C}^n$, the operator norm $\| \ \|_{\mathrm{op}}$ that it induces applies to matrices in $\mathrm{M}_n(\mathbb{C})$. If we are careful to denote vectors and matrices so that no confusion arises, for example, by using lower case letters for vectors and upper case letters for matrices, it should be clear that $\|A\|_{\mathrm{op}}$ is the operator norm of the matrix A and that $\|x\|$ is the vector norm of x. Consequently, following common practice to alleviate notation, we will drop the subscript "op" and simply write $\|A\|$ instead of $\|A\|_{\mathrm{op}}$.

The notion of subordinate norm can be slightly generalized.

Definition 9.8. If $K = \mathbb{R}$ or $K = \mathbb{C}$, for any norm $\| \ \|$ on $\mathrm{M}_{m,n}(K)$, and for any two norms $\| \ \|_a$ on K^n and $\| \ \|_b$ on K^m, we say that the norm $\| \ \|$ is *subordinate* to the norms $\| \ \|_a$ and $\| \ \|_b$ if

$$\|Ax\|_b \leq \|A\| \, \|x\|_a \quad \text{for all } A \in \mathrm{M}_{m,n}(K) \text{ and all } x \in K^n.$$

Remark: For any norm $\| \ \|$ on $\mathbb{C}^n$, we can define the function $\| \ \|_{\mathbb{R}}$ on $\mathrm{M}_n(\mathbb{R})$ by

$$\|A\|_{\mathbb{R}} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|=1}} \|Ax\|.$$

The function $A \mapsto \|A\|_{\mathbb{R}}$ is a matrix norm on $\mathrm{M}_n(\mathbb{R})$, and

$$\|A\|_{\mathbb{R}} \leq \|A\|,$$

for all real matrices $A \in \mathrm{M}_n(\mathbb{R})$. However, it is possible to construct vector norms $\| \ \|$ on $\mathbb{C}^n$ and *real* matrices A such that

$$\|A\|_{\mathbb{R}} < \|A\|.$$

In order to avoid this kind of difficulties, we define subordinate matrix norms over $\mathrm{M}_n(\mathbb{C})$. Luckily, it turns out that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms, $\| \ \|_1$, $\| \ \|_2$, and $\| \ \|_\infty$.

We now prove Proposition 9.6 for real matrix norms.

Proposition 9.9. *For any matrix norm $\| \ \|$ on $\mathrm{M}_n(\mathbb{R})$ and for any square $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have*

$$\rho(A) \leq \|A\|.$$

Proof. We follow the proof in Denis Serre's book [156]. If A is a real matrix, the problem is that the eigenvectors associated with the eigenvalue of maximum modulus may be complex. We use a trick based on the fact that for every matrix A (real or complex),

$$\rho(A^k) = (\rho(A))^k,$$

which is left as an exercise (use Proposition 15.7 which shows that if $(\lambda_1, \ldots, \lambda_n)$ are the (not necessarily distinct) eigenvalues of A, then $(\lambda_1^k, \ldots, \lambda_n^k)$ are the eigenvalues of A^k, for $k \geq 1$).

Pick any complex matrix norm $\| \, \|_c$ on $\mathbb{C}^n$ (for example, the Frobenius norm, or any subordinate matrix norm induced by a norm on $\mathbb{C}^n$). The restriction of $\| \, \|_c$ to real matrices is a real norm that we also denote by $\| \, \|_c$. Now by Theorem 9.5, since $\mathrm{M}_n(\mathbb{R})$ has finite dimension n^2, there is some constant $C > 0$ so that

$$\|B\|_c \leq C \|B\|, \quad \text{for all} \quad B \in \mathrm{M}_n(\mathbb{R}).$$

Furthermore, for every $k \geq 1$ and for every real $n \times n$ matrix A, by Proposition 9.6, $\rho(A^k) \leq \left\|A^k\right\|_c$, and because $\| \, \|$ is a matrix norm, $\left\|A^k\right\| \leq \|A\|^k$, so we have

$$(\rho(A))^k = \rho(A^k) \leq \left\|A^k\right\|_c \leq C \left\|A^k\right\| \leq C \|A\|^k,$$

for all $k \geq 1$. It follows that

$$\rho(A) \leq C^{1/k} \|A\|, \quad \text{for all} \quad k \geq 1.$$

However because $C > 0$, we have $\lim_{k \mapsto \infty} C^{1/k} = 1$ (we have $\lim_{k \mapsto \infty} \frac{1}{k} \log(C) = 0$). Therefore, we conclude that

$$\rho(A) \leq \|A\|,$$

as desired. $\square$

We now determine explicitly what are the subordinate matrix norms associated with the vector norms $\| \, \|_1$, $\| \, \|_2$, and $\| \, \|_\infty$.

Proposition 9.10. *For every square matrix $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, we have*

$$\begin{aligned}
\|A\|_1 &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_1 = 1}} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}| \\
\|A\|_\infty &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_\infty = 1}} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \\
\|A\|_2 &= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_2 = 1}} \|Ax\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}.
\end{aligned}$$

Note that $\|A\|_1$ is the maximum of the ℓ^1-norms of the columns of A and $\|A\|_\infty$ is the maximum of the ℓ^1-norms of the rows of A. Furthermore, $\|A^*\|_2 = \|A\|_2$, the norm $\| \, \|_2$ is unitarily invariant, which means that

$$\|A\|_2 = \|UAV\|_2$$

for all unitary matrices U, V, and if A is a normal matrix, then $\|A\|_2 = \rho(A)$.

Proof. For every vector u, we have

$$\|Au\|_1 = \sum_i \left| \sum_j a_{ij} u_j \right| \leq \sum_j |u_j| \sum_i |a_{ij}| \leq \left(\max_j \sum_i |a_{ij}| \right) \|u\|_1 ,$$

which implies that

$$\|A\|_1 \leq \max_j \sum_{i=1}^n |a_{ij}|.$$

It remains to show that equality can be achieved. For this let j_0 be some index such that

$$\max_j \sum_i |a_{ij}| = \sum_i |a_{ij_0}|,$$

and let $u_i = 0$ for all $i \neq j_0$ and $u_{j_0} = 1$.

In a similar way, we have

$$\|Au\|_\infty = \max_i \left| \sum_j a_{ij} u_j \right| \leq \left(\max_i \sum_j |a_{ij}| \right) \|u\|_\infty ,$$

which implies that

$$\|A\|_\infty \leq \max_i \sum_{j=1}^n |a_{ij}|.$$

To achieve equality, let i_0 be some index such that

$$\max_i \sum_j |a_{ij}| = \sum_j |a_{i_0 j}|.$$

The reader should check that the vector given by

$$u_j = \begin{cases} \frac{\overline{a}_{i_0 j}}{|a_{i_0 j}|} & \text{if } a_{i_0 j} \neq 0 \\ 1 & \text{if } a_{i_0 j} = 0 \end{cases}$$

works.

We have

$$\|A\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} \|Ax\|_2^2 = \sup_{\substack{x \in \mathbb{C}^n \\ x^* x = 1}} x^* A^* A x.$$

Since the matrix $A^* A$ is symmetric, it has real eigenvalues and it can be diagonalized with respect to a unitary matrix. These facts can be used to prove that the function $x \mapsto x^* A^* A x$ has a maximum on the sphere $x^* x = 1$ equal to the largest eigenvalue of $A^* A$, namely, $\rho(A^* A)$. We postpone the proof until we discuss optimizing quadratic functions. Therefore,

$$\|A\|_2 = \sqrt{\rho(A^* A)}.$$

Let use now prove that $\rho(A^*A) = \rho(AA^*)$. First assume that $\rho(A^*A) > 0$. In this case, there is some eigenvector $u\,(\neq 0)$ such that

$$A^*Au = \rho(A^*A)u,$$

and since $\rho(A^*A) > 0$, we must have $Au \neq 0$. Since $Au \neq 0$,

$$AA^*(Au) = A(A^*Au) = \rho(A^*A)Au$$

which means that $\rho(A^*A)$ is an eigenvalue of AA^*, and thus

$$\rho(A^*A) \leq \rho(AA^*).$$

Because $(A^*)^* = A$, by replacing A by A^*, we get

$$\rho(AA^*) \leq \rho(A^*A),$$

and so $\rho(A^*A) = \rho(AA^*)$.

If $\rho(A^*A) = 0$, then we must have $\rho(AA^*) = 0$, since otherwise by the previous reasoning we would have $\rho(A^*A) = \rho(AA^*) > 0$. Hence, in all case

$$\|A\|_2^2 = \rho(A^*A) = \rho(AA^*) = \|A^*\|_2^2.$$

For any unitary matrices U and V, it is an easy exercise to prove that V^*A^*AV and A^*A have the same eigenvalues, so

$$\|A\|_2^2 = \rho(A^*A) = \rho(V^*A^*AV) = \|AV\|_2^2,$$

and also

$$\|A\|_2^2 = \rho(A^*A) = \rho(A^*U^*UA) = \|UA\|_2^2.$$

Finally, if A is a normal matrix ($AA^* = A^*A$), it can be shown that there is some unitary matrix U so that

$$A = UDU^*,$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix consisting of the eigenvalues of A, and thus

$$A^*A = (UDU^*)^*UDU^* = UD^*U^*UDU^* = UD^*DU^*.$$

However, $D^*D = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$, which proves that

$$\rho(A^*A) = \rho(D^*D) = \max_i |\lambda_i|^2 = (\rho(A))^2,$$

so that $\|A\|_2 = \rho(A)$. □

Definition 9.9. For $A = (a_{ij}) \in \mathrm{M}_n(\mathbb{C})$, the norm $\|A\|_2$ is often called the *spectral norm*.

Observe that Property (3) of Proposition 9.7 says that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2\,,$$

which shows that the Frobenius norm is an upper bound on the spectral norm. The Frobenius norm is much easier to compute than the spectral norm.

The reader will check that the above proof still holds if the matrix A is real (change unitary to orthogonal), confirming the fact that $\|A\|_{\mathbb{R}} = \|A\|$ for the vector norms $\| \|_1$, $\| \|_2$, and $\| \|_\infty$. It is also easy to verify that the proof goes through for *rectangular* $m \times n$ matrices, with the same formulae. Similarly, the Frobenius norm given by

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\mathrm{tr}(AA^*)}$$

is also a norm on rectangular matrices. For these norms, whenever AB makes sense, we have

$$\|AB\| \leq \|A\|\,\|B\|\,.$$

Remark: It can be shown that for any two real numbers $p, q \geq 1$ such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$, we have

$$\|A^*\|_q = \|A\|_p = \sup\{\Re(y^*Ax) \mid \|x\|_p = 1, \|y\|_q = 1\} = \sup\{|\langle Ax, y\rangle| \mid \|x\|_p = 1, \|y\|_q = 1\},$$

where $\|A^*\|_q$ and $\|A\|_p$ are the operator norms.

Remark: Let $(E, \| \|)$ and $(F, \| \|)$ be two normed vector spaces (for simplicity of notation, we use the same symbol $\| \|$ for the norms on E and F; this should not cause any confusion). Recall that a function $f \colon E \to F$ is *continuous* if for every $a \in E$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for all $x \in E$,

$$\text{if} \quad \|x - a\| \leq \eta \quad \text{then} \quad \|f(x) - f(a)\| \leq \epsilon.$$

It is not hard to show that a *linear map* $f \colon E \to F$ is continuous iff there is some constant $C \geq 0$ such that

$$\|f(x)\| \leq C\,\|x\| \text{ for all } x \in E.$$

If so, we say that f is *bounded* (or a *linear bounded operator*). We let $\mathcal{L}(E; F)$ denote the set of all continuous (equivalently, bounded) linear maps from E to F. Then we can define the *operator norm* (or *subordinate norm*) $\| \|$ on $\mathcal{L}(E; F)$ as follows: for every $f \in \mathcal{L}(E; F)$,

$$\|f\| = \sup_{\substack{x \in E \\ x \neq 0}} \frac{\|f(x)\|}{\|x\|} = \sup_{\substack{x \in E \\ \|x\| = 1}} \|f(x)\|\,,$$

or equivalently by

$$\|f\| = \inf\{\lambda \in \mathbb{R} \mid \|f(x)\| \leq \lambda \|x\|, \text{ for all } x \in E\}.$$

Here because E may be infinite-dimensional, sup can't be replaced by max and inf can't be replaced by min. It is not hard to show that the map $f \mapsto \|f\|$ is a norm on $\mathcal{L}(E;F)$ satisfying the property

$$\|f(x)\| \leq \|f\| \|x\|$$

for all $x \in E$, and that if $f \in \mathcal{L}(E;F)$ and $g \in \mathcal{L}(F;G)$, then

$$\|g \circ f\| \leq \|g\| \|f\|.$$

Operator norms play an important role in functional analysis, especially when the spaces E and F are *complete*.

9.4 Inequalities Involving Subordinate Norms

In this section we discuss two technical inequalities which will be needed for certain proofs in the last three sections of this chapter. First we prove a proposition which will be needed when we deal with the condition number of a matrix.

Proposition 9.11. *Let $\| \ \|$ be any matrix norm, and let $B \in \mathrm{M}_n(\mathbb{C})$ such that $\|B\| < 1$.*

(1) If $\| \ \|$ is a subordinate matrix norm, then the matrix $I + B$ is invertible and

$$\left\|(I + B)^{-1}\right\| \leq \frac{1}{1 - \|B\|}.$$

(2) If a matrix of the form $I + B$ is singular, then $\|B\| \geq 1$ for every matrix norm (not necessarily subordinate).

Proof. (1) Observe that $(I + B)u = 0$ implies $Bu = -u$, so

$$\|u\| = \|Bu\|.$$

Recall that

$$\|Bu\| \leq \|B\| \|u\|$$

for every subordinate norm. Since $\|B\| < 1$, if $u \neq 0$, then

$$\|Bu\| < \|u\|,$$

which contradicts $\|u\| = \|Bu\|$. Therefore, we must have $u = 0$, which proves that $I + B$ is injective, and thus bijective, i.e., invertible. Then we have

$$(I + B)^{-1} + B(I + B)^{-1} = (I + B)(I + B)^{-1} = I,$$

so we get

$$(I+B)^{-1} = I - B(I+B)^{-1},$$

which yields

$$\left\|(I+B)^{-1}\right\| \leq 1 + \|B\| \left\|(I+B)^{-1}\right\|,$$

and finally,

$$\left\|(I+B)^{-1}\right\| \leq \frac{1}{1-\|B\|}.$$

(2) If $I+B$ is singular, then -1 is an eigenvalue of B, and by Proposition 9.6, we get $\rho(B) \leq \|B\|$, which implies $1 \leq \rho(B) \leq \|B\|$. $\square$

The second inequality is a result is that is needed to deal with the convergence of sequences of powers of matrices.

Proposition 9.12. *For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \ \|$ such that*

$$\|A\| \leq \rho(A) + \epsilon.$$

Proof. By Theorem 15.5, there exists some invertible matrix U and some upper triangular matrix T such that

$$A = UTU^{-1},$$

and say that

$$T = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. For every $\delta \neq 0$, define the diagonal matrix

$$D_\delta = \mathrm{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1}),$$

and consider the matrix

$$(UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}TD_\delta = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1} t_{1n} \\ 0 & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2} t_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \delta t_{n-1\,n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Now define the function $\| \ \| \colon \mathrm{M}_n(\mathbb{C}) \to \mathbb{R}$ by

$$\|B\| = \left\|(UD_\delta)^{-1}B(UD_\delta)\right\|_\infty,$$

for every $B \in \mathrm{M}_n(\mathbb{C})$. Then it is easy to verify that the above function is the matrix norm subordinate to the vector norm

$$v \mapsto \left\|(UD_\delta)^{-1}v\right\|_\infty.$$

Furthermore, for every $\epsilon > 0$, we can pick δ so that

$$\sum_{j=i+1}^{n} |\delta^{j-i} t_{ij}| \leq \epsilon, \quad 1 \leq i \leq n-1,$$

and by definition of the norm $\| \, \|_\infty$, we get

$$\|A\| \leq \rho(A) + \epsilon,$$

which shows that the norm that we have constructed satisfies the required properties. $\square$

Note that equality is generally not possible; consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

for which $\rho(A) = 0 < \|A\|$, since $A \neq 0$.

9.5 Condition Numbers of Matrices

Unfortunately, there exist linear systems $Ax = b$ whose solutions are not stable under small perturbations of either b or A. For example, consider the system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

The reader should check that it has the solution $x = (1, 1, 1, 1)$. If we perturb slightly the right-hand side as $b + \Delta b$, where

$$\Delta b = \begin{pmatrix} 0.1 \\ -0.1 \\ 0.1 \\ -0.1 \end{pmatrix},$$

we obtain the new system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix}.$$

The new solution turns out to be $x + \Delta x = (9.2, -12.6, 4.5, -1.1)$, where

$$\Delta x = (9.2, -12.6, 4.5, -1.1) - (1, 1, 1, 1) = (8.2, -13.6, 3.5, -2.1).$$

Then a relative error of the data in terms of the one-norm,

$$\frac{\|\Delta b\|_1}{\|b\|_1} = \frac{0.4}{119} = \frac{4}{1190} \approx \frac{1}{300},$$

produces a relative error in the input

$$\frac{\|\Delta x\|_1}{\|x\|_1} = \frac{27.4}{4} \approx 7.$$

So a relative error of the order $1/300$ in the data produces a relative error of the order $7/1$ in the solution, which represents an amplification of the relative error of the order 2100.

Now let us perturb the matrix slightly, obtaining the new system

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.98 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \\ x_3 + \Delta x_3 \\ x_4 + \Delta x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

This time the solution is $x + \Delta x = (-81, 137, -34, 22)$. Again a small change in the data alters the result rather drastically. Yet the original system is symmetric, has determinant 1, and has integer entries. The problem is that the matrix of the system is badly conditioned, a concept that we will now explain.

Given an invertible matrix A, first assume that we perturb b to $b + \Delta b$, and let us analyze the change between the two exact solutions x and $x + \Delta x$ of the two systems

$$\begin{aligned} Ax &= b \\ A(x + \Delta x) &= b + \Delta b. \end{aligned}$$

We also assume that we have some norm $\| \ \|$ and we use the *subordinate* matrix norm on matrices. From

$$\begin{aligned} Ax &= b \\ Ax + A\Delta x &= b + \Delta b, \end{aligned}$$

we get

$$\Delta x = A^{-1} \Delta b,$$

and we conclude that

$$\begin{aligned} \|\Delta x\| &\leq \left\|A^{-1}\right\| \|\Delta b\| \\ \|b\| &\leq \|A\| \|x\|. \end{aligned}$$

Consequently, the relative error in the result $\|\Delta x\| / \|x\|$ is bounded in terms of the relative error $\|\Delta b\| / \|b\|$ in the data as follows:

$$\frac{\|\Delta x\|}{\|x\|} \leq \left(\|A\| \left\|A^{-1}\right\| \right) \frac{\|\Delta b\|}{\|b\|}.$$

Now let us assume that A is perturbed to $A + \Delta A$, and let us analyze the change between the exact solutions of the two systems

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b. \end{aligned}$$

The second equation yields $Ax + A\Delta x + \Delta A(x + \Delta x) = b$, and by subtracting the first equation we get

$$\Delta x = -A^{-1} \Delta A(x + \Delta x).$$

It follows that

$$\|\Delta x\| \leq \left\|A^{-1}\right\| \|\Delta A\| \|x + \Delta x\|,$$

which can be rewritten as

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \left(\|A\| \left\|A^{-1}\right\| \right) \frac{\|\Delta A\|}{\|A\|}.$$

Observe that the above reasoning is valid even if the matrix $A + \Delta A$ is singular, as long as $x + \Delta x$ is a solution of the second system. Furthermore, if $\|\Delta A\|$ is small enough, it is not unreasonable to expect that the ratio $\|\Delta x\| / \|x + \Delta x\|$ is close to $\|\Delta x\| / \|x\|$. This will be made more precise later.

In summary, for each of the two perturbations, we see that the relative error in the result is bounded by the relative error in the data, *multiplied the number* $\|A\| \|A^{-1}\|$. In fact, this factor turns out to be optimal and this suggests the following definition:

Definition 9.10. For any subordinate matrix norm $\| \, \|$, for any invertible matrix A, the number

$$\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

is called the *condition number* of A relative to $\| \, \|$.

The condition number $\mathrm{cond}(A)$ measures the sensitivity of the linear system $Ax = b$ to variations in the data b and A; a feature referred to as the *condition* of the system. Thus, when we says that a linear system is *ill-conditioned*, we mean that the condition number of its matrix is large. We can sharpen the preceding analysis as follows:

Proposition 9.13. *Let A be an invertible matrix and let x and $x + \Delta x$ be the solutions of the linear systems*

$$\begin{aligned} Ax &= b \\ A(x + \Delta x) &= b + \Delta b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta b\|}{\|b\|}$$

holds and is the best possible. This means that for a given matrix A, there exist some vectors $b \neq 0$ and $\Delta b \neq 0$ for which equality holds.

Proof. We already proved the inequality. Now, because $\| \, \|$ is a subordinate matrix norm, there exist some vectors $x \neq 0$ and $\Delta b \neq 0$ for which

$$\left\|A^{-1}\Delta b\right\| = \left\|A^{-1}\right\| \|\Delta b\| \quad \text{and} \quad \|Ax\| = \|A\| \, \|x\| \, .$$

$\square$

Proposition 9.14. *Let A be an invertible matrix and let x and $x + \Delta x$ be the solutions of the two systems*

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b. \end{aligned}$$

If $b \neq 0$, then the inequality

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

holds and is the best possible. This means that given a matrix A, there exist a vector $b \neq 0$ and a matrix $\Delta A \neq 0$ for which equality holds. Furthermore, if $\|\Delta A\|$ is small enough (for instance, if $\|\Delta A\| < 1/\left\|A^{-1}\right\|$), we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}(1 + O(\|\Delta A\|));$$

in fact, we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}\left(\frac{1}{1 - \|A^{-1}\| \, \|\Delta A\|}\right).$$

Proof. The first inequality has already been proven. To show that equality can be achieved, let w be any vector such that $w \neq 0$ and

$$\left\|A^{-1}w\right\| = \left\|A^{-1}\right\| \|w\| \, ,$$

and let $\beta \neq 0$ be any real number. Now the vectors

$$\begin{aligned} \Delta x &= -\beta A^{-1}w \\ x + \Delta x &= w \\ b &= (A + \beta I)w \end{aligned}$$

and the matrix

$$\Delta A = \beta I$$

sastisfy the equations

$$\begin{aligned} Ax &= b \\ (A + \Delta A)(x + \Delta x) &= b \\ \|\Delta x\| &= |\beta| \left\|A^{-1} w\right\| = \|\Delta A\| \left\|A^{-1}\right\| \|x + \Delta x\|. \end{aligned}$$

Finally we can pick β so that $-\beta$ is not equal to any of the eigenvalues of A, so that $A + \Delta A = A + \beta I$ is invertible and b is is nonzero.

If $\|\Delta A\| < 1/\left\|A^{-1}\right\|$, then

$$\left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| < 1,$$

so by Proposition 9.11, the matrix $I + A^{-1}\Delta A$ is invertible and

$$\left\|(I + A^{-1}\Delta A)^{-1}\right\| \leq \frac{1}{1 - \|A^{-1}\Delta A\|} \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|}.$$

Recall that we proved earlier that

$$\Delta x = -A^{-1}\Delta A(x + \Delta x),$$

and by adding x to both sides and moving the right-hand side to the left-hand side yields

$$(I + A^{-1}\Delta A)(x + \Delta x) = x,$$

and thus

$$x + \Delta x = (I + A^{-1}\Delta A)^{-1} x,$$

which yields

$$\begin{aligned} \Delta x &= ((I + A^{-1}\Delta A)^{-1} - I)x = (I + A^{-1}\Delta A)^{-1}(I - (I + A^{-1}\Delta A))x \\ &= -(I + A^{-1}\Delta A)^{-1} A^{-1} (\Delta A) x. \end{aligned}$$

From this and

$$\left\|(I + A^{-1}\Delta A)^{-1}\right\| \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|},$$

we get

$$\|\Delta x\| \leq \frac{\|A^{-1}\| \|\Delta A\|}{1 - \|A^{-1}\| \|\Delta A\|} \|x\|,$$

which can be written as

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\Delta A\|}{\|A\|} \left(\frac{1}{1 - \|A^{-1}\| \|\Delta A\|} \right),$$

which is the kind of inequality that we were seeking. $\square$

Remark: If A and b are perturbed simultaneously, so that we get the "perturbed" system

$$(A + \Delta A)(x + \Delta x) = b + \Delta b,$$

it can be shown that if $\|\Delta A\| < 1/\|A^{-1}\|$ (and $b \neq 0$), then

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\| \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right);$$

see Demmel [48], Section 2.2 and Horn and Johnson [95], Section 5.8.

We now list some properties of condition numbers and figure out what $\mathrm{cond}(A)$ is in the case of the spectral norm (the matrix norm induced by $\| \|_2$). First, we need to introduce a very important factorization of matrices, the *singular value decomposition*, for short, *SVD*.

It can be shown (see Section 22.2) that given any $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, there exist two unitary matrices U and V, and a *real* diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, such that

$$A = V \Sigma U^*.$$

Definition 9.11. Given a complex $n \times n$ matrix A, a triple (U, V, Σ) such that $A = V \Sigma U^*$, where U and V are $n \times n$ unitary matrices and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix of real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, is called a *singular decomposition* (for short *SVD*) of A. If A is a real matrix, then U and V are orthogonal matrices The nonnegative numbers $\sigma_1, \ldots, \sigma_n$ are called the *singular values* of A.

The factorization $A = V \Sigma U^*$ implies that

$$A^*A = U \Sigma^2 U^* \quad \text{and} \quad AA^* = V \Sigma^2 V^*,$$

which shows that $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of *both* A^*A and AA^*, that the columns of U are corresponding eigenvectors for A^*A, and that the columns of V are corresponding eigenvectors for AA^*.

Since σ_1^2 is the largest eigenvalue of A^*A (and AA^*), note that $\sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \sigma_1$.

Corollary 9.15. *The spectral norm $\|A\|_2$ of a matrix A is equal to the largest singular value of A. Equivalently, the spectral norm $\|A\|_2$ of a matrix A is equal to the ℓ^∞-norm of its vector of singular values,*

$$\|A\|_2 = \max_{1 \leq i \leq n} \sigma_i = \|(\sigma_1, \ldots, \sigma_n)\|_\infty .$$

Since the Frobenius norm of a matrix A is defined by $\|A\|_F = \sqrt{\mathrm{tr}(A^*A)}$ and since

$$\mathrm{tr}(A^*A) = \sigma_1^2 + \cdots + \sigma_n^2$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of A^*A, we see that

$$\|A\|_F = (\sigma_1^2 + \cdots + \sigma_n^2)^{1/2} = \|(\sigma_1, \ldots, \sigma_n)\|_2 .$$

Corollary 9.16. *The Frobenius norm of a matrix is given by the ℓ^2-norm of its vector of singular values;* $\|A\|_F = \|(\sigma_1, \ldots, \sigma_n)\|_2$.

In the case of a normal matrix if $\lambda_1, \ldots, \lambda_n$ are the (complex) eigenvalues of A, then

$$\sigma_i = |\lambda_i|, \quad 1 \leq i \leq n.$$

Proposition 9.17. *For every invertible matrix $A \in \mathrm{M}_n(\mathbb{C})$, the following properties hold:*

(1)

$$\begin{aligned} \operatorname{cond}(A) &\geq 1, \\ \operatorname{cond}(A) &= \operatorname{cond}(A^{-1}) \\ \operatorname{cond}(\alpha A) &= \operatorname{cond}(A) \quad \textit{for all } \alpha \in \mathbb{C} - \{0\}. \end{aligned}$$

(2) If $\operatorname{cond}_2(A)$ denotes the condition number of A with respect to the spectral norm, then

$$\operatorname{cond}_2(A) = \frac{\sigma_1}{\sigma_n},$$

where $\sigma_1 \geq \cdots \geq \sigma_n$ are the singular values of A.

(3) If the matrix A is normal, then

$$\operatorname{cond}_2(A) = \frac{|\lambda_1|}{|\lambda_n|},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A sorted so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$.

(4) If A is a unitary or an orthogonal matrix, then

$$\operatorname{cond}_2(A) = 1.$$

(5) The condition number $\operatorname{cond}_2(A)$ is invariant under unitary transformations, which means that

$$\operatorname{cond}_2(A) = \operatorname{cond}_2(UA) = \operatorname{cond}_2(AV),$$

for all unitary matrices U and V.

Proof. The properties in (1) are immediate consequences of the properties of subordinate matrix norms. In particular, $AA^{-1} = I$ implies

$$1 = \|I\| \leq \|A\| \left\|A^{-1}\right\| = \operatorname{cond}(A).$$

(2) We showed earlier that $\|A\|_2^2 = \rho(A^*A)$, which is the square of the modulus of the largest eigenvalue of A^*A. Since we just saw that the eigenvalues of A^*A are $\sigma_1^2 \geq \cdots \geq \sigma_n^2$, where $\sigma_1, \ldots, \sigma_n$ are the singular values of A, we have

$$\|A\|_2 = \sigma_1.$$

Now if A is invertible, then $\sigma_1 \geq \cdots \geq \sigma_n > 0$, and it is easy to show that the eigenvalues of $(A^*A)^{-1}$ are $\sigma_n^{-2} \geq \cdots \geq \sigma_1^{-2}$, which shows that

$$\left\|A^{-1}\right\|_2 = \sigma_n^{-1},$$

and thus

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_n}.$$

(3) This follows from the fact that $\|A\|_2 = \rho(A)$ for a normal matrix.

(4) If A is a unitary matrix, then $A^*A = AA^* = I$, so $\rho(A^*A) = 1$, and $\|A\|_2 = \sqrt{\rho(A^*A)} = 1$. We also have $\|A^{-1}\|_2 = \|A^*\|_2 = \sqrt{\rho(AA^*)} = 1$, and thus $\mathrm{cond}(A) = 1$.

(5) This follows immediately from the unitary invariance of the spectral norm. $\square$

Proposition 9.17 (4) shows that unitary and orthogonal transformations are very well-conditioned, and Part (5) shows that unitary transformations preserve the condition number.

In order to compute $\mathrm{cond}_2(A)$, we need to compute the top and bottom singular values of A, which may be hard. The inequality

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2\,,$$

may be useful in getting an approximation of $\mathrm{cond}_2(A) = \|A\|_2\,\|A^{-1}\|_2$, if A^{-1} can be determined.

Remark: There is an interesting geometric characterization of $\mathrm{cond}_2(A)$. If $\theta(A)$ denotes the least angle between the vectors Au and Av as u and v range over all pairs of orthonormal vectors, then it can be shown that

$$\mathrm{cond}_2(A) = \cot(\theta(A)/2)).$$

Thus if A is nearly singular, then there will be some orthonormal pair u, v such that Au and Av are nearly parallel; the angle $\theta(A)$ will the be small and $\cot(\theta(A)/2))$ will be large. For more details, see Horn and Johnson [95] (Section 5.8 and Section 7.4).

It should be noted that in general (if A is not a normal matrix) a matrix could have a very large condition number even if all its eigenvalues are identical! For example, if we consider the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 2 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & 2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 & 2 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 & 2 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{pmatrix},$$

it turns out that $\mathrm{cond}_2(A) \geq 2^{n-1}$.

A classical example of matrix with a very large condition number is the *Hilbert matrix* $H^{(n)}$, the $n \times n$ matrix with

$$H_{ij}^{(n)} = \left(\frac{1}{i+j-1}\right).$$

For example, when $n = 5$,

$$H^{(5)} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix}.$$

It can be shown that

$$\mathrm{cond}_2(H^{(5)}) \approx 4.77 \times 10^5.$$

Hilbert introduced these matrices in 1894 while studying a problem in approximation theory. The Hilbert matrix $H^{(n)}$ is symmetric positive definite. A closed-form formula can be given for its determinant (it is a special form of the so-called *Cauchy determinant*); see Problem 9.15. The inverse of $H^{(n)}$ can also be computed explicitly; see Problem 9.15. It can be shown that

$$\mathrm{cond}_2(H^{(n)}) = O((1+\sqrt{2})^{4n}/\sqrt{n}).$$

Going back to our matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix},$$

which is a symmetric positive definite matrix, it can be shown that its eigenvalues, which in this case are also its singular values because A is SPD, are

$$\lambda_1 \approx 30.2887 > \lambda_2 \approx 3.858 > \lambda_3 \approx 0.8431 > \lambda_4 \approx 0.01015,$$

so that

$$\mathrm{cond}_2(A) = \frac{\lambda_1}{\lambda_4} \approx 2984.$$

The reader should check that for the perturbation of the right-hand side b used earlier, the relative errors $\|\Delta x\| / \|x\|$ and $\|\Delta x\| / \|x\|$ satisfy the inequality

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\Delta b\|}{\|b\|}$$

and comes close to equality.

9.6 An Application of Norms: Solving Inconsistent Linear Systems

The problem of solving an inconsistent linear system $Ax = b$ often arises in practice. This is a system where b does not belong to the column space of A, usually with more equations than variables. Thus, such a system has no solution. Yet we would still like to "solve" such a system, at least approximately.

Such systems often arise when trying to fit some data. For example, we may have a set of 3D data points

$$\{p_1, \ldots, p_n\},$$

and we have reason to believe that these points are nearly coplanar. We would like to find a plane that best fits our data points. Recall that the equation of a plane is

$$\alpha x + \beta y + \gamma z + \delta = 0,$$

with $(\alpha, \beta, \gamma) \neq (0, 0, 0)$. Thus, every plane is either not parallel to the x-axis ($\alpha \neq 0$) or not parallel to the y-axis ($\beta \neq 0$) or not parallel to the z-axis ($\gamma \neq 0$).

Say we have reasons to believe that the plane we are looking for is not parallel to the z-axis. If we are wrong, in the least squares solution, one of the coefficients, α, β, will be very large. If $\gamma \neq 0$, then we may assume that our plane is given by an equation of the form

$$z = ax + by + d,$$

and we would like this equation to be satisfied for all the p_i's, which leads to a system of n equations in 3 unknowns a, b, d, with $p_i = (x_i, y_i, z_i)$;

$$\begin{aligned} ax_1 + by_1 + d &= z_1 \\ \vdots \qquad & \quad \vdots \\ ax_n + by_n + d &= z_n. \end{aligned}$$

However, if n is larger than 3, such a system generally has *no solution*. Since the above system can't be solved exactly, we can try to find a solution (a, b, d) that *minimizes the least-squares error*

$$\sum_{i=1}^{n} (ax_i + by_i + d - z_i)^2.$$

This is what Legendre and Gauss figured out in the early 1800's!

In general, given a linear system

$$Ax = b,$$

we solve the *least squares problem*: minimize $\|Ax - b\|_2^2$.

Fortunately, every $n \times m$-matrix A can be written as

$$A = VDU^\top$$

where U and V are orthogonal and D is a rectangular diagonal matrix with non-negative entries (*singular value decomposition, or SVD*); see Chapter 22.

The SVD can be used to solve an inconsistent system. It is shown in Chapter 23 that there is a vector x of smallest norm minimizing $\|Ax - b\|_2$. It is given by the (Penrose) *pseudo-inverse* of A (itself given by the SVD).

It has been observed that solving in the least-squares sense may give too much weight to "outliers," that is, points clearly outside the best-fit plane. In this case, it is preferable to minimize (the ℓ^1-norm)

$$\sum_{i=1}^{n} |ax_i + by_i + d - z_i|.$$

This does not appear to be a linear problem, but we can use a trick to convert this minimization problem into a linear program (which means a problem involving linear constraints).

Note that $|x| = \max\{x, -x\}$. So by introducing new variables $e_1, \ldots, e_n$, our minimization problem is equivalent to the linear program (LP):

$$\begin{aligned} \text{minimize} \qquad & e_1 + \cdots + e_n \\ \text{subject to} \qquad & ax_i + by_i + d - z_i \leq e_i \\ & -(ax_i + by_i + d - z_i) \leq e_i \\ & 1 \leq i \leq n. \end{aligned}$$

Observe that the constraints are equivalent to

$$e_i \geq |ax_i + by_i + d - z_i|, \qquad 1 \leq i \leq n.$$

For an optimal solution, we must have equality, since otherwise we could decrease some e_i and get an even better solution. Of course, we are no longer dealing with "pure" linear algebra, since our constraints are inequalities.

We prefer not getting into linear programming right now, but the above example provides a good reason to learn more about linear programming!

9.7 Limits of Sequences and Series

If $x \in \mathbb{R}$ or $x \in \mathbb{C}$ and if $|x| < 1$, it is well known that the sums $\sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n$ converge to the limit $1/(1 - x)$ when n goes to infinity, and we write

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x}.$$

For example,

$$\sum_{k=0}^{\infty} \frac{1}{2^k} = 2.$$

Similarly, the sums

$$S_n = \sum_{k=0}^{n} \frac{x^k}{k!}$$

converge to e^x when n goes to infinity, for every x (in $\mathbb{R}$ or $\mathbb{C}$). What if we replace x by a real or complex $n \times n$ matrix A?

The partial sums $\sum_{k=0}^{n} A^k$ and $\sum_{k=0}^{n} \frac{A^k}{k!}$ still make sense, but we have to define what is the limit of a sequence of matrices. This can be done in any normed vector space.

Definition 9.12. Let $(E, \|\|)$ be a normed vector space. A *sequence* $(u_n)_{n\in\mathbb{N}}$ in E is any function $u\colon \mathbb{N} \to E$. For any $v \in E$, the sequence (u_n) *converges to* v (and v *is the limit of the sequence* (u_n)) if for every $\epsilon > 0$, there is some integer $N > 0$ such that

$$\|u_n - v\| < \epsilon \quad \text{for all } n \geq N.$$

Often we assume that a sequence is indexed by $\mathbb{N} - \{0\}$, that is, its first term is u_1 rather than u_0.

If the sequence (u_n) converges to v, then since by the triangle inequality

$$\|u_m - u_n\| \leq \|u_m - v\| + \|v - u_n\|,$$

we see that for every $\epsilon > 0$, we can find $N > 0$ such that $\|u_m - v\| < \epsilon/2$ and $\|u_n - v\| < \epsilon/2$ for all $m, n \geq N$, and so

$$\|u_m - u_n\| < \epsilon \quad \text{for all } m, n \geq N.$$

The above property is *necessary* for a convergent sequence, but *not necessarily* sufficient. For example, if $E = \mathbb{Q}$, there are sequences of rationals satisfying the above condition, but whose limit is not a rational number. For example, the sequence $\sum_{k=1}^{n} \frac{1}{k!}$ converges to e, and the sequence $\sum_{k=0}^{n} (-1)^k \frac{1}{2k+1}$ converges to $\pi/4$, but e and $\pi/4$ are not rational (in fact, they are transcendental). However, $\mathbb{R}$ is constructed from $\mathbb{Q}$ to guarantee that sequences with the above property converge, and so is $\mathbb{C}$.

Definition 9.13. Given a normed vector space $(E, \| \|)$, a sequence (u_n) is a *Cauchy sequence* if for every $\epsilon > 0$, there is some $N > 0$ such that

$$\|u_m - u_n\| < \epsilon \quad \text{for all } m, n \geq N.$$

If every Cauchy sequence converges, then we say that E is *complete*. A complete normed vector spaces is also called a *Banach space*.

A fundamental property of $\mathbb{R}$ is that *it is complete*. It follows immediately that $\mathbb{C}$ is also complete. If E is a finite-dimensional real or complex vector space, since any two norms are equivalent, we can pick the ℓ^∞ norm, and then by picking a basis in E, a sequence (u_n) of vectors in E converges iff the n sequences of coordinates (u_n^i) $(1 \leq i \leq n)$ converge, so *any finite-dimensional real or complex vector space is a Banach space*.

Let us now consider the convergence of series.

Definition 9.14. Given a normed vector space $(E, \| \|)$, a *series* is an infinite sum $\sum_{k=0}^\infty u_k$ of elements $u_k \in E$. We denote by S_n the partial sum of the first $n+1$ elements,

$$S_n = \sum_{k=0}^n u_k.$$

Definition 9.15. We say that the series $\sum_{k=0}^\infty u_k$ *converges* to the limit $v \in E$ if the sequence (S_n) converges to v, i.e., given any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$,

$$\|S_n - v\| < \epsilon.$$

In this case, we say that the series is *convergent*. We say that the series $\sum_{k=0}^\infty u_k$ *converges absolutely* if the series of norms $\sum_{k=0}^\infty \|u_k\|$ is convergent.

If the series $\sum_{k=0}^\infty u_k$ converges to v, since for all m, n with $m > n$ we have

$$\sum_{k=0}^m u_k - S_n = \sum_{k=0}^m u_k - \sum_{k=0}^n u_k = \sum_{k=n+1}^m u_k,$$

if we let m go to infinity (with n fixed), we see that the series $\sum_{k=n+1}^\infty u_k$ converges and that

$$v - S_n = \sum_{k=n+1}^\infty u_k.$$

There are series that are convergent but not absolutely convergent; for example, the series

$$\sum_{k=1}^\infty \frac{(-1)^{k-1}}{k}$$

converges to $\ln 2$, but $\sum_{k=1}^\infty \frac{1}{k}$ does not converge (this sum is infinite).

If E is complete, the converse is an enormously useful result.

Proposition 9.18. *Assume $(E, \| \|)$ is a complete normed vector space. If a series $\sum_{k=0}^\infty u_k$ is absolutely convergent, then it is convergent.*

Proof. If $\sum_{k=0}^{\infty} u_k$ is absolutely convergent, then we prove that the sequence (S_m) is a Cauchy sequence; that is, for every $\epsilon > 0$, there is some $p > 0$ such that for all $n \geq m \geq p$,

$$\|S_n - S_m\| \leq \epsilon.$$

Observe that

$$\|S_n - S_m\| = \|u_{m+1} + \cdots + u_n\| \leq \|u_{m+1}\| + \cdots + \|u_n\|,$$

and since the sequence $\sum_{k=0}^{\infty} \|u_k\|$ converges, it satisfies Cauchy's criterion. Thus, the sequence (S_m) also satisfies Cauchy's criterion, and since E is a complete vector space, the sequence (S_m) converges. ∎

Remark: It can be shown that if $(E, \| \|)$ is a normed vector space such that every absolutely convergent series is also convergent, then E must be complete (see Schwartz [150]).

An important corollary of absolute convergence is that if the terms in series $\sum_{k=0}^{\infty} u_k$ are rearranged, then the resulting series is still absolutely convergent and has the *same sum*. More precisely, let σ be any permutation (bijection) of the natural numbers. The series $\sum_{k=0}^{\infty} u_{\sigma(k)}$ is called a *rearrangement* of the original series. The following result can be shown (see Schwartz [150]).

Proposition 9.19. *Assume $(E, \| \|)$ is a normed vector space. If a series $\sum_{k=0}^{\infty} u_k$ is convergent as well as absolutely convergent, then for every permutation σ of $\mathbb{N}$, the series $\sum_{k=0}^{\infty} u_{\sigma(k)}$ is convergent and absolutely convergent, and its sum is equal to the sum of the original series:*

$$\sum_{k=0}^{\infty} u_{\sigma(k)} = \sum_{k=0}^{\infty} u_k.$$

In particular, if $(E, \| \|)$ is a complete normed vector space, then Proposition 9.19 holds.

We now apply Proposition 9.18 to the matrix exponential.

9.8 The Matrix Exponential

Proposition 9.20. *For any $n \times n$ real or complex matrix A, the series*

$$\sum_{k=0}^{\infty} \frac{A^k}{k!}$$

converges absolutely for any operator norm on $\mathrm{M}_n(\mathbb{C})$ (or $\mathrm{M}_n(\mathbb{R})$).

Proof. Pick any norm on $\mathbb{C}^n$ (or $\mathbb{R}^n$) and let $\|\|$ be the corresponding operator norm on $\mathrm{M}_n(\mathbb{C})$. Since $\mathrm{M}_n(\mathbb{C})$ has dimension n^2, it is complete. By Proposition 9.18, it suffices to show that the series of nonnegative reals $\sum_{k=0}^{n}\left\|\frac{A^k}{k!}\right\|$ converges. Since $\|\ \|$ is an operator norm, this a matrix norm, so we have

$$\sum_{k=0}^{n}\left\|\frac{A^k}{k!}\right\| \leq \sum_{k=0}^{n}\frac{\|A\|^k}{k!} \leq e^{\|A\|}.$$

Thus, the nondecreasing sequence of positive real numbers $\sum_{k=0}^{n}\left\|\frac{A^k}{k!}\right\|$ is bounded by $e^{\|A\|}$, and by a fundamental property of $\mathbb{R}$, it has a least upper bound which is its limit. □

Definition 9.16. Let E be a finite-dimensional real or complex normed vector space. For any $n\times n$ matrix A, the limit of the series

$$\sum_{k=0}^{\infty}\frac{A^k}{k!}$$

is the *exponential of* A and is denoted e^A.

A basic property of the exponential $x\mapsto e^x$ with $x\in\mathbb{C}$ is

$$e^{x+y} = e^x e^y, \quad \text{for all } x,y\in\mathbb{C}.$$

As a consequence, e^x is always invertible and $(e^x)^{-1} = e^{-x}$. For matrices, because matrix multiplication is not commutative, in general,

$$e^{A+B} = e^A e^B$$

fails! This result is salvaged as follows.

Proposition 9.21. *For any two $n\times n$ complex matrices A and B, if A and B commute, that is, $AB = BA$, then*

$$e^{A+B} = e^A e^B.$$

A proof of Proposition 9.21 can be found in Gallier [72].

Since A and $-A$ commute, as a corollary of Proposition 9.21, we see that e^A is always invertible and that

$$(e^A)^{-1} = e^{-A}.$$

It is also easy to see that

$$(e^A)^\top = e^{A^\top}.$$

In general, there is no closed-form formula for the exponential e^A of a matrix A, but for skew symmetric matrices of dimension 2 and 3, there are explicit formulae. Everyone should enjoy computing the exponential e^A where

$$A = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}.$$

If we write

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

then

$$A = \theta J$$

The key property is that

$$J^2 = -I.$$

Proposition 9.22. *If $A = \theta J$, then*

$$e^A = \cos\theta I + \sin\theta J = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Proof. We have

$$\begin{aligned} A^{4n} &= \theta^{4n} I_2, \\ A^{4n+1} &= \theta^{4n+1} J, \\ A^{4n+2} &= -\theta^{4n+2} I_2, \\ A^{4n+3} &= -\theta^{4n+3} J, \end{aligned}$$

and so

$$e^A = I_2 + \frac{\theta}{1!} J - \frac{\theta^2}{2!} I_2 - \frac{\theta^3}{3!} J + \frac{\theta^4}{4!} I_2 + \frac{\theta^5}{5!} J - \frac{\theta^6}{6!} I_2 - \frac{\theta^7}{7!} J + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) I_2 + \left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right) J.$$

We recognize the power series for $\cos\theta$ and $\sin\theta$, and thus

$$e^A = \cos\theta I_2 + \sin\theta J,$$

that is

$$e^A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

as claimed. $\square$

Thus, we see that the exponential of a 2×2 skew-symmetric matrix is a rotation matrix. This property generalizes to any dimension. An explicit formula when $n = 3$ (the Rodrigues' formula) is given in Section 12.7.

Proposition 9.23. *If B is an $n \times n$ (real) skew symmetric matrix, that is, $B^\top = -B$, then $Q = e^B$ is an orthogonal matrix, that is*

$$Q^\top Q = QQ^\top = I.$$

Proof. Since $B^\top = -B$, we have

$$Q^\top = (e^B)^\top = e^{B^\top} = e^{-B}.$$

Since B and $-B$ commute, we have

$$Q^\top Q = e^{-B} e^B = e^{-B+B} = e^0 = I.$$

Similarly,

$$QQ^\top = e^B e^{-B} = e^{B-B} = e^0 = I,$$

which concludes the proof. $\square$

It can also be shown that $\det(Q) = \det(e^B) = 1$, but this requires a better understanding of the eigenvalues of e^B (see Section 15.5). Furthermore, for every $n \times n$ rotation matrix Q (an orthogonal matrix Q such that $\det(Q) = 1$), there is a skew symmetric matrix B such that $Q = e^B$. This is a fundamental property which has applications in robotics for $n = 3$.

All familiar series have matrix analogs. For example, if $\|A\| < 1$ (where $\| \ \|$ is an operator norm), then the series $\sum_{k=0}^\infty A^k$ converges absolutely, and it can be shown that its limit is $(I - A)^{-1}$.

Another interesting series is the logarithm. For any $n \times n$ complex matrix A, if $\|A\| < 1$ (where $\| \ \|$ is an operator norm), then the series

$$\log(I + A) = \sum_{k=1}^\infty (-1)^{k+1} \frac{A^k}{k}$$

converges absolutely.

9.9 Summary

The main concepts and results of this chapter are listed below:

- *Norms* and *normed vector spaces*.
- The *triangle inequality*.

- The *Euclidean norm*; the *ℓ^p-norms*.
- *Hölder's inequality*; the *Cauchy–Schwarz inequality*; *Minkowski's inequality*.
- *Hermitian inner product* and *Euclidean inner product*.
- *Equivalent* norms.
- *All norms on a finite-dimensional vector space are equivalent* (Theorem 9.5).
- *Matrix norms*.
- *Hermitian, symmetric* and *normal* matrices. *Orthogonal* and *unitary* matrices.
- The *trace* of a matrix.
- *Eigenvalues* and *eigenvectors* of a matrix.
- The *characteristic polynomial* of a matrix.
- The *spectral radius* $\rho(A)$ of a matrix A.
- The *Frobenius norm*.
- The Frobenius norm is a *unitarily invariant* matrix norm.
- *Bounded* linear maps.
- *Subordinate matrix norms*.
- Characterization of the subordinate matrix norms for the vector norms $\| \|_1$, $\| \|_2$, and $\| \|_\infty$.
- The *spectral norm*.
- For every matrix $A \in \mathrm{M}_n(\mathbb{C})$ and for every $\epsilon > 0$, there is some subordinate matrix norm $\| \|$ such that $\|A\| \leq \rho(A) + \epsilon$.
- *Condition numbers* of matrices.
- Perturbation analysis of linear systems.
- The *singular value decomposition* (SVD).
- Properties of conditions numbers. Characterization of $\mathrm{cond}_2(A)$ in terms of the largest and smallest singular values of A.
- The *Hilbert matrix*: a very badly conditioned matrix.
- Solving inconsistent linear systems by the method of *least-squares*; *linear programming*.

- Convergence of sequences of vectors in a normed vector space.
- Cauchy sequences, complex normed vector spaces, Banach spaces.
- Convergence of series. Absolute convergence.
- The matrix exponential.
- Skew symmetric matrices and orthogonal matrices.

9.10 Problems

Problem 9.1. Let A be the following matrix:

$$A = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 3/2 \end{pmatrix}.$$

Compute the operator 2-norm $\|A\|_2$ of A.

Problem 9.2. Prove Proposition 9.3, namely that the following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):

$$\begin{aligned} \|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty, \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \\ \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\|x\|_2. \end{aligned}$$

Problem 9.3. For any $p \geq 1$, prove that for all $x \in \mathbb{R}^n$,

$$\lim_{p \mapsto \infty} \|x\|_p = \|x\|_\infty .$$

Problem 9.4. Let A be an $n \times n$ matrix which is strictly row diagonally dominant, which means that

$$|a_{i\,i}| > \sum_{j \neq i} |a_{i\,j}|,$$

for $i = 1, \ldots, n$, and let

$$\delta = \min_i \left\{ |a_{i\,i}| - \sum_{j \neq i} |a_{i\,j}| \right\}.$$

The fact that A is strictly row diagonally dominant is equivalent to the condition $\delta > 0$.

(1) For any nonzero vector v, prove that

$$\|Av\|_\infty \geq \|v\|_\infty \,\delta.$$

Use the above to prove that A is invertible.

(2) Prove that

$$\left\|A^{-1}\right\|_\infty \leq \delta^{-1}.$$

Hint. Prove that

$$\sup_{v\neq 0} \frac{\|A^{-1}v\|_\infty}{\|v\|_\infty} = \sup_{w\neq 0} \frac{\|w\|_\infty}{\|Aw\|_\infty}.$$

Problem 9.5. Let A be any invertible complex $n \times n$ matrix.

(1) For any vector norm $\| \;\|$ on $\mathbb{C}^n$, prove that the function $\| \;\|_A : \mathbb{C}^n \to \mathbb{R}$ given by

$$\|x\|_A = \|Ax\| \quad \text{for all} \quad x \in \mathbb{C}^n,$$

is a vector norm.

(2) Prove that the operator norm induced by $\| \;\|_A$, also denoted by $\| \;\|_A$, is given by

$$\|B\|_A = \left\|ABA^{-1}\right\| \quad \text{for every } n \times n \text{ matrix} \quad B,$$

where $\|ABA^{-1}\|$ uses the operator norm induced by $\| \;\|$.

Problem 9.6. Give an example of a norm on $\mathbb{C}^n$ and of a *real* matrix A such that

$$\|A\|_{\mathbb{R}} < \|A\|,$$

where $\|-\|_{\mathbb{R}}$ and $\|-\|$ are the operator norms associated with the vector norm $\|-\|$.

Hint. This can already be done for $n = 2$.

Problem 9.7. Let $\| \;\|$ be any operator norm. Given an invertible $n \times n$ matrix A, if $c = 1/(2\|A^{-1}\|)$, then for every $n \times n$ matrix H, if $\|H\| \leq c$, then $A + H$ is invertible. Furthermore, show that if $\|H\| \leq c$, then $\|(A + H)^{-1}\| \leq 1/c$.

Problem 9.8. Let A be any $m \times n$ matrix and let $\lambda \in \mathbb{R}$ be any positive real number $\lambda > 0$.

(1) Prove that $A^\top A + \lambda I_n$ and $AA^\top + \lambda I_m$ are invertible.

(2) Prove that

$$A^\top (AA^\top + \lambda I_m)^{-1} = (A^\top A + \lambda I_n)^{-1} A^\top.$$

Remark: The expressions above correspond to the matrix for which the function

$$\Phi(x) = (Ax - b)^\top (Ax - b) + \lambda x^\top x$$

achieves a minimum. It shows up in machine learning (kernel methods).

Problem 9.9. Let Z be a $q \times p$ real matrix. Prove that if $I_p - Z^\top Z$ is positive definite, then the $(p + q) \times (p + q)$ matrix

$$S = \begin{pmatrix} I_p & Z^\top \\ Z & I_q \end{pmatrix}$$

is symmetric positive definite.

Problem 9.10. Prove that for any real or complex square matrix A, we have

$$\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty,$$

where the above norms are operator norms.

Hint. Use Proposition 9.10 (among other things, it shows that $\|A\|_1 = \left\|A^\top\right\|_\infty$).

Problem 9.11. Show that the map $A \mapsto \rho(A)$ (where $\rho(A)$ is the spectral radius of A) is neither a norm nor a matrix norm. In particular, find two 2×2 matrices A and B such that

$$\rho(A + B) > \rho(A) + \rho(B) = 0 \quad \text{and} \quad \rho(AB) > \rho(A)\rho(B) = 0.$$

Problem 9.12. Define the map $A \mapsto M(A)$ (defined on $n \times n$ real or complex $n \times n$ matrices) by

$$M(A) = \max\{|a_{ij}| \mid 1 \leq i, j \leq n\}.$$

(1) Prove that

$$M(AB) \leq nM(A)M(B)$$

for all $n \times n$ matrices A and B.

(2) Give a counter-example of the inequality

$$M(AB) \leq M(A)M(B).$$

(3) Prove that the map $A \mapsto \|A\|_M$ given by

$$\|A\|_M = nM(A) = n \max\{|a_{ij}| \mid 1 \leq i, j \leq n\}$$

is a matrix norm.

Problem 9.13. Let S be a real symmetric positive definite matrix.

(1) Use the Cholesky factorization to prove that there is some upper-triangular matrix C, unique if its diagonal elements are strictly positive, such that $S = C^\top C$.

(2) For any $x \in \mathbb{R}^n$, define

$$\|x\|_S = (x^\top S x)^{1/2}.$$

Prove that

$$\|x\|_S = \|Cx\|_2,$$

and that the map $x \mapsto \|x\|_S$ is a norm.

Problem 9.14. Let A be a real 2×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

(1) Prove that the squares of the singular values $\sigma_1 \geq \sigma_2$ of A are the roots of the quadratic equation
$$X^2 - \mathrm{tr}(A^\top A)X + |\det(A)|^2 = 0.$$

(2) If we let
$$\mu(A) = \frac{a_{1\,1}^2 + a_{1\,2}^2 + a_{2\,1}^2 + a_{2\,2}^2}{2|a_{1\,1}a_{2\,2} - a_{1\,2}a_{2\,1}|},$$
prove that
$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_2} = \mu(A) + (\mu(A)^2 - 1)^{1/2}.$$

(3) Consider the subset $\mathcal{S}$ of 2×2 invertible matrices whose entries $a_{i\,j}$ are integers such that $0 \leq a_{ij} \leq 100$.

Prove that the functions $\mathrm{cond}_2(A)$ and $\mu(A)$ reach a maximum on the set $\mathcal{S}$ for the same values of A.

Check that for the matrix
$$A_m = \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix}$$
we have
$$\mu(A_m) = 19{,}603 \quad \det(A_m) = -1$$
and
$$\mathrm{cond}_2(A_m) \approx 39{,}206.$$

(4) Prove that for all $A \in \mathcal{S}$, if $|\det(A)| \geq 2$ then $\mu(A) \leq 10{,}000$. Conclude that the maximum of $\mu(A)$ on $\mathcal{S}$ is achieved for matrices such that $\det(A) = \pm 1$. Prove that finding matrices that maximize μ on $\mathcal{S}$ is equivalent to finding some integers n_1, n_2, n_3, n_4 such that
$$\begin{aligned}
&0 \leq n_4 \leq n_3 \leq n_2 \leq n_1 \leq 100 \\
&n_1^2 + n_2^2 + n_3^2 + n_4^2 \geq 100^2 + 99^2 + 99^2 + 98^2 = 39{,}206 \\
&|n_1 n_4 - n_2 n_3| = 1.
\end{aligned}$$

You may use without proof that the fact that the only solution to the above constraints is the multiset
$$\{100, 99, 99, 98\}.$$

(5) Deduce from part (4) that the matrices in $\mathcal{S}$ for which μ has a maximum value are
$$A_m = \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix} \quad \begin{pmatrix} 98 & 99 \\ 99 & 100 \end{pmatrix} \quad \begin{pmatrix} 99 & 100 \\ 98 & 99 \end{pmatrix} \quad \begin{pmatrix} 99 & 98 \\ 100 & 99 \end{pmatrix}$$
and check that μ has the same value for these matrices. Conclude that
$$\max_{A \in \mathcal{S}} \mathrm{cond}_2(A) = \mathrm{cond}_2(A_m).$$

(6) Solve the system

$$\begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 199 \\ 197 \end{pmatrix}.$$

Perturb the right-hand side b by

$$\Delta b = \begin{pmatrix} -0.0097 \\ 0.0106 \end{pmatrix}$$

and solve the new system

$$A_m y = b + \Delta b$$

where $y = (y_1, y_2)$. Check that

$$\Delta x = y - x = \begin{pmatrix} 2 \\ -2.0203 \end{pmatrix}.$$

Compute $\|x\|_2$, $\|\Delta x\|_2$, $\|b\|_2$, $\|\Delta b\|_2$, and estimate

$$c = \frac{\|\Delta x\|_2}{\|x\|_2} \left(\frac{\|\Delta b\|_2}{\|b\|_2} \right)^{-1}.$$

Check that

$$c \approx \mathrm{cond}_2(A_m) \approx 39,206.$$

Problem 9.15. Consider a real 2×2 matrix with zero trace of the form

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}.$$

(1) Prove that

$$A^2 = (a^2 + bc)I_2 = -\det(A)I_2.$$

If $a^2 + bc = 0$, prove that

$$e^A = I_2 + A.$$

(2) If $a^2 + bc < 0$, let $\omega > 0$ be such that $\omega^2 = -(a^2 + bc)$. Prove that

$$e^A = \cos\omega\, I_2 + \frac{\sin\omega}{\omega} A.$$

(3) If $a^2 + bc > 0$, let $\omega > 0$ be such that $\omega^2 = a^2 + bc$. Prove that

$$e^A = \cosh\omega\, I_2 + \frac{\sinh\omega}{\omega} A.$$

(3) Prove that in all cases

$$\det\left(e^A\right) = 1 \quad \text{and} \quad \mathrm{tr}(A) \geq -2.$$

(4) Prove that there exist some real 2×2 matrix B with $\det(B) = 1$ such that there is no real 2×2 matrix A with zero trace such that $e^A = B$.

Problem 9.16. Recall that the Hilbert matrix is given by

$$H_{ij}^{(n)} = \left(\frac{1}{i+j-1}\right).$$

(1) Prove that

$$\det(H^{(n)}) = \frac{(1!2!\cdots(n-1)!)^4}{1!2!\cdots(2n-1)!},$$

thus the reciprocal of an integer.

Hint. Use Problem 7.13.

(2) Amazingly, the entries of the inverse of $H^{(n)}$ are integers. Prove that $(H^{(n)})^{-1} = (\alpha_{ij})$, with

$$\alpha_{ij} = (-1)^{i+j}(i+j-1)\binom{n+i-1}{n-j}\binom{n+j-1}{n-i}\binom{i+j-2}{i-1}^2.$$

Chapter 10

Iterative Methods for Solving Linear Systems

10.1 Convergence of Sequences of Vectors and Matrices

In Chapter 8 we discussed some of the main methods for solving systems of linear equations. These methods are *direct methods*, in the sense that they yield exact solutions (assuming infinite precision!).

Another class of methods for solving linear systems consists in approximating solutions using *iterative methods*. The basic idea is this: Given a linear system $Ax = b$ (with A a square invertible matrix in $\mathrm{M}_n(\mathbb{C})$), find another matrix $B \in \mathrm{M}_n(\mathbb{C})$ and a vector $c \in \mathbb{C}^n$, such that

1. The matrix $I - B$ is invertible

2. The unique solution $\widetilde{x}$ of the system $Ax = b$ is *identical* to the unique solution $\widetilde{u}$ of the system
$$u = Bu + c,$$

and then starting from any vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N}.$$

Under certain conditions (to be clarified soon), the sequence (u_k) converges to a limit $\widetilde{u}$ which is the unique solution of $u = Bu + c$, and thus of $Ax = b$.

Consequently, it is important to find conditions that ensure the convergence of the above sequences and to have tools to compare the "rate" of convergence of these sequences. Thus, we begin with some general results about the convergence of sequences of vectors and matrices.

Let $(E, \| \|)$ be a normed vector space. Recall from Section 9.7 that a sequence (u_k) of vectors $u_k \in E$ *converges to a limit* $u \in E$, if for every $\epsilon > 0$, there some natural number N such that

$$\|u_k - u\| \leq \epsilon, \quad \textit{for all } \ k \geq N.$$

We write

$$u = \lim_{k \mapsto \infty} u_k.$$

If E is a finite-dimensional vector space and $\dim(E) = n$, we know from Theorem 9.5 that any two norms are equivalent, and if we choose the norm $\| \|_\infty$, we see that the convergence of the sequence of vectors u_k is equivalent to the convergence of the n sequences of scalars formed by the components of these vectors (over any basis). The same property applies to the finite-dimensional vector space $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices (with $K = \mathbb{R}$ or $K = \mathbb{C}$), which means that the convergence of a sequence of matrices $A_k = (a_{ij}^{(k)})$ is equivalent to the convergence of the $m \times n$ sequences of scalars $(a_{ij}^{(k)})$, with i, j fixed $(1 \leq i \leq m,\ 1 \leq j \leq n)$.

The first theorem below gives a necessary and sufficient condition for the sequence (B^k) of powers of a matrix B to converge to the zero matrix. Recall that the spectral radius $\rho(B)$ of a matrix B is the maximum of the moduli $|\lambda_i|$ of the eigenvalues of B.

Theorem 10.1. *For any square matrix B, the following conditions are equivalent:*

(1) $\lim_{k \mapsto \infty} B^k = 0$,

(2) $\lim_{k \mapsto \infty} B^k v = 0$, *for all vectors* v,

(3) $\rho(B) < 1$,

(4) $\|B\| < 1$, *for some subordinate matrix norm* $\| \|$.

Proof. Assume (1) and let $\| \|$ be a vector norm on E and $\| \|$ be the corresponding matrix norm. For every vector $v \in E$, because $\| \|$ is a matrix norm, we have

$$\|B^k v\| \leq \|B^k\| \|v\|,$$

and since $\lim_{k \mapsto \infty} B^k = 0$ means that $\lim_{k \mapsto \infty} \|B^k\| = 0$, we conclude that $\lim_{k \mapsto \infty} \|B^k v\| = 0$, that is, $\lim_{k \mapsto \infty} B^k v = 0$. This proves that (1) implies (2).

Assume (2). If we had $\rho(B) \geq 1$, then there would be some eigenvector $u \, (\neq 0)$ and some eigenvalue λ such that

$$Bu = \lambda u, \quad |\lambda| = \rho(B) \geq 1,$$

but then the sequence $(B^k u)$ would not converge to 0, because $B^k u = \lambda^k u$ and $|\lambda^k| = |\lambda|^k \geq 1$. It follows that (2) implies (3).

Assume that (3) holds, that is, $\rho(B) < 1$. By Proposition 9.12, we can find $\epsilon > 0$ small enough that $\rho(B) + \epsilon < 1$, and a subordinate matrix norm $\| \|$ such that

$$\|B\| \leq \rho(B) + \epsilon,$$

which is (4).

Finally, assume (4). Because $\| \ \|$ is a matrix norm,

$$\|B^k\| \leq \|B\|^k,$$

and since $\|B\| < 1$, we deduce that (1) holds. $\square$

The following proposition is needed to study the rate of convergence of iterative methods.

Proposition 10.2. *For every square matrix $B \in \mathrm{M}_n(\mathbb{C})$ and every matrix norm $\| \ \|$, we have*

$$\lim_{k \mapsto \infty} \|B^k\|^{1/k} = \rho(B).$$

Proof. We know from Proposition 9.6 that $\rho(B) \leq \|B\|$, and since $\rho(B) = (\rho(B^k))^{1/k}$, we deduce that

$$\rho(B) \leq \|B^k\|^{1/k} \quad \text{for all } k \geq 1.$$

Now let us prove that for every $\epsilon > 0$, there is some integer $N(\epsilon)$ such that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon \quad \text{for all } k \geq N(\epsilon).$$

Together with the fact that

$$\rho(B) \leq \|B^k\|^{1/k} \quad \text{for all } k \geq 1,$$

we deduce that $\lim_{k \mapsto \infty} \|B^k\|^{1/k}$ exists and that

$$\lim_{k \mapsto \infty} \|B^k\|^{1/k} = \rho(B).$$

For any given $\epsilon > 0$, let B_ϵ be the matrix

$$B_\epsilon = \frac{B}{\rho(B) + \epsilon}.$$

Since $\rho(B_\epsilon) < 1$, Theorem 10.1 implies that $\lim_{k \mapsto \infty} B_\epsilon^k = 0$. Consequently, there is some integer $N(\epsilon)$ such that for all $k \geq N(\epsilon)$, we have

$$\|B_\epsilon^k\| = \frac{\|B^k\|}{(\rho(B) + \epsilon)^k} \leq 1,$$

which implies that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon,$$

as claimed. $\square$

We now apply the above results to the convergence of iterative methods.

10.2 Convergence of Iterative Methods

Recall that iterative methods for solving a linear system $Ax = b$ (with $A \in \mathrm{M}_n(\mathbb{C})$ invertible) consists in finding some matrix B and some vector c, such that $I - B$ is invertible, and the unique solution $\widetilde{x}$ of $Ax = b$ is equal to the unique solution $\widetilde{u}$ of $u = Bu + c$. Then starting from *any* vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

and say that the iterative method is *convergent* iff

$$\lim_{k \mapsto \infty} u_k = \widetilde{u},$$

for *every* initial vector u_0.

Here is a fundamental criterion for the convergence of any iterative methods based on a matrix B, called the *matrix of the iterative method*.

Theorem 10.3. *Given a system $u = Bu + c$ as above, where $I - B$ is invertible, the following statements are equivalent:*

(1) The iterative method is convergent.

(2) $\rho(B) < 1$.

(3) $\|B\| < 1$, for some subordinate matrix norm $\| \ \|$.

Proof. Define the vector e_k (*error vector*) by

$$e_k = u_k - \widetilde{u},$$

where $\widetilde{u}$ is the unique solution of the system $u = Bu + c$. Clearly, the iterative method is convergent iff

$$\lim_{k \mapsto \infty} e_k = 0.$$

We claim that

$$e_k = B^k e_0, \quad k \geq 0,$$

where $e_0 = u_0 - \widetilde{u}$.

This is proven by induction on k. The base case $k = 0$ is trivial. By the induction hypothesis, $e_k = B^k e_0$, and since $u_{k+1} = Bu_k + c$, we get

$$u_{k+1} - \widetilde{u} = Bu_k + c - \widetilde{u},$$

and because $\widetilde{u} = B\widetilde{u} + c$ and $e_k = B^k e_0$ (by the induction hypothesis), we obtain

$$u_{k+1} - \widetilde{u} = Bu_k - B\widetilde{u} = B(u_k - \widetilde{u}) = Be_k = BB^k e_0 = B^{k+1} e_0,$$

proving the induction step. Thus, the iterative method converges iff

$$\lim_{k \mapsto \infty} B^k e_0 = 0.$$

Consequently, our theorem follows by Theorem 10.1. $\square$

The next proposition is needed to compare the rate of convergence of iterative methods. It shows that *asymptotically, the error vector* $e_k = B^k e_0$ *behaves at worst like* $(\rho(B))^k$.

Proposition 10.4. *Let* $\| \|$ *be any vector norm, let* $B \in \mathrm{M}_n(\mathbb{C})$ *be a matrix such that* $I - B$ *is invertible, and let* $\widetilde{u}$ *be the unique solution of* $u = Bu + c$.

(1) If (u_k) *is any sequence defined iteratively by*

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

then

$$\lim_{k \mapsto \infty} \left[\sup_{\|u_0 - \widetilde{u}\| = 1} \|u_k - \widetilde{u}\|^{1/k} \right] = \rho(B).$$

(2) Let B_1 *and* B_2 *be two matrices such that* $I - B_1$ *and* $I - B_2$ *are invertible, assume that both* $u = B_1 u + c_1$ *and* $u = B_2 u + c_2$ *have the same unique solution* $\widetilde{u}$, *and consider any two sequences* (u_k) *and* (v_k) *defined inductively by*

$$\begin{aligned} u_{k+1} &= B_1 u_k + c_1 \\ v_{k+1} &= B_2 v_k + c_2, \end{aligned}$$

with $u_0 = v_0$. *If* $\rho(B_1) < \rho(B_2)$, *then for any* $\epsilon > 0$, *there is some integer* $N(\epsilon)$, *such that for all* $k \geq N(\epsilon)$, *we have*

$$\sup_{\|u_0 - \widetilde{u}\| = 1} \left[\frac{\|v_k - \widetilde{u}\|}{\|u_k - \widetilde{u}\|} \right]^{1/k} \geq \frac{\rho(B_2)}{\rho(B_1) + \epsilon}.$$

Proof. Let $\| \|$ be the subordinate matrix norm. Recall that

$$u_k - \widetilde{u} = B^k e_0,$$

with $e_0 = u_0 - \widetilde{u}$. For every $k \in \mathbb{N}$, we have

$$(\rho(B))^k = \rho(B^k) \leq \|B^k\| = \sup_{\|e_0\| = 1} \|B^k e_0\|,$$

which implies

$$\rho(B) = \sup_{\|e_0\| = 1} \|B^k e_0\|^{1/k} = \|B^k\|^{1/k},$$

and Statement (1) follows from Proposition 10.2.

Because $u_0 = v_0$, we have

$$\begin{aligned} u_k - \widetilde{u} &= B_1^k e_0 \\ v_k - \widetilde{u} &= B_2^k e_0, \end{aligned}$$

with $e_0 = u_0 - \widetilde{u} = v_0 - \widetilde{u}$. Again, by Proposition 10.2, for every $\epsilon > 0$, there is some natural number $N(\epsilon)$ such that if $k \geq N(\epsilon)$, then

$$\sup_{\|e_0\|=1} \|B_1^k e_0\|^{1/k} \leq \rho(B_1) + \epsilon.$$

Furthermore, for all $k \geq N(\epsilon)$, there exists a vector $e_0 = e_0(k)$ (for some suitable choice of u_0) such that

$$\|e_0\| = 1 \quad \text{and} \quad \|B_2^k e_0\|^{1/k} = \|B_2^k\|^{1/k} \geq \rho(B_2),$$

which implies Statement (2). $\square$

In light of the above, we see that when we investigate new iterative methods, we have to deal with the following two problems:

1. Given an iterative method with matrix B, determine whether the method is convergent. This involves determining whether $\rho(B) < 1$, or equivalently whether there is a subordinate matrix norm such that $\|B\| < 1$. By Proposition 9.11, this implies that $I - B$ is invertible (since $\| - B\| = \|B\|$, Proposition 9.11 applies).

2. Given two convergent iterative methods, compare them. The iterative method which is faster is that whose matrix has the smaller spectral radius.

We now discuss three iterative methods for solving linear systems:

1. Jacobi's method

2. Gauss–Seidel's method

3. The relaxation method.

10.3 Description of the Methods of Jacobi, Gauss–Seidel, and Relaxation

The methods described in this section are instances of the following scheme: Given a linear system $Ax = b$, with A invertible, suppose we can write A in the form

$$A = M - N,$$

with M invertible, and "easy to invert," which means that M is close to being a diagonal or a triangular matrix (perhaps by blocks). Then $Au = b$ is equivalent to

$$Mu = Nu + b,$$

that is,

$$u = M^{-1}Nu + M^{-1}b.$$

Therefore, we are in the situation described in the previous sections with $B = M^{-1}N$ and $c = M^{-1}b$. In fact, since $A = M - N$, we have

$$B = M^{-1}N = M^{-1}(M - A) = I - M^{-1}A, \tag{$*$}$$

which shows that $I - B = M^{-1}A$ is invertible. The iterative method associated with the matrix $B = M^{-1}N$ is given by

$$u_{k+1} = M^{-1}Nu_k + M^{-1}b, \quad k \geq 0, \tag{$\dagger$}$$

starting from any arbitrary vector u_0. From a practical point of view, we do not invert M, and instead we solve iteratively the systems

$$Mu_{k+1} = Nu_k + b, \quad k \geq 0.$$

Various methods correspond to various ways of choosing M and N from A. The first two methods choose M and N as disjoint submatrices of A, but the relaxation method allows some overlapping of M and N.

To describe the various choices of M and N, it is convenient to write A in terms of three submatrices D, E, F, as

$$A = D - E - F,$$

where the only nonzero entries in D are the diagonal entries in A, the only nonzero entries in E are the negatives of nonzero entries in A below the the diagonal, and the only nonzero entries in F are the negatives of nonzero entries in A above the diagonal. More explicitly, if

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix},$$

then

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{22} & 0 & \cdots & 0 & 0 \\ 0 & 0 & a_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & a_{n\,n} \end{pmatrix},$$

$$-E = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \ddots & 0 & 0 \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & 0 \end{pmatrix}, \ -F = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ 0 & 0 & 0 & \ddots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

In *Jacobi's method*, we assume that *all* diagonal entries in A are nonzero, and we pick

$$\begin{aligned} M &= D \\ N &= E + F, \end{aligned}$$

so that by $(*)$,

$$B = M^{-1}N = D^{-1}(E+F) = I - D^{-1}A.$$

As a matter of notation, we let

$$J = I - D^{-1}A = D^{-1}(E+F),$$

which is called *Jacobi's matrix*. The corresponding method, *Jacobi's iterative method*, computes the sequence (u_k) using the recurrence

$$u_{k+1} = D^{-1}(E+F)u_k + D^{-1}b, \quad k \geq 0.$$

In practice, we iteratively solve the systems

$$Du_{k+1} = (E+F)u_k + b, \quad k \geq 0.$$

If we write $u_k = (u_1^k, \ldots, u_n^k)$, we solve iteratively the following system:

$$\begin{array}{cccccccc} a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & -a_{1n}u_n^k & +\,b_1 \\ a_{22}u_2^{k+1} & = & -a_{21}u_1^k & & -a_{23}u_3^k & \cdots & -a_{2n}u_n^k & +\,b_2 \\ \vdots & \vdots & \vdots & & & & & \\ a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^k & \cdots & -a_{n-1\,n-2}u_{n-2}^k & & -a_{n-1\,n}u_n^k & +\,b_{n-1} \\ a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^k & -a_{n\,2}u_2^k & \cdots & -a_{n\,n-1}u_{n-1}^k & & +\,b_n \end{array}.$$

In `Matlab` one step of Jacobi iteration is achieved by the following function:

```
function v = Jacobi2(A,b,u)
   n = size(A,1);
   v = zeros(n,1);
   for i = 1:n
      v(i,1)  = u(i,1) + (-A(i,:)*u + b(i))/A(i,i);
   end
end
```

In order to run m iteration steps, run the following function:

```
function u = jacobi(A,b,u0,m)
  u = u0;
  for j = 1:m
    u = Jacobi2(A,b,u);
  end
end
```

Example 10.1. Consider the linear system

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 25 \\ -24 \\ 21 \\ -15 \end{pmatrix}.$$

We check immediately that the solution is

$$x_1 = 11,\ x_2 = -3,\ x_3 = 7,\ x_4 = -4.$$

It is easy to see that the Jacobi matrix is

$$J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

After 10 Jacobi iterations, we find the approximate solution

$$x_1 = 10.2588,\ x_2 = -2.5244,\ x_3 = 5.8008,\ x_4 = -3.7061.$$

After 20 iterations, we find the approximate solution

$$x_1 = 10.9110,\ x_2 = -2.9429,\ x_3 = 6.8560,\ x_4 = -3.9647.$$

After 50 iterations, we find the approximate solution

$$x_1 = 10.9998,\ x_2 = -2.9999,\ x_3 = 6.9998,\ x_4 = -3.9999,$$

and After 60 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals.

It can be shown (see Problem 10.6) that the eigenvalues of J are

$$\cos\left(\frac{\pi}{5}\right),\ \cos\left(\frac{2\pi}{5}\right),\ \cos\left(\frac{3\pi}{5}\right),\ \cos\left(\frac{4\pi}{5}\right),$$

so the spectral radius of $J = B$ is

$$\rho(J) = \cos\left(\frac{\pi}{5}\right) = 0.8090 < 1.$$

By Theorem 10.3, Jacobi's method converges for the matrix of this example.

Observe that we can try to "speed up" the method by using the new value u_1^{k+1} instead of u_1^k in solving for u_2^{k+2} using the second equations, and more generally, use $u_1^{k+1}, \ldots, u_{i-1}^{k+1}$ instead of $u_1^k, \ldots, u_{i-1}^k$ in solving for u_i^{k+1} in the ith equation. This observation leads to the system

$$\begin{array}{rclllllll}
a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & & -a_{1n}u_n^k & +\, b_1 \\
a_{22}u_2^{k+1} & = & -a_{21}u_1^{k+1} & & -a_{23}u_3^k & \cdots & & -a_{2n}u_n^k & +\, b_2 \\
\vdots & \vdots & \vdots & & & & & & \\
a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^{k+1} & \cdots & -a_{n-1\,n-2}u_{n-2}^{k+1} & & & -a_{n-1\,n}u_n^k & +\, b_{n-1} \\
a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^{k+1} & -a_{n\,2}u_2^{k+1} & \cdots & & -a_{n\,n-1}u_{n-1}^{k+1} & & +\, b_n,
\end{array}$$

which, in matrix form, is written

$$Du_{k+1} = Eu_{k+1} + Fu_k + b.$$

Because D is invertible and E is lower triangular, the matrix $D - E$ is invertible, so the above equation is equivalent to

$$u_{k+1} = (D - E)^{-1}Fu_k + (D - E)^{-1}b, \quad k \geq 0.$$

The above corresponds to choosing M and N to be

$$\begin{aligned} M &= D - E \\ N &= F, \end{aligned}$$

and the matrix B is given by

$$B = M^{-1}N = (D - E)^{-1}F.$$

Since $M = D - E$ is invertible, we know that $I - B = M^{-1}A$ is also invertible.

The method that we just described is the *iterative method of Gauss–Seidel*, and the matrix B is called the *matrix of Gauss–Seidel* and denoted by $\mathcal{L}_1$, with

$$\mathcal{L}_1 = (D - E)^{-1}F.$$

One of the advantages of the method of Gauss–Seidel is that is requires only half of the memory used by Jacobi's method, since we only need

$$u_1^{k+1}, \ldots, u_{i-1}^{k+1}, u_{i+1}^k, \ldots, u_n^k$$

to compute u_i^{k+1}. We also show that in certain important cases (for example, if A is a tridiagonal matrix), the method of Gauss–Seidel converges faster than Jacobi's method (in this case, they both converge or diverge simultaneously).

In `Matlab` one step of Gauss–Seidel iteration is achieved by the following function:

```
function u = GaussSeidel3(A,b,u)
  n = size(A,1);
  for i = 1:n
     u(i,1)  = u(i,1) + (-A(i,:)*u + b(i))/A(i,i);
  end
end
```

It is remarkable that the only difference with `Jacobi2` is that the same variable u is used on both sides of the assignment. In order to run m iteration steps, run the following function:

```
function u = GaussSeidel1(A,b,u0,m)
  u = u0;
  for j = 1:m
    u = GaussSeidel3(A,b,u);
  end
end
```

Example 10.2. Consider the same linear system

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 25 \\ -24 \\ 21 \\ -15 \end{pmatrix}$$

as in Example 10.1, whose solution is

$$x_1 = 11,\ x_2 = -3,\ x_3 = 7,\ x_4 = -4.$$

After 10 Gauss–Seidel iterations, we find the approximate solution

$$x_1 = 10.9966,\ x_2 = -3.0044,\ x_3 = 6.9964,\ x_4 = -4.0018.$$

After 20 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0001,\ x_3 = 6.9999,\ x_4 = -4.0000.$$

After 25 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals. We observe that for this example, Gauss–Seidel's method converges about twice as fast as Jacobi's method. It will be shown in Proposition 10.8 that for a tridiagonal matrix, the spectral radius of the Gauss–Seidel matrix $\mathcal{L}_1$ is given by

$$\rho(\mathcal{L}_1) = (\rho(J))^2,$$

so our observation is consistent with the theory.

The new ingredient in the *relaxation method* is to incorporate part of the matrix D into N: we define M and N by

$$M = \frac{D}{\omega} - E$$

$$N = \frac{1-\omega}{\omega} D + F,$$

where $\omega \neq 0$ is a real parameter to be suitably chosen. Actually, we show in Section 10.4 that for the relaxation method to converge, we must have $\omega \in (0, 2)$. Note that the case $\omega = 1$ corresponds to the method of Gauss–Seidel.

If we assume that *all* diagonal entries of D are nonzero, the matrix M is invertible. The matrix B is denoted by $\mathcal{L}_\omega$ and called the *matrix of relaxation*, with

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1} \left(\frac{1-\omega}{\omega} D + F\right) = (D - \omega E)^{-1}((1-\omega)D + \omega F).$$

The number ω is called the *parameter of relaxation*.

When $\omega > 1$, the relaxation method is known as *successive overrelaxation*, abbreviated as *SOR*.

At first glance the relaxation matrix $\mathcal{L}_\omega$ seems a lot more complicated than the Gauss–Seidel matrix $\mathcal{L}_1$, but the iterative system associated with the relaxation method is very similar to the method of Gauss–Seidel, and is quite simple. Indeed, the system associated with the relaxation method is given by

$$\left(\frac{D}{\omega} - E\right) u_{k+1} = \left(\frac{1-\omega}{\omega} D + F\right) u_k + b,$$

which is equivalent to

$$(D - \omega E) u_{k+1} = ((1-\omega)D + \omega F) u_k + \omega b,$$

and can be written

$$D u_{k+1} = D u_k - \omega (D u_k - E u_{k+1} - F u_k - b).$$

Explicitly, this is the system

$$\begin{aligned}
a_{11} u_1^{k+1} &= a_{11} u_1^k - \omega(a_{11} u_1^k + a_{12} u_2^k + a_{13} u_3^k + \cdots + a_{1n-2} u_{n-2}^k + a_{1n-1} u_{n-1}^k + a_{1n} u_n^k - b_1) \\
a_{22} u_2^{k+1} &= a_{22} u_2^k - \omega(a_{21} u_1^{k+1} + a_{22} u_2^k + a_{23} u_3^k + \cdots + a_{2n-2} u_{n-2}^k + a_{2n-1} u_{n-1}^k + a_{2n} u_n^k - b_2) \\
&\vdots \\
a_{n\,n} u_n^{k+1} &= a_{n\,n} u_n^k - \omega(a_{n\,1} u_1^{k+1} + a_{n\,2} u_2^{k+1} + \cdots + a_{n\,n-2} u_{n-2}^{k+1} + a_{n\,n-1} u_{n-1}^{k+1} + a_{n\,n} u_n^k - b_n).
\end{aligned}$$

In `Matlab` one step of relaxation iteration is achieved by the following function:

```
function u = relax3(A,b,u,omega)
  n = size(A,1);
  for i = 1:n
     u(i,1)  = u(i,1) + omega*(-A(i,:)*u + b(i))/A(i,i);
  end
end
```

Observe that function `relax3` is obtained from the function `GaussSeidel3` by simply inserting ω in front of the expression $(-A(i,:)*u+b(i))/A(i,i)$. In order to run m iteration steps, run the following function:

```
function u = relax(A,b,u0,omega,m)
  u = u0;
  for j = 1:m
    u = relax3(A,b,u,omega);
  end
end
```

Example 10.3. Consider the same linear system as in Examples 10.1 and 10.2, whose solution is

$$x_1 = 11,\ x_2 = -3,\ x_3 = 7,\ x_4 = -4.$$

After 10 relaxation iterations with $\omega = 1.1$, we find the approximate solution

$$x_1 = 11.0026,\ x_2 = -2.9968,\ x_3 = 7.0024,\ x_4 = -3.9989.$$

After 10 iterations with $\omega = 1.2$, we find the approximate solution

$$x_1 = 11.0014,\ x_2 = -2.9985,\ x_3 = 7.0010,\ x_4 = -3.9996.$$

After 10 iterations with $\omega = 1.3$, we find the approximate solution

$$x_1 = 10.9996,\ x_2 = -3.0001,\ x_3 = 6.9999,\ x_4 = -4.0000.$$

After 10 iterations with $\omega = 1.27$, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals. We observe that for this example the method of relaxation with $\omega = 1.27$ converges faster than the method of Gauss–Seidel. This observation will be confirmed by Proposition 10.10.

What remains to be done is to find conditions that ensure the convergence of the relaxation method (and the Gauss–Seidel method), that is:

1. Find conditions on ω, namely some interval $I \subseteq \mathbb{R}$ so that $\omega \in I$ implies $\rho(\mathcal{L}_\omega) < 1$; we will prove that $\omega \in (0, 2)$ is a necessary condition.

2. Find if there exist some *optimal value* ω_0 of $\omega \in I$, so that

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{\omega \in I} \rho(\mathcal{L}_\omega).$$

We will give partial answers to the above questions in the next section.

It is also possible to extend the methods of this section by using *block decompositions* of the form $A = D - E - F$, where D, E, and F consist of blocks, and D is an invertible block-diagonal matrix. See Figure 10.1.

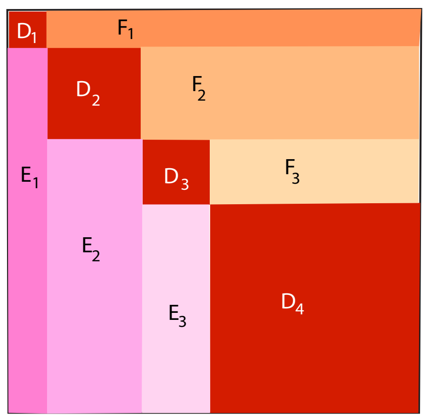


Figure 10.1: A schematic representation of a block decomposition $A = D - E - F$, where $D = \cup_{i=1}^4 D_i$, $E = \cup_{i=1}^3 E_i$, and $F = \cup_{i=1}^3 F_i$.

10.4 Convergence of the Methods of Gauss–Seidel and Relaxation

We begin with a general criterion for the convergence of an iterative method associated with a (complex) Hermitian positive definite matrix, $A = M - N$. Next we apply this result to the relaxation method.

Proposition 10.5. *Let A be any Hermitian positive definite matrix, written as*

$$A = M - N,$$

with M invertible. Then $M^* + N$ is Hermitian, and if it is positive definite, then

$$\rho(M^{-1}N) < 1,$$

so that the iterative method converges.

Proof. Since $M = A + N$ and A is Hermitian, $A^* = A$, so we get

$$M^* + N = A^* + N^* + N = A + N + N^* = M + N^* = (M^* + N)^*,$$

which shows that $M^* + N$ is indeed Hermitian.

Because A is Hermitian positive definite, the function

$$v \mapsto (v^* A v)^{1/2}$$

from $\mathbb{C}^n$ to $\mathbb{R}$ is a vector norm $\| \|$, and let $\| \|$ also denote its subordinate matrix norm. We prove that

$$\|M^{-1}N\| < 1,$$

which by Theorem 10.1 proves that $\rho(M^{-1}N) < 1$. By definition

$$\|M^{-1}N\| = \|I - M^{-1}A\| = \sup_{\|v\|=1} \|v - M^{-1}Av\|,$$

which leads us to evaluate $\|v - M^{-1}Av\|$ when $\|v\| = 1$. If we write $w = M^{-1}Av$, using the facts that $\|v\| = 1$, $v = A^{-1}Mw$, $A^* = A$, and $A = M - N$, we have

$$\begin{aligned}
\|v - w\|^2 &= (v - w)^* A (v - w) \\
&= \|v\|^2 - v^* A w - w^* A v + w^* A w \\
&= 1 - w^* M^* w - w^* M w + w^* A w \\
&= 1 - w^* (M^* + N) w.
\end{aligned}$$

Now since we assumed that $M^* + N$ is positive definite, if $w \neq 0$, then $w^*(M^* + N)w > 0$, and we conclude that

$$\text{if} \quad \|v\| = 1, \quad \text{then} \quad \|v - M^{-1}Av\| < 1.$$

Finally, the function

$$v \mapsto \|v - M^{-1}Av\|$$

is continuous as a composition of continuous functions, therefore it achieves its maximum on the compact subset $\{v \in \mathbb{C}^n \mid \|v\| = 1\}$, which proves that

$$\sup_{\|v\|=1} \|v - M^{-1}Av\| < 1,$$

and completes the proof. $\square$

Now as in the previous sections, we assume that A is written as $A = D - E - F$, with D invertible, possibly in block form. The next theorem provides a sufficient condition (which turns out to be also necessary) for the relaxation method to converge (and thus, for the method of Gauss–Seidel to converge). This theorem is known as the *Ostrowski-Reich theorem*.

Theorem 10.6. *If $A = D - E - F$ is Hermitian positive definite, and if $0 < \omega < 2$, then the relaxation method converges. This also holds for a block decomposition of A.*

Proof. Recall that for the relaxation method, $A = M - N$ with

$$M = \frac{D}{\omega} - E$$

$$N = \frac{1-\omega}{\omega} D + F,$$

and because $D^* = D$, $E^* = F$ (since A is Hermitian) and $\omega \neq 0$ is real, we have

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega} D + F = \frac{2-\omega}{\omega} D.$$

If D consists of the diagonal entries of A, then we know from Section 8.8 that these entries are all positive, and since $\omega \in (0, 2)$, we see that the matrix $((2-\omega)/\omega)D$ is positive definite. If D consists of diagonal blocks of A, because A is positive, definite, by choosing vectors z obtained by picking a nonzero vector for each block of D and padding with zeros, we see that each block of D is positive definite, and thus D itself is positive definite. Therefore, in all cases, $M^* + N$ is positive definite, and we conclude by using Proposition 10.5. □

Remark: What if we allow the parameter ω to be a nonzero complex number $\omega \in \mathbb{C}$? In this case, we get

$$M^* + N = \frac{D^*}{\overline{\omega}} - E^* + \frac{1-\omega}{\omega} D + F = \left(\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1\right) D.$$

But,

$$\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1 = \frac{\omega + \overline{\omega} - \omega\overline{\omega}}{\omega\overline{\omega}} = \frac{1 - (\omega - 1)(\overline{\omega} - 1)}{|\omega|^2} = \frac{1 - |\omega - 1|^2}{|\omega|^2},$$

so the relaxation method also converges for $\omega \in \mathbb{C}$, provided that

$$|\omega - 1| < 1.$$

This condition reduces to $0 < \omega < 2$ if ω is real.

Unfortunately, Theorem 10.6 does not apply to Jacobi's method, but in special cases, Proposition 10.5 can be used to prove its convergence. On the positive side, if a matrix is strictly column (or row) diagonally dominant, then it can be shown that the method of Jacobi and the method of Gauss–Seidel both converge. The relaxation method also converges if $\omega \in (0, 1]$, but this is not a very useful result because the speed-up of convergence usually occurs for $\omega > 1$.

We now prove that, without *any* assumption on $A = D - E - F$, other than the fact that A and D are invertible, in order for the relaxation method to converge, we must have $\omega \in (0, 2)$.

Proposition 10.7. *Given any matrix $A = D - E - F$, with A and D invertible, for any $\omega \neq 0$, we have*

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|,$$

where $\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$. Therefore, the relaxation method (possibly by blocks) does not converge unless $\omega \in (0, 2)$. If we allow ω to be complex, then we must have

$$|\omega - 1| < 1$$

for the relaxation method to converge.

Proof. Observe that the product $\lambda_1 \cdots \lambda_n$ of the eigenvalues of $\mathcal{L}_\omega$, which is equal to $\det(\mathcal{L}_\omega)$, is given by

$$\lambda_1 \cdots \lambda_n = \det(\mathcal{L}_\omega) = \frac{\det\left(\dfrac{1-\omega}{\omega}D + F\right)}{\det\left(\dfrac{D}{\omega} - E\right)} = (1-\omega)^n.$$

It follows that

$$\rho(\mathcal{L}_\omega) \geq |\lambda_1 \cdots \lambda_n|^{1/n} = |\omega - 1|.$$

The proof is the same if $\omega \in \mathbb{C}$. ☐

10.5 Convergence of the Methods of Jacobi, Gauss–Seidel, and Relaxation for Tridiagonal Matrices

We now consider the case where A is a *tridiagonal matrix*, possibly by blocks. In this case, we obtain precise results about the spectral radius of J and $\mathcal{L}_\omega$, and as a consequence, about the convergence of these methods. We also obtain some information about the rate of convergence of these methods. We begin with the case $\omega = 1$, which is technically easier to deal with. The following proposition gives us the precise relationship between the spectral radii $\rho(J)$ and $\rho(\mathcal{L}_1)$ of the Jacobi matrix and the Gauss–Seidel matrix.

Proposition 10.8. *Let A be a tridiagonal matrix (possibly by blocks). If $\rho(J)$ is the spectral radius of the Jacobi matrix and $\rho(\mathcal{L}_1)$ is the spectral radius of the Gauss–Seidel matrix, then we have*

$$\rho(\mathcal{L}_1) = (\rho(J))^2.$$

Consequently, the method of Jacobi and the method of Gauss–Seidel both converge or both diverge simultaneously (even when A is tridiagonal by blocks); when they converge, the method of Gauss–Seidel converges faster than Jacobi's method.

Proof. We begin with a preliminary result. Let $A(\mu)$ be a tridiagonal matrix by block of the form

$$A(\mu) = \begin{pmatrix} A_1 & \mu^{-1}C_1 & 0 & 0 & \cdots & 0 \\ \mu B_1 & A_2 & \mu^{-1}C_2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mu B_{p-2} & A_{p-1} & \mu^{-1}C_{p-1} \\ 0 & \cdots & \cdots & 0 & \mu B_{p-1} & A_p \end{pmatrix},$$

then

$$\det(A(\mu)) = \det(A(1)), \quad \mu \neq 0.$$

To prove this fact, form the block diagonal matrix

$$P(\mu) = \mathrm{diag}(\mu I_1, \mu^2 I_2, \ldots, \mu^p I_p),$$

where I_j is the identity matrix of the same dimension as the block A_j. Then it is easy to see that

$$A(\mu) = P(\mu)A(1)P(\mu)^{-1},$$

and thus,

$$\det(A(\mu)) = \det(P(\mu)A(1)P(\mu)^{-1}) = \det(A(1)).$$

Since the Jacobi matrix is $J = D^{-1}(E+F)$, the eigenvalues of J are the zeros of the characteristic polynomial

$$p_J(\lambda) = \det(\lambda I - D^{-1}(E+F)),$$

and thus, they are also the zeros of the polynomial

$$q_J(\lambda) = \det(\lambda D - E - F) = \det(D)p_J(\lambda).$$

Similarly, since the Gauss–Seidel matrix is $\mathcal{L}_1 = (D-E)^{-1}F$, the zeros of the characteristic polynomial

$$p_{\mathcal{L}_1}(\lambda) = \det(\lambda I - (D-E)^{-1}F)$$

are also the zeros of the polynomial

$$q_{\mathcal{L}_1}(\lambda) = \det(\lambda D - \lambda E - F) = \det(D-E)p_{\mathcal{L}_1}(\lambda).$$

Since $A = D - E - F$ is tridiagonal (or tridiagonal by blocks), $\lambda^2 D - \lambda^2 E - F$ is also tridiagonal (or tridiagonal by blocks), and by using our preliminary result with $\mu = \lambda \neq 0$ starting with the matrix $\lambda^2 D - \lambda E - \lambda F$, we get

$$\lambda^n q_J(\lambda) = \det(\lambda^2 D - \lambda E - \lambda F) = \det(\lambda^2 D - \lambda^2 E - F) = q_{\mathcal{L}_1}(\lambda^2).$$

By continuity, the above equation also holds for $\lambda = 0$. But then we deduce that:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_1$, then $\beta^{1/2}$ and $-\beta^{1/2}$ are both eigenvalues of J, where $\beta^{1/2}$ is one of the complex square roots of β.

2. For any $\alpha \neq 0$, α is an eigenvalues of J iff $-\alpha$ is an eigenvalues of J, and if α is an eigenvalues of J, then α^2 is an eigenvalue of $\mathcal{L}_1$.

The above immediately implies that $\rho(\mathcal{L}_1) = (\rho(J))^2$. $\square$

We now consider the more general situation where ω is any real in $(0, 2)$.

Proposition 10.9. *Let A be a tridiagonal matrix (possibly by blocks), and assume that the eigenvalues of the Jacobi matrix are all real. If $\omega \in (0, 2)$, then the method of Jacobi and the method of relaxation both converge or both diverge simultaneously (even when A is tridiagonal by blocks). When they converge, the function $\omega \mapsto \rho(\mathcal{L}_\omega)$ (for $\omega \in (0, 2)$) has a unique minimum equal to $\omega_0 - 1$ for*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

where $1 < \omega_0 < 2$ if $\rho(J) > 0$.

Proof. The proof is very technical and can be found in Serre [156] and Ciarlet [41]. As in the proof of the previous proposition, we begin by showing that the eigenvalues of the matrix $\mathcal{L}_\omega$ are the zeros of the polynomial

$$q_{\mathcal{L}_\omega}(\lambda) = \det\left(\frac{\lambda + \omega - 1}{\omega} D - \lambda E - F\right) = \det\left(\frac{D}{\omega} - E\right) p_{\mathcal{L}_\omega}(\lambda),$$

where $p_{\mathcal{L}_\omega}(\lambda)$ is the characteristic polynomial of $\mathcal{L}_\omega$. Then using the preliminary fact from Proposition 10.8, it is easy to show that

$$q_{\mathcal{L}_\omega}(\lambda^2) = \lambda^n q_J\left(\frac{\lambda^2 + \omega - 1}{\lambda\omega}\right),$$

for all $\lambda \in \mathbb{C}$, with $\lambda \neq 0$. This time we cannot extend the above equation to $\lambda = 0$. This leads us to consider the equation

$$\frac{\lambda^2 + \omega - 1}{\lambda\omega} = \alpha,$$

which is equivalent to

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0,$$

for all $\lambda \neq 0$. Since $\lambda \neq 0$, the above equivalence does not hold for $\omega = 1$, but this is not a problem since the case $\omega = 1$ has already been considered in the previous proposition. Then we can show the following:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_\omega$, then
$$\frac{\beta + \omega - 1}{\beta^{1/2}\omega}, \quad -\frac{\beta + \omega - 1}{\beta^{1/2}\omega}$$
are eigenvalues of J.

2. For every $\alpha \neq 0$, if α and $-\alpha$ are eigenvalues of J, then $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are eigenvalues of $\mathcal{L}_\omega$, where $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are the squares of the roots of the equation
$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$

It follows that
$$\rho(\mathcal{L}_\omega) = \max_{\alpha \,|\, p_J(\alpha)=0} \{\max(|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|)\},$$
and since we are assuming that J has real roots, we are led to study the function
$$M(\alpha, \omega) = \max\{|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|\},$$
where $\alpha \in \mathbb{R}$ and $\omega \in (0, 2)$. Actually, because $M(-\alpha, \omega) = M(\alpha, \omega)$, it is only necessary to consider the case where $\alpha \geq 0$.

Note that for $\alpha \neq 0$, the roots of the equation
$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$
are
$$\frac{\alpha\omega \pm \sqrt{\alpha^2\omega^2 - 4\omega + 4}}{2}.$$
In turn, this leads to consider the roots of the equation
$$\omega^2\alpha^2 - 4\omega + 4 = 0,$$
which are
$$\frac{2(1 \pm \sqrt{1 - \alpha^2})}{\alpha^2},$$
for $\alpha \neq 0$. Since we have
$$\frac{2(1 + \sqrt{1 - \alpha^2})}{\alpha^2} = \frac{2(1 + \sqrt{1 - \alpha^2})(1 - \sqrt{1 - \alpha^2})}{\alpha^2(1 - \sqrt{1 - \alpha^2})} = \frac{2}{1 - \sqrt{1 - \alpha^2}}$$
and
$$\frac{2(1 - \sqrt{1 - \alpha^2})}{\alpha^2} = \frac{2(1 + \sqrt{1 - \alpha^2})(1 - \sqrt{1 - \alpha^2})}{\alpha^2(1 + \sqrt{1 - \alpha^2})} = \frac{2}{1 + \sqrt{1 - \alpha^2}},$$
these roots are
$$\omega_0(\alpha) = \frac{2}{1 + \sqrt{1 - \alpha^2}}, \quad \omega_1(\alpha) = \frac{2}{1 - \sqrt{1 - \alpha^2}}.$$

Observe that the expression for $\omega_0(\alpha)$ is exactly the expression in the statement of our proposition! The rest of the proof consists in analyzing the variations of the function $M(\alpha, \omega)$ by considering various cases for α. In the end, we find that the minimum of $\rho(\mathcal{L}_\omega)$ is obtained for $\omega_0(\rho(J))$. The details are tedious and we omit them. The reader will find complete proofs in Serre [156] and Ciarlet [41]. $\square$

Combining the results of Theorem 10.6 and Proposition 10.9, we obtain the following result which gives precise information about the spectral radii of the matrices J, $\mathcal{L}_1$, and $\mathcal{L}_\omega$.

Proposition 10.10. *Let A be a tridiagonal matrix (possibly by blocks) which is Hermitian positive definite. Then the methods of Jacobi, Gauss–Seidel, and relaxation, all converge for $\omega \in (0, 2)$. There is a unique optimal relaxation parameter*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

such that

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{0<\omega<2} \rho(\mathcal{L}_\omega) = \omega_0 - 1.$$

Furthermore, if $\rho(J) > 0$, then

$$\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J),$$

and if $\rho(J) = 0$, then $\omega_0 = 1$ and $\rho(\mathcal{L}_1) = \rho(J) = 0$.

Proof. In order to apply Proposition 10.9, we have to check that $J = D^{-1}(E + F)$ has real eigenvalues. However, if α is any eigenvalue of J and if u is any corresponding eigenvector, then

$$D^{-1}(E + F)u = \alpha u$$

implies that

$$(E + F)u = \alpha D u,$$

and since $A = D - E - F$, the above shows that $(D - A)u = \alpha D u$, that is,

$$Au = (1 - \alpha)Du.$$

Consequently,

$$u^* A u = (1 - \alpha) u^* D u,$$

and since A and D are Hermitian positive definite, we have $u^* A u > 0$ and $u^* D u > 0$ since $u \neq 0$, which proves that $\alpha \in \mathbb{R}$. The rest follows from Theorem 10.6 and Proposition 10.9. $\square$

Remark: It is preferable to overestimate rather than underestimate the relaxation parameter when the optimum relaxation parameter is not known exactly.

10.6 Summary

The main concepts and results of this chapter are listed below:

- Iterative methods. Splitting A as $A = M - N$.
- *Convergence of a sequence of vectors or matrices*.
- A criterion for the convergence of the sequence (B^k) of powers of a matrix B to zero in terms of the spectral radius $\rho(B)$.
- A characterization of the spectral radius $\rho(B)$ as the limit of the sequence $(\|B^k\|^{1/k})$.
- A criterion of the convergence of iterative methods.
- Asymptotic behavior of iterative methods.
- Splitting A as $A = D - E - F$, and the methods of *Jacobi*, *Gauss–Seidel*, and *relaxation* (and *SOR*).
- The *Jacobi matrix*, $J = D^{-1}(E + F)$.
- The *Gauss–Seidel matrix*, $\mathcal{L}_1 = (D - E)^{-1}F$.
- The *matrix of relaxation*, $\mathcal{L}_\omega = (D - \omega E)^{-1}((1 - \omega)D + \omega F)$.
- Convergence of iterative methods: a general result when $A = M - N$ is Hermitian positive definite.
- A sufficient condition for the convergence of the methods of Jacobi, Gauss–Seidel, and relaxation. The *Ostrowski-Reich theorem*: A is Hermitian positive definite and $\omega \in (0, 2)$.
- A necessary condition for the convergence of the methods of Jacobi , Gauss–Seidel, and relaxation: $\omega \in (0, 2)$.
- The case of tridiagonal matrices (possibly by blocks). Simultaneous convergence or divergence of Jacobi's method and Gauss–Seidel's method, and comparison of the spectral radii of $\rho(J)$ and $\rho(\mathcal{L}_1)$: $\rho(\mathcal{L}_1) = (\rho(J))^2$.
- The case of tridiagonal Hermitian positive definite matrices (possibly by blocks). The methods of Jacobi, Gauss–Seidel, and relaxation, all converge.
- In the above case, there is a unique optimal relaxation parameter for which $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J)$ (if $\rho(J) \neq 0$).

10.7 Problems

Problem 10.1. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}.$$

Prove that $\rho(J) = 0$ and $\rho(\mathcal{L}_1) = 2$, so

$$\rho(J) < 1 < \rho(\mathcal{L}_1),$$

where J is Jacobi's matrix and $\mathcal{L}_1$ is the matrix of Gauss–Seidel.

Problem 10.2. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

Prove that $\rho(J) = \sqrt{5}/2$ and $\rho(\mathcal{L}_1) = 1/2$, so

$$\rho(\mathcal{L}_1) < \rho(J),$$

where where J is Jacobi's matrix and $\mathcal{L}_1$ is the matrix of Gauss–Seidel.

Problem 10.3. Consider the following linear system:

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 19 \\ 19 \\ -3 \\ -12 \end{pmatrix}.$$

(1) Solve the above system by Gaussian elimination.

(2) Compute the sequences of vectors $u_k = (u_1^k, u_2^k, u_3^k, u_4^k)$ for $k = 1, \ldots, 10$, using the methods of Jacobi, Gauss–Seidel, and relaxation for the following values of ω: $\omega = 1.1, 1.2, \ldots, 1.9$. In all cases, the initial vector is $u_0 = (0, 0, 0, 0)$.

Problem 10.4. Recall that a complex or real $n \times n$ matrix A is *strictly row diagonally dominant* if $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1, \ldots, n$.

(1) Prove that if A is strictly row diagonally dominant, then Jacobi's method converges.

(2) Prove that if A is strictly row diagonally dominant, then Gauss–Seidel's method converges.

Problem 10.5. Prove that the converse of Proposition 10.5 holds. That is, if A is an invertible Hermitian matrix with the splitting $A = M - N$ where M is invertible, if the Hermitian matrix $M^* + N$ is positive definite and if $\rho(M^{-1}N) < 1$, then A is positive definite.

Problem 10.6. Consider the following tridiagonal $n \times n$ matrix:

$$A = \frac{1}{(n+1)^2}\begin{pmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix}.$$

(1) Prove that the eigenvalues of the Jacobi matrix J are given by

$$\lambda_k = \cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n.$$

Hint. First show that the Jacobi matrix is

$$J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & 0 & 1 & 0 \end{pmatrix}.$$

Then the eigenvalues and the eigenvectors of J are solutions of the system of equations

$$\begin{aligned} y_0 &= 0 \\ y_{k+1} + y_{k-1} &= 2\lambda y_k, \quad k = 1, \ldots, n \\ y_{n+1} &= 0, \end{aligned}$$

where the variables y_0 and y_{n+1} are introduced so that the same equation applies for $k = 1, \ldots, n$. It is well known that the general solution to the above recurrence is given by

$$y_k = \alpha z_1^k + \beta z_2^k, \quad k = 0, \ldots, n+1,$$

(with $\alpha, \beta \neq 0$) where z_1 and z_2 are the zeros of the equation

$$z^2 - 2\lambda z + 1 = 0.$$

It follows that $z_2 = z_1^{-1}$ and $z_1 + z_2 = 2\lambda$. The boundary condition $y_0 = 0$ yields $\alpha + \beta = 0$, so $y_k = \alpha(z_1^k - z_1^{-k})$, and the boundary condition $y_{n+1} = 0$ yields

$$z_1^{2(n+1)} = 1.$$

Deduce that we may assume that the n possible values $(z_1)_k$ for z_1 are given by

$$(z_1)_k = e^{\frac{k\pi i}{n+1}}, \quad k = 1, \ldots, n,$$

and find

$$2\lambda_k = (z_1)_k + (z_1)_k^{-1}.$$

Show that an eigenvector $(y_1^{(k)}, \ldots, y_n^{(k)})$ associated with the eigenvalue λ_k is given by

$$y_j^{(k)} = \sin\left(\frac{kj\pi}{n+1}\right), \quad j = 1, \ldots, n.$$

(2) Find the spectral radius $\rho(J)$, $\rho(\mathcal{L}_1)$, and $\rho(\mathcal{L}_{\omega_0})$, as functions of $h = 1/(n+1)$.

Chapter 11

The Dual Space and Duality

In this chapter all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$.

11.1 The Dual Space E^* and Linear Forms

In Section 3.9 we defined linear forms, the dual space $E^* = \mathrm{Hom}(E, K)$ of a vector space E, and showed the existence of dual bases for vector spaces of finite dimension.

In this chapter we take a deeper look at the connection between a space E and its dual space E^*. As we will see shortly, every linear map $f\colon E \to F$ gives rise to a linear map $f^\top\colon F^* \to E^*$, and it turns out that in a suitable basis, the matrix of $f^\top$ is the transpose of the matrix of f. *Thus, the notion of dual space provides a conceptual explanation of the phenomena associated with transposition.*

But it does more, because it allows us to view a linear equation as an element of the dual space E^*, and thus to view subspaces of E as solutions of sets of linear equations and vice-versa. The relationship between subspaces and sets of linear forms is the essence of *duality*, a term which is often used loosely, but can be made precise as a bijection between the set of subspaces of a given vector space E and the set of subspaces of its dual E^*. In this correspondence, a subspace V of E yields the subspace V^0 of E^* consisting of all linear forms that vanish on V (that is, have the value zero for all input in V).

Consider the following set of two "linear equations" in $\mathbb{R}^3$,

$$\begin{aligned} x - y + z &= 0 \\ x - y - z &= 0, \end{aligned}$$

and let us find out what is their set V of common solutions $(x, y, z) \in \mathbb{R}^3$. By subtracting the second equation from the first, we get $2z = 0$, and by adding the two equations, we find that $2(x - y) = 0$, so the set V of solutions is given by

$$\begin{aligned} y &= x \\ z &= 0. \end{aligned}$$

This is a one dimensional subspace of $\mathbb{R}^3$. Geometrically, this is the line of equation $y = x$ in the plane $z = 0$ as illustrated by Figure 11.1.

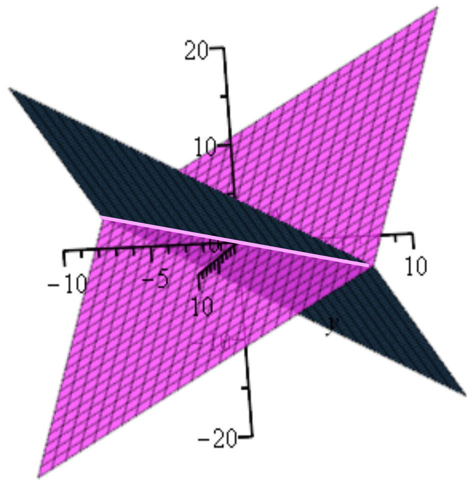


Figure 11.1: The intersection of the magenta plane $x - y + z = 0$ with the blue-gray plane $x - y - z = 0$ is the pink line $y = x$.

Now why did we say that the above equations are linear? Because as functions of (x, y, z), both maps $f_1 \colon (x, y, z) \mapsto x - y + z$ and $f_2 \colon (x, y, z) \mapsto x - y - z$ are linear. The set of all such linear functions from $\mathbb{R}^3$ to $\mathbb{R}$ is a vector space; we used this fact to form linear combinations of the "equations" f_1 and f_2. Observe that the dimension of the subspace V is 1. The ambient space has dimension $n = 3$ and there are two "independent" equations f_1, f_2, so it appears that the dimension $\dim(V)$ of the subspace V defined by m independent equations is

$$\dim(V) = n - m,$$

which is indeed a general fact (proven in Theorem 11.4).

More generally, in $\mathbb{R}^n$, a linear equation is determined by an n-tuple $(a_1, \ldots, a_n) \in \mathbb{R}^n$, and the solutions of this linear equation are given by the n-tuples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ such that

$$a_1 x_1 + \cdots + a_n x_n = 0;$$

these solutions constitute the kernel of the linear map $(x_1, \ldots, x_n) \mapsto a_1 x_1 + \cdots + a_n x_n$. The above considerations assume that we are working in the canonical basis $(e_1, \ldots, e_n)$ of

$\mathbb{R}^n$, but we can define "linear equations" independently of bases and in any dimension, by viewing them as elements of the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K.

Definition 11.1. Given a vector space E, the vector space $\mathrm{Hom}(E, K)$ of linear maps from E to the field K is called the *dual space (or dual)* of E. The space $\mathrm{Hom}(E, K)$ is also denoted by E^*, and the linear maps in E^* are called *the linear forms*, or *covectors*. The dual space E^{**} of the space E^* is called the *bidual* of E.

As a matter of notation, linear forms $f\colon E \to K$ will also be denoted by starred symbol, such as u^*, x^*, *etc.*

Given a vector space E and any basis $(u_i)_{i\in I}$ for E, we can associate to each u_i a linear form $u_i^* \in E^*$, and the u_i^* have some remarkable properties.

Definition 11.2. Given a vector space E and any basis $(u_i)_{i\in I}$ for E, by Proposition 3.18, for every $i \in I$, there is a unique linear form u_i^* such that

$$u_i^*(u_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$

for every $j \in I$. The linear form u_i^* is called the *coordinate form* of index i w.r.t. the basis $(u_i)_{i\in I}$.

The reason for the terminology *coordinate form* was explained in Section 3.9.

We proved in Theorem 3.23 that if $(u_1, \ldots, u_n)$ is a basis of E, then $(u_1^*, \ldots, u_n^*)$ is a basis of E^* called the *dual basis*.

If $(u_1, \ldots, u_n)$ is a basis of $\mathbb{R}^n$ (more generally K^n), it is possible to find explicitly the dual basis $(u_1^*, \ldots, u_n^*)$, where each u_i^* is represented by a row vector.

Example 11.1. For example, consider the columns of the Bézier matrix

$$B_4 = \begin{pmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In other words, we have the basis

$$u_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad u_2 = \begin{pmatrix} -3 \\ 3 \\ 0 \\ 0 \end{pmatrix} \qquad u_3 = \begin{pmatrix} 3 \\ -6 \\ 3 \\ 0 \end{pmatrix} \qquad u_4 = \begin{pmatrix} -1 \\ 3 \\ -3 \\ 1 \end{pmatrix}.$$

Since the form u_1^* is defined by the conditions $u_1^*(u_1) = 1, u_1^*(u_2) = 0, u_1^*(u_3) = 0, u_1^*(u_4) = 0$, it is represented by a row vector $(\lambda_1 \;\; \lambda_2 \;\; \lambda_3 \;\; \lambda_4)$ such that

$$\begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3 & \lambda_4\end{pmatrix}\begin{pmatrix}1 & -3 & 3 & -1\\ 0 & 3 & -6 & 3\\ 0 & 0 & 3 & -3\\ 0 & 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}.$$

This implies that u_1^* is the first row of the inverse of B_4. Since

$$B_4^{-1} = \begin{pmatrix}1 & 1 & 1 & 1\\ 0 & 1/3 & 2/3 & 1\\ 0 & 0 & 1/3 & 1\\ 0 & 0 & 0 & 1\end{pmatrix},$$

the linear forms $(u_1^*, u_2^*, u_3^*, u_4^*)$ correspond to the rows of B_4^{-1}. In particular, u_1^* is represented by $(1 \;\; 1 \;\; 1 \;\; 1)$.

The above method works for any n. Given any basis $(u_1, \ldots, u_n)$ of $\mathbb{R}^n$, if P is the $n \times n$ matrix whose jth column is u_j, then the dual form u_i^* is given by the ith row of the matrix P^{-1}.

When E is of finite dimension n and $(u_1, \ldots, u_n)$ is a basis of E, by Theorem 11.4 (1), the family $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^*. Let us see how the coordinates of a linear form $\varphi^* \in E^*$ over the dual basis $(u_1^*, \ldots, u_n^*)$ vary under a change of basis.

Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two bases of E, and let $P = (a_{i\,j})$ be the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, so that

$$v_j = \sum_{i=1}^n a_{i\,j} u_i,$$

and let $P^{-1} = (b_{i\,j})$ be the inverse of P, so that

$$u_i = \sum_{j=1}^n b_{j\,i} v_j.$$

For fixed j, where $1 \leq j \leq n$, we want to find scalars $(c_i)_{i=1}^n$ such that

$$v_j^* = c_1 u_1^* + c_2 u_2^* + \cdots + c_n u_n^*.$$

To find each c_i, we evaluate the above expression at u_i. Since $u_i^*(u_j) = \delta_{i\,j}$ and $v_i^*(v_j) = \delta_{i\,j}$, we get

$$\begin{aligned} v_j^*(u_i) &= (c_1 u_1^* + c_2 u_2^* + \cdots + c_n u_n^*)(u_i) = c_i\\ v_j^*(u_i) &= v_j^*(\sum_{k=1}^n b_{k\,i} v_k) = b_{j\,i}, \end{aligned}$$

and thus

$$v_j^* = \sum_{i=1}^n b_{j\,i} u_i^*.$$

Similar calculations show that

$$u_i^* = \sum_{j=1}^n a_{i\,j} v_j^*.$$

This means that the change of basis from the dual basis $(u_1^*, \ldots, u_n^*)$ to the dual basis $(v_1^*, \ldots, v_n^*)$ is $(P^{-1})^\top$. Since

$$\varphi^* = \sum_{i=1}^n \varphi_i u_i^* = \sum_{i=1}^n \varphi_i \sum_{j=1}^n a_{ij} v_j^* = \sum_{j=1}^n \left(\sum_{i=1}^n a_{ij} \varphi_i \right) v_j = \sum_{i=1}^n \varphi_i' v_i^*,$$

we get

$$\varphi_j' = \sum_{i=1}^n a_{i\,j} \varphi_i,$$

so the new coordinates φ_j' are expressed in terms of the old coordinates φ_i using the matrix $P^\top$. If we use the row vectors $(\varphi_1, \ldots, \varphi_n)$ and $(\varphi_1', \ldots, \varphi_n')$, we have

$$(\varphi_1', \ldots, \varphi_n') = (\varphi_1, \ldots, \varphi_n) P.$$

These facts are summarized in the following proposition.

Proposition 11.1. *Let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be two bases of E, and let $P = (a_{i\,j})$ be the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, so that*

$$v_j = \sum_{i=1}^n a_{i\,j} u_i.$$

Then the change of basis from the dual basis $(u_1^*, \ldots, u_n^*)$ to the dual basis $(v_1^*, \ldots, v_n^*)$ is $(P^{-1})^\top$, and for any linear form φ, the new coordinates φ_j' of φ are expressed in terms of the old coordinates φ_i of φ using the matrix $P^\top$; that is,

$$(\varphi_1', \ldots, \varphi_n') = (\varphi_1, \ldots, \varphi_n) P.$$

To best understand the preceding paragraph, recall Example 3.1, in which $E = \mathbb{R}^2$, $u_1 = (1, 0)$, $u_2 = (0, 1)$, and $v_1 = (1, 1)$, $v_2 = (-1, 1)$. Then P, the change of basis matrix from (u_1, u_2) to (v_1, v_2), is given by

$$P = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

with $(v_1, v_2) = (u_1, u_2)P$, and $(u_1, u_2) = (v_1, v_2)P^{-1}$, where

$$P^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Let (u_1^*, u_2^*) be the dual basis for (u_1, u_2) and (v_1^*, v_2^*) be the dual basis for (v_1, v_2). We claim that

$$(v_1^*, v_2^*) = (u_1^*, u_2^*)\begin{pmatrix}1/2 & -1/2\\ 1/2 & 1/2\end{pmatrix} = (u_1^*, u_2^*)(P^{-1})^\top,$$

Indeed, since $v_1^* = c_1 u_1^* + c_2 u_2^*$ and $v_2^* = C_1 u_1^* + C_2 u_2^*$ we find that

$$\begin{aligned} c_1 &= v_1^*(u_1) = v_1^*(1/2v_1 - 1/2v_2) = 1/2 & c_2 &= v_1^*(u_2) = v_1^*(1/2v_1 + 1/2v_2) = 1/2\\ C_1 &= v_2^*(u_1) = v_2^*(1/2v_1 - 1/2v_2) = -1/2 & C_2 &= v_2^*(u_2) = v_1^*(1/2v_1 + 1/2v_2) = 1/2. \end{aligned}$$

Furthermore, since $(u_1^*, u_2^*) = (v_1^*, v_2^*)P^\top$ (since $(v_1^*, v_2^*) = (u_1^*, u_2^*)(P^\top)^{-1}$), we find that

$$\varphi^* = \varphi_1 u_1^* + \varphi_2 u_2^* = \varphi_1(v_1^* - v_2^*) + \varphi_(v_1^* + v_2^*) = (\varphi_1 + \varphi_2)v_1^* + (-\varphi_1 + \varphi_2)v_2^* = \varphi_1' v_1^* + \varphi_2' v_2^*$$

Hence

$$\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}\varphi_1\\ \varphi_2\end{pmatrix} = \begin{pmatrix}\varphi_1'\\ \varphi_2'\end{pmatrix},$$

where

$$P^\top = \begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}.$$

Comparing with the change of basis

$$v_j = \sum_{i=1}^n a_{i\,j} u_i,$$

we note that this time, the coordinates (φ_i) of the linear form φ^* change in the *same direction* as the change of basis. For this reason, we say that the coordinates of linear forms are *covariant*. By abuse of language, it is often said that linear forms are *covariant*, which explains why the term *covector* is also used for a linear form.

Observe that if $(e_1, \ldots, e_n)$ is a basis of the vector space E, then, as a linear map from E to K, every linear form $f \in E^*$ is represented by a $1 \times n$ matrix, that is, by a *row vector*

$$(\lambda_1 \ \cdots \ \lambda_n),$$

with respect to the basis $(e_1, \ldots, e_n)$ of E, and 1 of K, where $f(e_i) = \lambda_i$. A vector $u = \sum_{i=1}^n u_i e_i \in E$ is represented by a $n \times 1$ matrix, that is, by a *column vector*

$$\begin{pmatrix}u_1\\ \vdots\\ u_n\end{pmatrix},$$

and the action of f on u, namely $f(u)$, is represented by the matrix product

$$\begin{pmatrix}\lambda_1 & \cdots & \lambda_n\end{pmatrix}\begin{pmatrix}u_1\\ \vdots\\ u_n\end{pmatrix} = \lambda_1 u_1 + \cdots + \lambda_n u_n.$$

On the other hand, with respect to the dual basis $(e_1^*, \ldots, e_n^*)$ of E^*, the linear form f is represented by the column vector

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

Remark: In many texts using tensors, vectors are often indexed with lower indices. If so, it is more convenient to write the coordinates of a vector x over the basis $(u_1, \ldots, u_n)$ as (x^i), using an upper index, so that

$$x = \sum_{i=1}^{n} x^i u_i,$$

and in a change of basis, we have

$$v_j = \sum_{i=1}^{n} a_j^i u_i$$

and

$$x^i = \sum_{j=1}^{n} a_j^i x'^j.$$

Dually, linear forms are indexed with upper indices. Then it is more convenient to write the coordinates of a covector φ^* over the dual basis $(u^{*1}, \ldots, u^{*n})$ as (φ_i), using a lower index, so that

$$\varphi^* = \sum_{i=1}^{n} \varphi_i u^{*i}$$

and in a change of basis, we have

$$u^{*i} = \sum_{j=1}^{n} a_j^i v^{*j}$$

and

$$\varphi'_j = \sum_{i=1}^{n} a_j^i \varphi_i.$$

With these conventions, the index of summation appears once in upper position and once in lower position, and the summation sign can be safely omitted, a trick due to *Einstein*. For example, we can write

$$\varphi'_j = a_j^i \varphi_i$$

as an abbreviation for

$$\varphi'_j = \sum_{i=1}^{n} a_j^i \varphi_i.$$

For another example of the use of Einstein's notation, if the vectors $(v_1, \ldots, v_n)$ are linear combinations of the vectors $(u_1, \ldots, u_n)$, with

$$v_i = \sum_{j=1}^{n} a_{ij} u_j, \quad 1 \leq i \leq n,$$

then the above equations are written as

$$v_i = a_i^j u_j, \quad 1 \leq i \leq n.$$

Thus, in Einstein's notation, the $n \times n$ matrix (a_{ij}) is denoted by (a_i^j), a $(1,1)$-*tensor*.

Beware that some authors view a matrix as a mapping between *coordinates*, in which case the matrix (a_{ij}) is denoted by (a_j^i).

11.2 Pairing and Duality Between E and E^*

Given a linear form $u^* \in E^*$ and a vector $v \in E$, the result $u^*(v)$ of applying u^* to v is also denoted by $\langle u^*, v \rangle$. This defines a binary operation $\langle -, - \rangle \colon E^* \times E \to K$ satisfying the following properties:

$$\begin{aligned}
\langle u_1^* + u_2^*, v \rangle &= \langle u_1^*, v \rangle + \langle u_2^*, v \rangle \\
\langle u^*, v_1 + v_2 \rangle &= \langle u^*, v_1 \rangle + \langle u^*, v_2 \rangle \\
\langle \lambda u^*, v \rangle &= \lambda \langle u^*, v \rangle \\
\langle u^*, \lambda v \rangle &= \lambda \langle u^*, v \rangle.
\end{aligned}$$

The above identities mean that $\langle -, - \rangle$ is a *bilinear map*, since it is linear in each argument. It is often called the *canonical pairing* between E^* and E. In view of the above identities, given any fixed vector $v \in E$, the map $\mathrm{eval}_v \colon E^* \to K$ (*evaluation at* v) defined such that

$$\mathrm{eval}_v(u^*) = \langle u^*, v \rangle = u^*(v) \quad \text{for every } u^* \in E^*$$

is a linear map from E^* to K, that is, eval_v is a linear form in E^{**}. Again, from the above identities, the map $\mathrm{eval}_E \colon E \to E^{**}$, defined such that

$$\mathrm{eval}_E(v) = \mathrm{eval}_v \quad \text{for every } v \in E,$$

is a linear map. Observe that

$$\mathrm{eval}_E(v)(u^*) = \mathrm{eval}_v(u^*) = \langle u^*, v \rangle = u^*(v), \quad \text{for all } v \in E \text{ and all } u^* \in E^*.$$

We shall see that the map eval_E is injective, and that it is an isomorphism when E has finite dimension.

We now formalize the notion of the set V^0 of linear equations vanishing on all vectors in a given subspace $V \subseteq E$, and the notion of the set U^0 of common solutions of a given set $U \subseteq E^*$ of linear equations. The duality theorem (Theorem 11.4) shows that the dimensions of V and V^0, and the dimensions of U and U^0, are related in a crucial way. It also shows that, in finite dimension, the maps $V \mapsto V^0$ and $U \mapsto U^0$ are inverse bijections from subspaces of E to subspaces of E^*.

Definition 11.3. Given a vector space E and its dual E^*, we say that a vector $v \in E$ and a linear form $u^* \in E^*$ are *orthogonal* iff $\langle u^*, v\rangle = 0$. Given a subspace V of E and a subspace U of E^*, we say that V *and* U *are orthogonal* iff $\langle u^*, v\rangle = 0$ for every $u^* \in U$ and every $v \in V$. Given a subset V of E (resp. a subset U of E^*), the *orthogonal* V^0 *of* V is the subspace V^0 of E^* defined such that

$$V^0 = \{u^* \in E^* \mid \langle u^*, v\rangle = 0, \text{ for every } v \in V\}$$

(resp. the *orthogonal* U^0 *of* U is the subspace U^0 of E defined such that

$$U^0 = \{v \in E \mid \langle u^*, v\rangle = 0, \text{ for every } u^* \in U\}).$$

The subspace $V^0 \subseteq E^*$ is also called the *annihilator* of V. The subspace $U^0 \subseteq E$ annihilated by $U \subseteq E^*$ does not have a special name. It seems reasonable to call it the *linear subspace (or linear variety) defined by* U.

Informally, V^0 is the *set of linear equations that vanish on* V, and U^0 is *the set of common zeros of all linear equations in* U. We can also define V^0 by

$$V^0 = \{u^* \in E^* \mid V \subseteq \operatorname{Ker} u^*\}$$

and U^0 by

$$U^0 = \bigcap_{u^* \in U} \operatorname{Ker} u^*.$$

Observe that $E^0 = \{0\} = (0)$, and $\{0\}^0 = E^*$.

Proposition 11.2. *If $V_1 \subseteq V_2 \subseteq E$, then $V_2^0 \subseteq V_1^0 \subseteq E^*$, and if $U_1 \subseteq U_2 \subseteq E^*$, then $U_2^0 \subseteq U_1^0 \subseteq E$. See Figure 11.2.*

Proof. Indeed, if $V_1 \subseteq V_2 \subseteq E$, then for any $f^* \in V_2^0$ we have $f^*(v) = 0$ for all $v \in V_2$, and thus $f^*(v) = 0$ for all $v \in V_1$, so $f^* \in V_1^0$. Similarly, if $U_1 \subseteq U_2 \subseteq E^*$, then for any $v \in U_2^0$, we have $f^*(v) = 0$ for all $f^* \in U_2$, so $f^*(v) = 0$ for all $f^* \in U_1$, which means that $v \in U_1^0$. $\square$

Here are some examples.

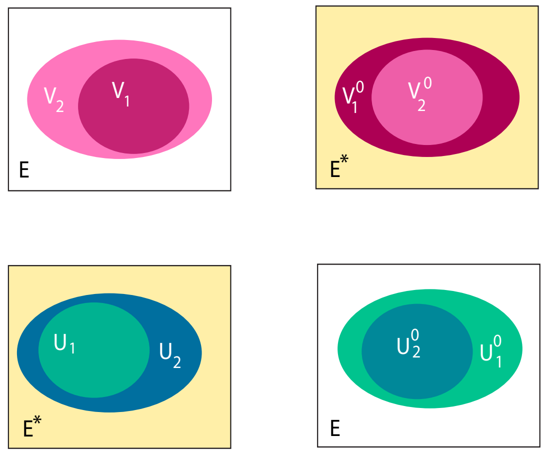


Figure 11.2: The top pair of figures schematically illustrates the relation if $V_1 \subseteq V_2 \subseteq E$, then $V_2^0 \subseteq V_1^0 \subseteq E^*$, while the bottom pair of figures illustrates the relationship if $U_1 \subseteq U_2 \subseteq E^*$, then $U_2^0 \subseteq U_1^0 \subseteq E$.

Example 11.2. Let $E = \mathrm{M}_2(\mathbb{R})$, the space of real 2×2 matrices, and let V be the subspace of $\mathrm{M}_2(\mathbb{R})$ spanned by the matrices

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We check immediately that the subspace V consists of all matrices of the form

$$\begin{pmatrix} b & a \\ a & c \end{pmatrix},$$

that is, all symmetric matrices. The matrices

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

in V satisfy the equation

$$a_{12} - a_{21} = 0,$$

and all scalar multiples of these equations, so V^0 is the subspace of E^* spanned by the linear form given by $u^*(a_{11}, a_{12}, a_{21}, a_{22}) = a_{12} - a_{21}$. By the duality theorem (Theorem 11.4) we have

$$\dim(V^0) = \dim(E) - \dim(V) = 4 - 3 = 1.$$

Example 11.3. The above example generalizes to $E = \mathrm{M}_n(\mathbb{R})$ for any $n \geq 1$, but this time, consider the space U of linear forms asserting that a matrix A is symmetric; these are the linear forms spanned by the $n(n-1)/2$ equations

$$a_{ij} - a_{ji} = 0, \quad 1 \leq i < j \leq n;$$

Note there are no constraints on diagonal entries, and half of the equations

$$a_{ij} - a_{ji} = 0, \quad 1 \leq i \neq j \leq n$$

are redundant. It is easy to check that the equations (linear forms) for which $i < j$ are linearly independent. To be more precise, let U be the space of linear forms in E^* spanned by the linear forms

$$u^*_{ij}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) = a_{ij} - a_{ji}, \quad 1 \leq i < j \leq n.$$

The dimension of U is $n(n-1)/2$. Then the set U^0 of common solutions of these equations is the space $\mathbf{S}(n)$ of symmetric matrices. By the duality theorem (Theorem 11.4), this space has dimension

$$\frac{n(n+1)}{2} = n^2 - \frac{n(n-1)}{2}.$$

We leave it as an exercise to find a basis of $\mathbf{S}(n)$.

Example 11.4. If $E = \mathrm{M}_n(\mathbb{R})$, consider the subspace U of linear forms in E^* spanned by the linear forms

$$\begin{aligned} u^*_{ij}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) &= a_{ij} + a_{ji}, \quad 1 \leq i < j \leq n \\ u^*_{ii}(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn}) &= a_{ii}, \quad 1 \leq i \leq n. \end{aligned}$$

It is easy to see that these linear forms are linearly independent, so $\dim(U) = n(n+1)/2$. The space U^0 of matrices $A \in \mathrm{M}_n(\mathbb{R})$ satifying all of the above equations is clearly the space $\mathbf{Skew}(n)$ of skew-symmetric matrices. By the duality theorem (Theorem 11.4), the dimension of U^0 is

$$\frac{n(n-1)}{2} = n^2 - \frac{n(n+1)}{2}.$$

We leave it as an exercise to find a basis of $\mathbf{Skew}(n)$.

Example 11.5. For yet another example with $E = \mathrm{M}_n(\mathbb{R})$, for any $A \in \mathrm{M}_n(\mathbb{R})$, consider the linear form in E^* given by

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn},$$

called the *trace* of A. The subspace U^0 of E consisting of all matrices A such that $\mathrm{tr}(A) = 0$ is a space of dimension $n^2 - 1$. We leave it as an exercise to find a basis of this space.

The dimension equations

$$\begin{aligned} \dim(V) + \dim(V^0) &= \dim(E) \\ \dim(U) + \dim(U^0) &= \dim(E) \end{aligned}$$

are always true (if E is finite-dimensional). This is part of the duality theorem (Theorem 11.4).

Remark: In contrast with the previous examples, given a matrix $A \in \mathrm{M}_n(\mathbb{R})$, the equations asserting that $A^\top A = I$ are not linear constraints. For example, for $n = 2$, we have

$$\begin{aligned} a_{11}^2 + a_{21}^2 &= 1 \\ a_{21}^2 + a_{22}^2 &= 1 \\ a_{11}a_{12} + a_{21}a_{22} &= 0. \end{aligned}$$

Remarks:

(1) The notation V^0 (resp. U^0) for the orthogonal of a subspace V of E (resp. a subspace U of E^*) is not universal. Other authors use the notation $V^\perp$ (resp. $U^\perp$). However, the notation $V^\perp$ is also used to denote the orthogonal complement of a subspace V with respect to an inner product on a space E, in which case $V^\perp$ is a subspace of E and not a subspace of E^* (see Chapter 12). To avoid confusion, we prefer using the notation V^0.

(2) Since linear forms can be viewed as linear equations (at least in finite dimension), given a subspace (or even a subset) U of E^*, we can define the set $\mathcal{Z}(U)$ of *common zeros* of the equations in U by

$$\mathcal{Z}(U) = \{v \in E \mid u^*(v) = 0, \text{ for all } u^* \in U\}.$$

Of course $\mathcal{Z}(U) = U^0$, but the notion $\mathcal{Z}(U)$ can be generalized to more general kinds of equations, namely polynomial equations. In this more general setting, U is a set of *polynomials* in n variables with coefficients in a field K (where $n = \dim(E)$). Sets of the form $\mathcal{Z}(U)$ are called *algebraic varieties*. Linear forms correspond to the special case where homogeneous polynomials of degree 1 are considered.

If V is a subset of E, it is natural to associate with V the *set of polynomials in* $K[X_1, \ldots, X_n]$ *that vanish on* V. This set, usually denoted $\mathcal{I}(V)$, has some special properties that make it an *ideal*. If V is a linear subspace of E, it is natural to restrict our attention to the space V^0 of linear forms that vanish on V, and in this case we identify $\mathcal{I}(V)$ and V^0 (although technically, $\mathcal{I}(V)$ is no longer an ideal).

For any arbitrary set of polynomials $U \subseteq K[X_1, \ldots, X_n]$ (resp. subset $V \subseteq E$), the relationship between $\mathcal{I}(\mathcal{Z}(U))$ and U (resp. $\mathcal{Z}(\mathcal{I}(V))$ and V) is generally not simple, even though we always have

$$U \subseteq \mathcal{I}(\mathcal{Z}(U)) \quad (\text{resp.} \quad V \subseteq \mathcal{Z}(\mathcal{I}(V))).$$

However, when the field K is algebraically closed, then $\mathcal{I}(\mathcal{Z}(U))$ is equal to the *radical* of the ideal U, a famous result due to Hilbert known as the *Nullstellensatz* (see Lang [109] or Dummit and Foote [54]). The study of algebraic varieties is the main subject of *algebraic geometry*, a beautiful but formidable subject. For a taste of algebraic geometry, see Lang [109] or Dummit and Foote [54].

The duality theorem (Theorem 11.4) shows that the situation is much simpler if we restrict our attention to linear subspaces; in this case

$$U = \mathcal{I}(\mathcal{Z}(U)) \quad \text{and} \quad V = \mathcal{Z}(\mathcal{I}(V)).$$

Proposition 11.3. *We have $V \subseteq V^{00}$ for every subspace V of E, and $U \subseteq U^{00}$ for every subspace U of E^*.*

Proof. Indeed, for any $v \in V$, to show that $v \in V^{00}$ we need to prove that $u^*(v) = 0$ for all $u^* \in V^0$. However, V^0 consists of all linear forms u^* such that $u^*(y) = 0$ for *all* $y \in V$; in particular, for a fixed $v \in V$, we have $u^*(v) = 0$ for all $u^* \in V^0$, as required.

Similarly, for any $u^* \in U$, to show that $u^* \in U^{00}$ we need to prove that $u^*(v) = 0$ for all $v \in U^0$. However, U^0 consists of all vectors v such that $f^*(v) = 0$ for *all* $f^* \in U$; in particular, for a fixed $u^* \in U$, we have $u^*(v) = 0$ for all $v \in U^0$, as required. ☐

We will see shortly that in finite dimension, we have $V = V^{00}$ and $U = U^{00}$.

However, even though $V = V^{00}$ is always true, when E is of infinite dimension, it is not always true that $U = U^{00}$.

Given a vector space E and a subspace U of E, by Theorem 3.7, every basis $(u_i)_{i\in I}$ of U can be extended to a basis $(u_j)_{j\in I\cup J}$ of E, where $I \cap J = \emptyset$.

11.3 The Duality Theorem and Some Consequences

We have the following important theorem adapted from E. Artin [6] (Chapter 1).

Theorem 11.4. *(Duality theorem) Let E be a vector space. The following properties hold:*

(a) For every basis $(u_i)_{i\in I}$ of E, the family $(u_i^*)_{i\in I}$ of coordinate forms is linearly independent.

(b) For every subspace V of E, we have $V^{00} = V$.

(c) For every subspace V of finite codimension m of E, for every subspace W of E such that $E = V \oplus W$ (where W is of finite dimension m), for every basis $(u_i)_{i\in I}$ of E such that $(u_1, \ldots, u_m)$ is a basis of W, the family $(u_1^*, \ldots, u_m^*)$ is a basis of the orthogonal V^0 of V in E^*, so that

$$\dim(V^0) = \operatorname{codim}(V).$$

Furthermore, we have $V^{00} = V$.

(d) For every subspace U of finite dimension m of E^*, the orthogonal U^0 of U in E is of finite codimension m, so that

$$\operatorname{codim}(U^0) = \dim(U).$$

Furthermore, $U^{00} = U$.

Proof. (a) Assume that

$$\sum_{i \in I} \lambda_i u_i^* = 0,$$

for a family $(\lambda_i)_{i \in I}$ (of scalars in K). Since $(\lambda_i)_{i \in I}$ has finite support, there is a finite subset J of I such that $\lambda_i = 0$ for all $i \in I - J$, and we have

$$\sum_{j \in J} \lambda_j u_j^* = 0.$$

Applying the linear form $\sum_{j \in J} \lambda_j u_j^*$ to each u_j ($j \in J$), by Definition 11.2, since $u_i^*(u_j) = 1$ if $i = j$ and 0 otherwise, we get $\lambda_j = 0$ for all $j \in J$, that is $\lambda_i = 0$ for all $i \in I$ (by definition of J as the support). Thus, $(u_i^*)_{i \in I}$ is linearly independent.

(b) Clearly, we have $V \subseteq V^{00}$. If $V \neq V^{00}$, then let $(u_i)_{i \in I \cup J}$ be a basis of V^{00} such that $(u_i)_{i \in I}$ is a basis of V (where $I \cap J = \emptyset$). Since $V \neq V^{00}$, $u_{j_0} \in V^{00}$ for some $j_0 \in J$ (and thus, $j_0 \notin I$). Since $u_{j_0} \in V^{00}$, u_{j_0} is orthogonal to every linear form in V^0. Now, we have $u_{j_0}^*(u_i) = 0$ for all $i \in I$, and thus $u_{j_0}^* \in V^0$. However, $u_{j_0}^*(u_{j_0}) = 1$, contradicting the fact that u_{j_0} is orthogonal to every linear form in V^0. Thus, $V = V^{00}$.

(c) Let $J = I - \{1, \ldots, m\}$. Every linear form $f^* \in V^0$ is orthogonal to every u_j, for $j \in J$, and thus, $f^*(u_j) = 0$, for all $j \in J$. For such a linear form $f^* \in V^0$, let

$$g^* = f^*(u_1)u_1^* + \cdots + f^*(u_m)u_m^*.$$

We have $g^*(u_i) = f^*(u_i)$, for every i, $1 \leq i \leq m$. Furthermore, by definition, g^* vanishes on all u_j, where $j \in J$. Thus, f^* and g^* agree on the basis $(u_i)_{i \in I}$ of E, and so, $g^* = f^*$. This shows that $(u_1^*, \ldots, u_m^*)$ generates V^0, and since it is also a linearly independent family, $(u_1^*, \ldots, u_m^*)$ is a basis of V^0. It is then obvious that $\dim(V^0) = \operatorname{codim}(V)$, and by part (b), we have $V^{00} = V$.

(d) Let $(u_1^*, \ldots, u_m^*)$ be a basis of U. Note that the map $h\colon E \to K^m$ defined such that

$$h(v) = (u_1^*(v), \ldots, u_m^*(v))$$

for every $v \in E$, is a linear map, and that its kernel $\operatorname{Ker} h$ is precisely U^0. Then, by Proposition 6.16,

$$E \approx \operatorname{Ker}(h) \oplus \operatorname{Im} h = U^0 \oplus \operatorname{Im} h,$$

and since $\dim(\operatorname{Im} h) \leq m$, we deduce that U^0 is a subspace of E of finite codimension at most m, and by (c), we have $\dim(U^{00}) = \operatorname{codim}(U^0) \leq m = \dim(U)$. However, it is clear that $U \subseteq U^{00}$, which implies $\dim(U) \leq \dim(U^{00})$, and so $\dim(U^{00}) = \dim(U) = m$, and we must have $U = U^{00}$. $\square$

Part (a) of Theorem 11.4 shows that

$$\dim(E) \leq \dim(E^*).$$

When E is of finite dimension n and $(u_1, \ldots, u_n)$ is a basis of E, by part (c), the family $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^*, called the *dual basis* of $(u_1, \ldots, u_n)$. This fact was also proven directly in Theorem 3.23.

Define the function $\mathcal{E}$ ($\mathcal{E}$ for equations) from subspaces of E to subspaces of E^* and the function $\mathcal{Z}$ ($\mathcal{Z}$ for zeros) from subspaces of E^* to subspaces of E by

$$\begin{aligned} \mathcal{E}(V) &= V^0, \quad V \subseteq E \\ \mathcal{Z}(U) &= U^0, \quad U \subseteq E^*. \end{aligned}$$

By Parts (c) and (d) of Theorem 11.4,

$$\begin{aligned} (\mathcal{Z} \circ \mathcal{E})(V) &= V^{00} = V \\ (\mathcal{E} \circ \mathcal{Z})(U) &= U^{00} = U, \end{aligned}$$

so $\mathcal{Z} \circ \mathcal{E} = \mathrm{id}$ and $\mathcal{E} \circ \mathcal{Z} = \mathrm{id}$, and the maps $\mathcal{E}$ and $\mathcal{Z}$ are inverse bijections. These maps set up a *duality* between subspaces of E and subspaces of E^*. In particular, every subspace $V \subseteq E$ of dimension m is the set of common zeros of the space of linear forms (equations) V^0, which has dimension $n - m$. This confirms the claim we made about the dimension of the subpsace defined by a set of linear equations.

One should be careful that this bijection does not extend to subspaces of E^* of infinite dimension.

When E is of infinite dimension, for every basis $(u_i)_{i \in I}$ of E, the family $(u_i^*)_{i \in I}$ of coordinate forms is never a basis of E^*. It is linearly independent, but it is "too small" to generate E^*. For example, if $E = \mathbb{R}^{(\mathbb{N})}$, where $\mathbb{N} = \{0, 1, 2, \ldots\}$, the map $f \colon E \to \mathbb{R}$ that sums the nonzero coordinates of a vector in E is a linear form, but it is easy to see that it cannot be expressed as a linear combination of coordinate forms. As a consequence, when E is of infinite dimension, E and E^* are not isomorphic.

We now discuss some applications of the duality theorem.

Problem 1 . Suppose that V is a subspace of $\mathbb{R}^n$ of dimension m and that $(v_1, \ldots, v_m)$ is a basis of V. The problem is to find a basis of V^0.

We first extend $(v_1, \ldots, v_m)$ to a basis $(v_1, \ldots, v_n)$ of $\mathbb{R}^n$, and then by part (c) of Theorem 11.4, we know that $(v_{m+1}^*, \ldots, v_n^*)$ is a basis of V^0.

Example 11.6. For example, suppose that V is the subspace of $\mathbb{R}^4$ spanned by the two linearly independent vectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

the first two vectors of the Haar basis in $\mathbb{R}^4$. The four columns of the Haar matrix

$$W = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

form a basis of $\mathbb{R}^4$, and the inverse of W is given by

$$W^{-1} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -1/4 & -1/4 \\ 1/2 & -1/2 & 0 & 0 \\ 0 & 0 & 1/2 & -1/2 \end{pmatrix}.$$

Since the dual basis $(v_1^*, v_2^*, v_3^*, v_4^*)$ is given by the rows of W^{-1}, the last two rows of W^{-1},

$$\begin{pmatrix} 1/2 & -1/2 & 0 & 0 \\ 0 & 0 & 1/2 & -1/2 \end{pmatrix},$$

form a basis of V^0. We also obtain a basis by rescaling by the factor $1/2$, so the linear forms given by the row vectors

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

form a basis of V^0, the space of linear forms (linear equations) that vanish on the subspace V.

The method that we described to find V^0 requires first extending a basis of V and then inverting a matrix, but there is a more direct method. Indeed, let A be the $n \times m$ matrix whose columns are the basis vectors $(v_1, \ldots, v_m)$ of V. Then a linear form u represented by a row vector belongs to V^0 iff $uv_i = 0$ for $i = 1, \ldots, m$ iff

$$uA = 0$$

iff

$$A^\top u^\top = 0.$$

Therefore, all we need to do is to find a basis of the nullspace of $A^\top$. This can be done quite effectively using the reduction of a matrix to reduced row echelon form (rref); see Section 8.10.

Example 11.7. For example, if we reconsider the previous example, $A^\top u^\top = 0$ becomes

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Since the rref of $A^\top$ is

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

the above system is equivalent to

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} u_1 + u_2 \\ u_3 + u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where the free variables are associated with u_2 and u_4. Thus to determine a basis for the kernel of $A^\top$, we set $u_2 = 1, u_4 = 0$ and $u_2 = 0, u_4 = 1$ and obtain a basis for V^0 as

$$\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix}.$$

Problem 2 . Let us now consider the problem of finding a basis of the hyperplane H in $\mathbb{R}^n$ defined by the equation

$$c_1 x_1 + \cdots + c_n x_n = 0.$$

More precisely, if $u^*(x_1, \ldots, x_n)$ is the linear form in $(\mathbb{R}^n)^*$ given by $u^*(x_1, \ldots, x_n) = c_1 x_1 + \cdots + c_n x_n$, then the hyperplane H is the kernel of u^*. Of course we assume that some c_j is nonzero, in which case the linear form u^* spans a one-dimensional subspace U of $(\mathbb{R}^n)^*$, and $U^0 = H$ has dimension $n - 1$.

Since u^* is not the linear form which is identically zero, there is a smallest positive index $j \leq n$ such that $c_j \neq 0$, so our linear form is really $u^*(x_1, \ldots, x_n) = c_j x_j + \cdots + c_n x_n$. We claim that the following $n - 1$ vectors (in $\mathbb{R}^n$) form a basis of H:

$$\begin{array}{cc}
 & \begin{array}{cccccccc} 1 & 2 & \ldots & j-1 & j & j+1 & \ldots & n-1 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ \vdots \\ j-1 \\ j \\ j+1 \\ j+2 \\ \vdots \\ n \end{array} &
\begin{pmatrix}
1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & -c_{j+1}/c_j & -c_{j+2}/c_j & \ldots & -c_n/c_j \\
0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1
\end{pmatrix}
\end{array}.$$

Observe that the $(n-1) \times (n-1)$ matrix obtained by deleting row j is the identity matrix, so the columns of the above matrix are linearly independent. A simple calculation also shows that the linear form $u^*(x_1, \ldots, x_n) = c_j x_j + \cdots + c_n x_n$ vanishes on every column of the above matrix. For a concrete example in $\mathbb{R}^6$, if $u^*(x_1, \ldots, x_6) = x_3 + 2x_4 + 3x_5 + 4x_6$, we obtain the basis for the hyperplane H of equation

$$x_3 + 2x_4 + 3x_5 + 4x_6 = 0$$

given by the following matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & -3 & -4 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Problem 3 . Conversely, given a hyperplane H in $\mathbb{R}^n$ given as the span of $n-1$ linearly vectors $(u_1, \ldots, u_{n-1})$, it is possible using determinants to find a linear form $(\lambda_1, \ldots, \lambda_n)$ that vanishes on H.

In the case $n = 3$, we are looking for a row vector $(\lambda_1, \lambda_2, \lambda_3)$ such that if

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

are two linearly independent vectors, then

$$\begin{pmatrix} u_1 & u_2 & u_2 \\ v_1 & v_2 & v_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and the cross-product $u \times v$ of u and v given by

$$u \times v = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix}$$

is a solution. In other words, the equation of the plane spanned by u and v is

$$(u_2v_3 - u_3v_2)x + (u_3v_1 - u_1v_3)y + (u_1v_2 - u_2v_1)z = 0.$$

Problem 4 . Here is another example illustrating the power of Theorem 11.4. Let $E = \mathrm{M}_n(\mathbb{R})$, and consider the equations asserting that the sum of the entries in every row of a matrix $A \in \mathrm{M}_n(\mathbb{R})$ is equal to the same number. We have $n-1$ equations

$$\sum_{j=1}^{n} (a_{ij} - a_{i+1j}) = 0, \quad 1 \leq i \leq n-1,$$

and it is easy to see that they are linearly independent. Therefore, the space U of linear forms in E^* spanned by the above linear forms (equations) has dimension $n-1$, and the space U^0 of matrices satisfying all these equations has dimension n^2-n+1. It is not so obvious to find a basis for this space.

We will now pin down the relationship between a vector space E and its bidual E^{**}.

11.4 The Bidual and Canonical Pairings

Proposition 11.5. *Let E be a vector space. The following properties hold:*

(a) The linear map $\mathrm{eval}_E\colon E \to E^{**}$ defined such that

$$\mathrm{eval}_E(v) = \mathrm{eval}_v \quad \textit{for all } v \in E,$$

that is, $\mathrm{eval}_E(v)(u^*) = \langle u^*, v\rangle = u^*(v)$ for every $u^* \in E^*$, is injective.

(b) When E is of finite dimension n, the linear map $\mathrm{eval}_E\colon E \to E^{**}$ is an isomorphism (called the canonical isomorphism).

Proof. (a) Let $(u_i)_{i\in I}$ be a basis of E, and let $v = \sum_{i\in I} v_i u_i$. If $\mathrm{eval}_E(v) = 0$, then in particular $\mathrm{eval}_E(v)(u_i^*) = 0$ for all u_i^*, and since

$$\mathrm{eval}_E(v)(u_i^*) = \langle u_i^*, v\rangle = v_i,$$

we have $v_i = 0$ for all $i \in I$, that is, $v = 0$, showing that $\mathrm{eval}_E\colon E \to E^{**}$ is injective.

If E is of finite dimension n, by Theorem 11.4, for every basis $(u_1, \ldots, u_n)$, the family $(u_1^*, \ldots, u_n^*)$ is a basis of the dual space E^*, and thus the family $(u_1^{**}, \ldots, u_n^{**})$ is a basis of the bidual E^{**}. This shows that $\dim(E) = \dim(E^{**}) = n$, and since by Part (a), we know that $\mathrm{eval}_E\colon E \to E^{**}$ is injective, in fact, $\mathrm{eval}_E\colon E \to E^{**}$ is bijective (by Proposition 6.19). ☐

When a vector space E has infinite dimension, E and its bidual E^{**} are never isomorphic.

When E is of finite dimension and $(u_1, \ldots, u_n)$ is a basis of E, in view of the canonical isomorphism $\mathrm{eval}_E\colon E \to E^{**}$, the basis $(u_1^{**}, \ldots, u_n^{**})$ of the bidual is *identified* with $(u_1, \ldots, u_n)$.

Proposition 11.5 can be reformulated very fruitfully in terms of pairings, a remarkably useful concept discovered by Pontrjagin in 1931 (adapted from E. Artin [6], Chapter 1). Given two vector spaces E and F over a field K, we say that a function $\varphi\colon E \times F \to K$ is *bilinear* if for every $v \in V$, the map $u \mapsto \varphi(u, v)$ (from E to K) is linear, and for every $u \in E$, the map $v \mapsto \varphi(u, v)$ (from F to K) is linear.

Definition 11.4. Given two vector spaces E and F over K, a *pairing between E and F* is a bilinear map $\varphi\colon E \times F \to K$. Such a pairing is *nondegenerate* iff

(1) for every $u \in E$, if $\varphi(u, v) = 0$ for all $v \in F$, then $u = 0$, and

(2) for every $v \in F$, if $\varphi(u, v) = 0$ for all $u \in E$, then $v = 0$.

A pairing $\varphi\colon E \times F \to K$ is often denoted by $\langle -, - \rangle\colon E \times F \to K$. For example, the map $\langle -, - \rangle\colon E^* \times E \to K$ defined earlier is a nondegenerate pairing (use the proof of (a) in Proposition 11.5). If $E = F$ and $K = \mathbb{R}$, any inner product on E is a nondegenerate pairing (because an inner product is positive definite); see Chapter 12. Other interesting nondegenerate pairings arise in exterior algebra and differential geometry.

Given a pairing $\varphi\colon E \times F \to K$, we can define two maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ as follows: For every $u \in E$, we define the linear form $l_\varphi(u)$ in F^* such that

$$l_\varphi(u)(y) = \varphi(u, y) \quad \text{for every } y \in F,$$

and for every $v \in F$, we define the linear form $r_\varphi(v)$ in E^* such that

$$r_\varphi(v)(x) = \varphi(x, v) \quad \text{for every } x \in E.$$

We have the following useful proposition.

Proposition 11.6. *Given two vector spaces E and F over K, for every nondegenerate pairing $\varphi\colon E \times F \to K$ between E and F, the maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear and injective. Furthermore, if E and F have finite dimension, then this dimension is the same and $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are bijections.*

Proof. The maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear because a pairing is bilinear. If $l_\varphi(u) = 0$ (the null form), then

$$l_\varphi(u)(v) = \varphi(u, v) = 0 \quad \text{for every } v \in F,$$

and since φ is nondegenerate, $u = 0$. Thus, $l_\varphi\colon E \to F^*$ is injective. Similarly, $r_\varphi\colon F \to E^*$ is injective. When F has finite dimension n, we have seen that F and F^* have the same dimension. Since $l_\varphi\colon E \to F^*$ is injective, we have $m = \dim(E) \leq \dim(F) = n$. The same argument applies to E, and thus $n = \dim(F) \leq \dim(E) = m$. But then, $\dim(E) = \dim(F)$, and $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are bijections. $\square$

When E has finite dimension, the nondegenerate pairing $\langle -, - \rangle\colon E^* \times E \to K$ yields another proof of the existence of a natural isomorphism between E and E^{**}. When $E = F$, the nondegenerate pairing induced by an inner product on E yields a natural isomorphism between E and E^* (see Section 12.2).

We now show the relationship between hyperplanes and linear forms.

11.5 Hyperplanes and Linear Forms

Actually Proposition 11.7 below follows from Parts (c) and (d) of Theorem 11.4, but we feel that it is also interesting to give a more direct proof.

Proposition 11.7. *Let E be a vector space. The following properties hold:*

(a) Given any nonnull linear form $f^* \in E^*$, its kernel $H = \mathrm{Ker}\, f^*$ is a hyperplane.

(b) For any hyperplane H in E, there is a (nonnull) linear form $f^* \in E^*$ such that $H = \mathrm{Ker}\, f^*$.

(c) Given any hyperplane H in E and any (nonnull) linear form $f^* \in E^*$ such that $H = \mathrm{Ker}\, f^*$, for every linear form $g^* \in E^*$, $H = \mathrm{Ker}\, g^*$ iff $g^* = \lambda f^*$ for some $\lambda \neq 0$ in K.

Proof. (a) If $f^* \in E^*$ is nonnull, there is some vector $v_0 \in E$ such that $f^*(v_0) \neq 0$. Let $H = \mathrm{Ker}\, f^*$. For every $v \in E$, we have

$$f^*\left(v - \frac{f^*(v)}{f^*(v_0)} v_0\right) = f^*(v) - \frac{f^*(v)}{f^*(v_0)} f^*(v_0) = f^*(v) - f^*(v) = 0.$$

Thus,

$$v - \frac{f^*(v)}{f^*(v_0)} v_0 = h \in H,$$

and

$$v = h + \frac{f^*(v)}{f^*(v_0)} v_0,$$

that is, $E = H + Kv_0$. Also since $f^*(v_0) \neq 0$, we have $v_0 \notin H$, that is, $H \cap Kv_0 = 0$. Thus, $E = H \oplus Kv_0$, and H is a hyperplane.

(b) If H is a hyperplane, $E = H \oplus Kv_0$ for some $v_0 \notin H$. Then every $v \in E$ can be written in a unique way as $v = h + \lambda v_0$. Thus there is a well-defined function $f^* \colon E \to K$, such that, $f^*(v) = \lambda$, for every $v = h + \lambda v_0$. We leave as a simple exercise the verification that f^* is a linear form. Since $f^*(v_0) = 1$, the linear form f^* is nonnull. Also, by definition, it is clear that $\lambda = 0$ iff $v \in H$, that is, $\mathrm{Ker}\, f^* = H$.

(c) Let H be a hyperplane in E, and let $f^* \in E^*$ be any (nonnull) linear form such that $H = \mathrm{Ker}\, f^*$. Clearly, if $g^* = \lambda f^*$ for some $\lambda \neq 0$, then $H = \mathrm{Ker}\, g^*$. Conversely, assume that $H = \mathrm{Ker}\, g^*$ for some nonnull linear form g^*. From (a), we have $E = H \oplus Kv_0$, for some v_0 such that $f^*(v_0) \neq 0$ and $g^*(v_0) \neq 0$. Then observe that

$$g^* - \frac{g^*(v_0)}{f^*(v_0)} f^*$$

is a linear form that vanishes on H, since both f^* and g^* vanish on H, but also vanishes on Kv_0. Thus, $g^* = \lambda f^*$, with

$$\lambda = \frac{g^*(v_0)}{f^*(v_0)}.$$

$\square$

We leave as an exercise the fact that every subspace $V \neq E$ of a vector space E is the intersection of all hyperplanes that contain V. We now consider the notion of transpose of a linear map and of a matrix.

11.6 Transpose of a Linear Map and of a Matrix

Given a linear map $f\colon E \to F$, it is possible to define a map $f^\top\colon F^* \to E^*$ which has some interesting properties.

Definition 11.5. Given a linear map $f\colon E \to F$, the *transpose* $f^\top\colon F^* \to E^*$ *of* f is the linear map defined such that

$$f^\top(v^*) = v^* \circ f, \quad \text{for every } v^* \in F^*,$$

as shown in the diagram below:

$$\begin{tikzcd} E \arrow[r, "f"] \arrow[dr, "f^\top(v^*)"'] & F \arrow[d, "v^*"] \\ & K. \end{tikzcd}$$

Equivalently, the linear map $f^\top\colon F^* \to E^*$ is defined such that

$$\langle v^*, f(u)\rangle = \langle f^\top(v^*), u\rangle, \tag{$*$}$$

for all $u \in E$ and all $v^* \in F^*$.

It is easy to verify that the following properties hold:

$$\begin{aligned} (f+g)^\top &= f^\top + g^\top \\ (g \circ f)^\top &= f^\top \circ g^\top \\ \mathrm{id}_E^\top &= \mathrm{id}_{E^*}. \end{aligned}$$

Note the reversal of composition on the right-hand side of $(g \circ f)^\top = f^\top \circ g^\top$.

The equation $(g \circ f)^\top = f^\top \circ g^\top$ implies the following useful proposition.

Proposition 11.8. *If $f\colon E \to F$ is any linear map, then the following properties hold:*

(1) If f is injective, then $f^\top$ is surjective.

(2) If f is surjective, then $f^\top$ is injective.

Proof. If $f\colon E \to F$ is injective, then it has a retraction $r\colon F \to E$ such that $r \circ f = \mathrm{id}_E$, and if $f\colon E \to F$ is surjective, then it has a section $s\colon F \to E$ such that $f \circ s = \mathrm{id}_F$. Now if $f\colon E \to F$ is injective, then we have

$$(r \circ f)^\top = f^\top \circ r^\top = \mathrm{id}_{E^*},$$

which implies that $f^\top$ is surjective, and if f is surjective, then we have

$$(f \circ s)^\top = s^\top \circ f^\top = \mathrm{id}_{F^*},$$

which implies that $f^\top$ is injective. $\square$

The following proposition gives a natural interpretation of the dual $(E/U)^*$ of a quotient space E/U.

Proposition 11.9. *For any subspace U of a vector space E, if $p\colon E \to E/U$ is the canonical surjection onto E/U, then $p^\top$ is injective and*

$$\mathrm{Im}(p^\top) = U^0 = (\mathrm{Ker}\,(p))^0.$$

Therefore, $p^\top$ is a linear isomorphism between $(E/U)^*$ and U^0.

Proof. Since p is surjective, by Proposition 11.8, the map $p^\top$ is injective. Obviously, $U = \mathrm{Ker}\,(p)$. Observe that $\mathrm{Im}(p^\top)$ consists of all linear forms $\psi \in E^*$ such that $\psi = \varphi \circ p$ for some $\varphi \in (E/U)^*$, and since $\mathrm{Ker}\,(p) = U$, we have $U \subseteq \mathrm{Ker}\,(\psi)$. Conversely for any linear form $\psi \in E^*$, if $U \subseteq \mathrm{Ker}\,(\psi)$, then ψ factors through E/U as $\psi = \overline{\psi} \circ p$ as shown in the following commutative diagram

$$\begin{tikzcd} E \arrow[r, "p"] \arrow[dr, "\psi"'] & E/U \arrow[d, "\overline{\psi}"] \\ & K, \end{tikzcd}$$

where $\overline{\psi}\colon E/U \to K$ is given by

$$\overline{\psi}(\overline{v}) = \psi(v), \quad v \in E,$$

where $\overline{v} \in E/U$ denotes the equivalence class of $v \in E$. The map $\overline{\psi}$ does not depend on the representative chosen in the equivalence class $\overline{v}$, since if $\overline{v'} = \overline{v}$, that is $v' - v = u \in U$, then $\psi(v') = \psi(v + u) = \psi(v) + \psi(u) = \psi(v) + 0 = \psi(v)$. Therefore, we have

$$\begin{aligned} \mathrm{Im}(p^\top) &= \{\varphi \circ p \mid \varphi \in (E/U)^*\} \\ &= \{\psi\colon E \to K \mid U \subseteq \mathrm{Ker}\,(\psi)\} \\ &= U^0, \end{aligned}$$

which proves our result. $\square$

Proposition 11.9 yields another proof of part (b) of the duality theorem (theorem 11.4) that does not involve the existence of bases (in infinite dimension).

Proposition 11.10. *For any vector space E and any subspace V of E, we have $V^{00} = V$.*

Proof. We begin by observing that $V^0 = V^{000}$. This is because, for any subspace U of E^*, we have $U \subseteq U^{00}$, so $V^0 \subseteq V^{000}$. Furthermore, $V \subseteq V^{00}$ holds, and for any two subspaces M, N of E, if $M \subseteq N$ then $N^0 \subseteq N^0$, so we get $V^{000} \subseteq V^0$. Write $V_1 = V^{00}$, so that $V_1^0 = V^{000} = V^0$. We wish to prove that $V_1 = V$.

Since $V \subseteq V_1 = V^{00}$, the canonical projection $p_1 \colon E \to E/V_1$ factors as $p_1 = f \circ p$ as in the diagram below,

$$\begin{tikzcd} E \arrow[r, "p"] \arrow[dr, "p_1"'] & E/V \arrow[d, "f"] \\ & E/V_1 \end{tikzcd}$$

where $p\colon E \to E/V$ is the canonical projection onto E/V and $f\colon E/V \to E/V_1$ is the quotient map induced by p_1, with $f(\overline{u}_{E/V}) = p_1(u) = \overline{u}_{E/V_1}$, for all $u \in E$ (since $V \subseteq V_1$, if $u - u' = v \in V$, then $u - u' = v \in V_1$, so $p_1(u) = p_1(u')$). Since p_1 is surjective, so is f. We wish to prove that f is actually an isomorphism, and for this, it is enough to show that f is injective. By transposing all the maps, we get the commutative diagram

$$\begin{tikzcd} E^* & (E/V)^* \arrow[l, "p^\top"'] \\ & (E/V_1)^*, \arrow[ul, "p_1^\top"] \arrow[u, "f^\top"'] \end{tikzcd}$$

but by Proposition 11.9, the maps $p^\top\colon (E/V)^* \to V^0$ and $p_1^\top\colon (E/V_1)^* \to V_1^0$ are isomorphism, and since $V^0 = V_1^0$, we have the following diagram where both $p^\top$ and $p_1^\top$ are isomorphisms:

$$\begin{tikzcd} V^0 & (E/V)^* \arrow[l, "p^\top"'] \\ & (E/V_1)^*. \arrow[ul, "p_1^\top"] \arrow[u, "f^\top"'] \end{tikzcd}$$

Therefore, $f^\top = (p^\top)^{-1} \circ p_1^\top$ is an isomorphism. We claim that this implies that f is injective.

If f is not injective, then there is some $x \in E/V$ such that $x \neq 0$ and $f(x) = 0$, so for every $\varphi \in (E/V_1)^*$, we have $f^\top(\varphi)(x) = \varphi(f(x)) = 0$. However, there is linear form $\psi \in (E/V)^*$ such that $\psi(x) = 1$, so $\psi \neq f^\top(\varphi)$ for all $\varphi \in (E/V_1)^*$, contradicting the fact that $f^\top$ is surjective. To find such a linear form ψ, pick any supplement W of Kx in E/V, so that $E/V = Kx \oplus W$ (W is a hyperplane in E/V not containing x), and define ψ to be zero on W and 1 on x.[1] Therefore, f is injective, and since we already know that it is surjective, it is bijective. This means that the canonical map $f\colon E/V \to E/V_1$ with $V \subseteq V_1$ is an isomorphism, which implies that $V = V_1 = V^{00}$ (otherwise, if $v \in V_1 - V$, then $p_1(v) = 0$, so $f(p(v)) = p_1(v) = 0$, but $p(v) \neq 0$ since $v \notin V$, and f is not injective). ☐

The following proposition shows the relationship between orthogonality and transposition.

Proposition 11.11. *Given a linear map $f\colon E \to F$, for any subspace V of E, we have*

$$f(V)^0 = (f^\top)^{-1}(V^0) = \{w^* \in F^* \mid f^\top(w^*) \in V^0\}.$$

As a consequence,

$$\mathrm{Ker}\, f^\top = (\mathrm{Im}\, f)^0.$$

We also have

$$\mathrm{Ker}\, f = (\mathrm{Im}\, f^\top)^0.$$

Proof. We have

$$\langle w^*, f(v)\rangle = \langle f^\top(w^*), v\rangle,$$

for all $v \in E$ and all $w^* \in F^*$, and thus, we have $\langle w^*, f(v)\rangle = 0$ for every $v \in V$, i.e. $w^* \in f(V)^0$ iff $\langle f^\top(w^*), v\rangle = 0$ for every $v \in V$ iff $f^\top(w^*) \in V^0$, i.e. $w^* \in (f^\top)^{-1}(V^0)$, proving that

$$f(V)^0 = (f^\top)^{-1}(V^0).$$

Since we already observed that $E^0 = (0)$, letting $V = E$ in the above identity we obtain that

$$\mathrm{Ker}\, f^\top = (\mathrm{Im}\, f)^0.$$

From the equation

$$\langle w^*, f(v)\rangle = \langle f^\top(w^*), v\rangle,$$

we deduce that $v \in (\mathrm{Im}\, f^\top)^0$ iff $\langle f^\top(w^*), v\rangle = 0$ for all $w^* \in F^*$ iff $\langle w^*, f(v)\rangle = 0$ for all $w^* \in F^*$. Assume that $v \in (\mathrm{Im}\, f^\top)^0$. If we pick a basis $(w_i)_{i\in I}$ of F, then we have the linear forms $w_i^*\colon F \to K$ such that $w_i^*(w_j) = \delta_{ij}$, and since we must have $\langle w_i^*, f(v)\rangle = 0$ for all $i \in I$ and $(w_i)_{i\in I}$ is a basis of F, we conclude that $f(v) = 0$, and thus $v \in \mathrm{Ker}\, f$ (this is because $\langle w_i^*, f(v)\rangle$ is the coefficient of $f(v)$ associated with the basis vector w_i). Conversely, if $v \in \mathrm{Ker}\, f$, then $\langle w^*, f(v)\rangle = 0$ for all $w^* \in F^*$, so we conclude that $v \in (\mathrm{Im}\, f^\top)^0$. Therefore, $v \in (\mathrm{Im}\, f^\top)^0$ iff $v \in \mathrm{Ker}\, f$; that is,

$$\mathrm{Ker}\, f = (\mathrm{Im}\, f^\top)^0,$$

as claimed. ☐

[1] Using Zorn's lemma, we pick W maximal among all subspaces of E/V such that $Kx \cap W = (0)$; then, $E/V = Kx \oplus W$.

The following theorem shows the relationship between the rank of f and the rank of $f^\top$.

Theorem 11.12. *Given a linear map $f \colon E \to F$, the following properties hold.*

(a) The dual $(\operatorname{Im} f)^*$ of $\operatorname{Im} f$ is isomorphic to $\operatorname{Im} f^\top = f^\top(F^*)$; that is,

$$(\operatorname{Im} f)^* \cong \operatorname{Im} f^\top.$$

(b) If F is finite dimensional, then $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$.

Proof. (a) Consider the linear maps

$$E \xrightarrow{p} \operatorname{Im} f \xrightarrow{j} F,$$

where $E \xrightarrow{p} \operatorname{Im} f$ is the surjective map induced by $E \xrightarrow{f} F$, and $\operatorname{Im} f \xrightarrow{j} F$ is the injective inclusion map of $\operatorname{Im} f$ into F. By definition, $f = j \circ p$. To simplify the notation, let $I = \operatorname{Im} f$. By Proposition 11.8, since $E \xrightarrow{p} I$ is surjective, $I^* \xrightarrow{p^\top} E^*$ is injective, and since $\operatorname{Im} f \xrightarrow{j} F$ is injective, $F^* \xrightarrow{j^\top} I^*$ is surjective. Since $f = j \circ p$, we also have

$$f^\top = (j \circ p)^\top = p^\top \circ j^\top,$$

and since $F^* \xrightarrow{j^\top} I^*$ is surjective, and $I^* \xrightarrow{p^\top} E^*$ is injective, we have an isomorphism between $(\operatorname{Im} f)^*$ and $f^\top(F^*)$.

(b) We already noted that Part (a) of Theorem 11.4 shows that $\dim(F) = \dim(F^*)$, for every vector space F of finite dimension. Consequently, $\dim(\operatorname{Im} f) = \dim((\operatorname{Im} f)^*)$, and thus, by Part (a) we have $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$.

Remark: When both E and F are finite-dimensional, there is also a simple proof of (b) that doesn't use the result of Part (a). By Theorem 11.4(c)

$$\dim(\operatorname{Im} f) + \dim((\operatorname{Im} f)^0) = \dim(F),$$

and by Theorem 6.16

$$\dim(\operatorname{Ker} f^\top) + \dim(\operatorname{Im} f^\top) = \dim(F^*).$$

Furthermore, by Proposition 11.11, we have

$$\operatorname{Ker} f^\top = (\operatorname{Im} f)^0,$$

and since F is finite-dimensional $\dim(F) = \dim(F^*)$, so we deduce

$$\dim(\operatorname{Im} f) + \dim((\operatorname{Im} f)^0) = \dim((\operatorname{Im} f)^0) + \dim(\operatorname{Im} f^\top),$$

which yields $\dim(\operatorname{Im} f) = \dim(\operatorname{Im} f^\top)$; that is, $\operatorname{rk}(f) = \operatorname{rk}(f^\top)$. $\square$

Proposition 11.13. *If $f\colon E \to F$ is any linear map, then the following identities hold:*

$$\begin{aligned} \operatorname{Im} f^\top &= (\operatorname{Ker}(f))^0 \\ \operatorname{Ker}(f^\top) &= (\operatorname{Im} f)^0 \\ \operatorname{Im} f &= (\operatorname{Ker}(f^\top))^0 \\ \operatorname{Ker}(f) &= (\operatorname{Im} f^\top)^0. \end{aligned}$$

Proof. The equation $\operatorname{Ker}(f^\top) = (\operatorname{Im} f)^0$ has already been proven in Proposition 11.11.

By the duality theorem $(\operatorname{Ker}(f))^{00} = \operatorname{Ker}(f)$, so from $\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0$ we get $\operatorname{Ker}(f) = (\operatorname{Im} f^\top)^0$. Similarly, $(\operatorname{Im} f)^{00} = \operatorname{Im} f$, so from $\operatorname{Ker}(f^\top) = (\operatorname{Im} f)^0$ we get $\operatorname{Im} f = (\operatorname{Ker}(f^\top))^0$. Therefore, what is left to be proven is that $\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0$.

Let $p\colon E \to E/\operatorname{Ker}(f)$ be the canonical surjection, $\overline{f}\colon E/\operatorname{Ker}(f) \to \operatorname{Im} f$ be the isomorphism induced by f, and $j\colon \operatorname{Im} f \to F$ be the inclusion map. Then, we have

$$f = j \circ \overline{f} \circ p,$$

which implies that

$$f^\top = p^\top \circ \overline{f}^\top \circ j^\top.$$

Since p is surjective, $p^\top$ is injective, since j is injective, $j^\top$ is surjective, and since $\overline{f}$ is bijective, $\overline{f}^\top$ is also bijective. It follows that $(E/\operatorname{Ker}(f))^* = \operatorname{Im}(\overline{f}^\top \circ j^\top)$, and we have

$$\operatorname{Im} f^\top = \operatorname{Im} p^\top.$$

Since $p\colon E \to E/\operatorname{Ker}(f)$ is the canonical surjection, by Proposition 11.9 applied to $U = \operatorname{Ker}(f)$, we get

$$\operatorname{Im} f^\top = \operatorname{Im} p^\top = (\operatorname{Ker}(f))^0,$$

as claimed. $\square$

In summary, the equation

$$\operatorname{Im} f^\top = (\operatorname{Ker}(f))^0$$

applies in any dimension, and it implies that

$$\operatorname{Ker}(f) = (\operatorname{Im} f^\top)^0.$$

The following proposition shows the relationship between the matrix representing a linear map $f\colon E \to F$ and the matrix representing its transpose $f^\top\colon F^* \to E^*$.

Proposition 11.14. *Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis for E and $(v_1, \ldots, v_m)$ be a basis for F. Given any linear map $f\colon E \to F$, if $M(f)$ is the $m \times n$-matrix representing f w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$, then the $n \times m$-matrix $M(f^\top)$ representing $f^\top\colon F^* \to E^*$ w.r.t. the dual bases $(v_1^*, \ldots, v_m^*)$ and $(u_1^*, \ldots, u_n^*)$ is the transpose $M(f)^\top$ of $M(f)$.*

Proof. Recall that the entry $a_{i\,j}$ in row i and column j of $M(f)$ is the i-th coordinate of $f(u_j)$ over the basis $(v_1, \ldots, v_m)$. By definition of v_i^*, we have $\langle v_i^*, f(u_j)\rangle = a_{i\,j}$. The entry $a^\top_{j\,i}$ in row j and column i of $M(f^\top)$ is the j-th coordinate of

$$f^\top(v_i^*) = a^\top_{1\,i}u_1^* + \cdots + a^\top_{j\,i}u_j^* + \cdots + a^\top_{n\,i}u_n^*$$

over the basis $(u_1^*, \ldots, u_n^*)$, which is just $a^\top_{j\,i} = f^\top(v_i^*)(u_j) = \langle f^\top(v_i^*), u_j\rangle$. Since

$$\langle v_i^*, f(u_j)\rangle = \langle f^\top(v_i^*), u_j\rangle,$$

we have $a_{i\,j} = a^\top_{j\,i}$, proving that $M(f^\top) = M(f)^\top$. $\square$

We now can give a very short proof of the fact that the rank of a matrix is equal to the rank of its transpose.

Proposition 11.15. *Given an $m \times n$ matrix A over a field K, we have* $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$.

Proof. The matrix A corresponds to a linear map $f\colon K^n \to K^m$, and by Theorem 11.12, $\mathrm{rk}(f) = \mathrm{rk}(f^\top)$. By Proposition 11.14, the linear map $f^\top$ corresponds to $A^\top$. Since $\mathrm{rk}(A) = \mathrm{rk}(f)$, and $\mathrm{rk}(A^\top) = \mathrm{rk}(f^\top)$, we conclude that $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$. $\square$

Thus, given an $m\times n$-matrix A, the maximum number of linearly independent columns is equal to the maximum number of linearly independent rows. There are other ways of proving this fact that do not involve the dual space, but instead some elementary transformations on rows and columns.

Proposition 11.15 immediately yields the following criterion for determining the rank of a matrix:

Proposition 11.16. *Given any $m\times n$ matrix A over a field K (typically $K = \mathbb{R}$ or $K = \mathbb{C}$), the rank of A is the maximum natural number r such that there is an invertible $r\times r$ submatrix of A obtained by selecting r rows and r columns of A.*

For example, the 3×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

has rank 2 iff one of the three 2×2 matrices

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{pmatrix} \quad \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

is invertible.

If we combine Proposition 7.11 with Proposition 11.16, we obtain the following criterion for finding the rank of a matrix.

Proposition 11.17. *Given any $m \times n$ matrix A over a field K (typically $K = \mathbb{R}$ or $K = \mathbb{C}$), the rank of A is the maximum natural number r such that there is an $r \times r$ submatrix B of A obtained by selecting r rows and r columns of A, such that* $\det(B) \neq 0$.

This is not a very efficient way of finding the rank of a matrix. We will see that there are better ways using various decompositions such as LU, QR, or SVD.

11.7 Properties of the Double Transpose

First we have the following property showing the naturality of the eval map.

Proposition 11.18. *For any linear map $f \colon E \to F$, we have*

$$f^{\top\top} \circ \mathrm{eval}_E = \mathrm{eval}_F \circ f,$$

or equivalently the following diagram commutes:

$$\begin{tikzcd} E^{**} \arrow[r, "f^{\top\top}"] & F^{**} \\ E \arrow[u, "\mathrm{eval}_E"] \arrow[r, "f"'] & F. \arrow[u, "\mathrm{eval}_F"'] \end{tikzcd}$$

Proof. For every $u \in E$ and every $\varphi \in F^*$, we have

$$\begin{aligned} (f^{\top\top} \circ \mathrm{eval}_E)(u)(\varphi) &= \langle f^{\top\top}(\mathrm{eval}_E(u)), \varphi \rangle \\ &= \langle \mathrm{eval}_E(u), f^{\top}(\varphi) \rangle \\ &= \langle f^{\top}(\varphi), u \rangle \\ &= \langle \varphi, f(u) \rangle \\ &= \langle \mathrm{eval}_F(f(u)), \varphi \rangle \\ &= \langle (\mathrm{eval}_F \circ f)(u), \varphi \rangle \\ &= (\mathrm{eval}_F \circ f)(u)(\varphi), \end{aligned}$$

which proves that $f^{\top\top} \circ \mathrm{eval}_E = \mathrm{eval}_F \circ f$, as claimed. ◻

If E and F are finite-dimensional, then eval_E and eval_F are isomorphisms, so Proposition 11.18 shows that

$$f^{\top\top} = \mathrm{eval}_F \circ f \circ \mathrm{eval}_E^{-1}. \qquad (*)$$

The above equation is often interpreted as follows: if we identify E with its bidual E^{**} and F with its bidual F^{**}, then $f^{\top\top} = f$. This is an abuse of notation; the rigorous statement is $(*)$.

As a corollary of Proposition 11.18, we obtain the following result.

Proposition 11.19. *If* $\dim(E)$ *is finite, then we have*

$$\mathrm{Ker}\,(f^{\top\top}) = \mathrm{eval}_E(\mathrm{Ker}\,(f)).$$

Proof. Indeed, if E is finite-dimensional, the map $\mathrm{eval}_E \colon E \to E^{**}$ is an isomorphism, so every $\varphi \in E^{**}$ is of the form $\varphi = \mathrm{eval}_E(u)$ for some $u \in E$, the map $\mathrm{eval}_F \colon F \to F^{**}$ is injective, and we have

$$\begin{aligned} f^{\top\top}(\varphi) = 0 \quad & \text{iff} \quad f^{\top\top}(\mathrm{eval}_E(u)) = 0 \\ & \text{iff} \quad \mathrm{eval}_F(f(u)) = 0 \\ & \text{iff} \quad f(u) = 0 \\ & \text{iff} \quad u \in \mathrm{Ker}\,(f) \\ & \text{iff} \quad \varphi \in \mathrm{eval}_E(\mathrm{Ker}\,(f)), \end{aligned}$$

which proves that $\mathrm{Ker}\,(f^{\top\top}) = \mathrm{eval}_E(\mathrm{Ker}\,(f))$. ∎

Remarks: If $\dim(E)$ is finite, following an argument of Dan Guralnik, the fact that $\mathrm{rk}(f) = \mathrm{rk}(f^\top)$ can be proven using Proposition 11.19.

Proof. We know from Proposition 11.11 applied to $f^\top \colon F^* \to E^*$ that

$$\mathrm{Ker}\,(f^{\top\top}) = (\mathrm{Im}\, f^\top)^0,$$

and we showed in Proposition 11.19 that

$$\mathrm{Ker}\,(f^{\top\top}) = \mathrm{eval}_E(\mathrm{Ker}\,(f)).$$

It follows (since eval_E is an isomorphism) that

$$\dim((\mathrm{Im}\, f^\top)^0) = \dim(\mathrm{Ker}\,(f^{\top\top})) = \dim(\mathrm{Ker}\,(f)) = \dim(E) - \dim(\mathrm{Im}\, f),$$

and since

$$\dim(\mathrm{Im}\, f^\top) + \dim((\mathrm{Im}\, f^\top)^0) = \dim(E),$$

we get

$$\dim(\mathrm{Im}\, f^\top) = \dim(\mathrm{Im}\, f).$$ ∎

As indicated by Dan Guralnik, if $\dim(E)$ is finite, the above result can be used to prove the following result.

Proposition 11.20. *If* $\dim(E)$ *is finite, then for any linear map* $f \colon E \to F$*, we have*

$$\mathrm{Im}\, f^\top = (\mathrm{Ker}\,(f))^0.$$

Proof. From
$$\langle f^\top(\varphi), u\rangle = \langle \varphi, f(u)\rangle$$
for all $\varphi \in F^*$ and all $u \in E$, we see that if $u \in \mathrm{Ker}\,(f)$, then $\langle f^\top(\varphi), u\rangle = \langle \varphi, 0\rangle = 0$, which means that $f^\top(\varphi) \in (\mathrm{Ker}\,(f))^0$, and thus, $\mathrm{Im}\, f^\top \subseteq (\mathrm{Ker}\,(f))^0$. For the converse, since $\dim(E)$ is finite, we have

$$\dim((\mathrm{Ker}\,(f))^0) = \dim(E) - \dim(\mathrm{Ker}\,(f)) = \dim(\mathrm{Im}\, f),$$

but we just proved that $\dim(\mathrm{Im}\, f^\top) = \dim(\mathrm{Im}\, f)$, so we get

$$\dim((\mathrm{Ker}\,(f))^0) = \dim(\mathrm{Im}\, f^\top),$$

and since $\mathrm{Im}\, f^\top \subseteq (\mathrm{Ker}\,(f))^0$, we obtain

$$\mathrm{Im}\, f^\top = (\mathrm{Ker}\,(f))^0,$$

as claimed. ☐

Remarks:

1. By the duality theorem, since $(\mathrm{Ker}\,(f))^{00} = \mathrm{Ker}\,(f)$, the above equation yields another proof of the fact that
$$\mathrm{Ker}\,(f) = (\mathrm{Im}\, f^\top)^0,$$
when E is finite-dimensional.

2. The equation
$$\mathrm{Im}\, f^\top = (\mathrm{Ker}\,(f))^0$$
is actually valid even if when E if infinite-dimensional, but we will not prove this here.

11.8 The Four Fundamental Subspaces

Given a linear map $f\colon E \to F$ (where E and F are finite-dimensional), Proposition 11.11 revealed that the four spaces

$$\mathrm{Im}\, f,\ \mathrm{Im}\, f^\top,\ \mathrm{Ker}\, f,\ \mathrm{Ker}\, f^\top$$

play a special role. They are often called the *fundamental subspaces* associated with f. These spaces are related in an intimate manner, since Proposition 11.11 shows that

$$\begin{aligned}\mathrm{Ker}\, f &= (\mathrm{Im}\, f^\top)^0\\ \mathrm{Ker}\, f^\top &= (\mathrm{Im}\, f)^0,\end{aligned}$$

and Theorem 11.12 shows that

$$\mathrm{rk}(f) = \mathrm{rk}(f^\top).$$

It is instructive to translate these relations in terms of matrices (actually, certain linear algebra books make a big deal about this!). If $\dim(E) = n$ and $\dim(F) = m$, given any basis $(u_1, \ldots, u_n)$ of E and a basis $(v_1, \ldots, v_m)$ of F, we know that f is represented by an $m \times n$ matrix $A = (a_{i\,j})$, where the jth column of A is equal to $f(u_j)$ over the basis $(v_1, \ldots, v_m)$. Furthermore, the transpose map $f^\top$ is represented by the $n \times m$ matrix $A^\top$ (with respect to the dual bases). Consequently, the four fundamental spaces

$$\mathrm{Im}\, f,\ \mathrm{Im}\, f^\top,\ \mathrm{Ker}\, f,\ \mathrm{Ker}\, f^\top$$

correspond to

(1) The *column space* of A, denoted by $\mathrm{Im}\, A$ or $\mathcal{R}(A)$; this is the subspace of $\mathbb{R}^m$ spanned by the columns of A, which corresponds to the image $\mathrm{Im}\, f$ of f.

(2) The *kernel* or *nullspace* of A, denoted by $\mathrm{Ker}\, A$ or $\mathcal{N}(A)$; this is the subspace of $\mathbb{R}^n$ consisting of all vectors $x \in \mathbb{R}^n$ such that $Ax = 0$.

(3) The *row space* of A, denoted by $\mathrm{Im}\, A^\top$ or $\mathcal{R}(A^\top)$; this is the subspace of $\mathbb{R}^n$ spanned by the rows of A, or equivalently, spanned by the columns of $A^\top$, which corresponds to the image $\mathrm{Im}\, f^\top$ of $f^\top$.

(4) The *left kernel* or *left nullspace* of A denoted by $\mathrm{Ker}\, A^\top$ or $\mathcal{N}(A^\top)$; this is the kernel (nullspace) of $A^\top$, the subspace of $\mathbb{R}^m$ consisting of all vectors $y \in \mathbb{R}^m$ such that $A^\top y = 0$, or equivalently, $y^\top A = 0$.

Recall that the dimension r of $\mathrm{Im}\, f$, which is also equal to the dimension of the column space $\mathrm{Im}\, A = \mathcal{R}(A)$, is the *rank* of A (and f). Then, some our previous results can be reformulated as follows:

1. The column space $\mathcal{R}(A)$ of A has dimension r.
2. The nullspace $\mathcal{N}(A)$ of A has dimension $n - r$.
3. The row space $\mathcal{R}(A^\top)$ has dimension r.
4. The left nullspace $\mathcal{N}(A^\top)$ of A has dimension $m - r$.

The above statements constitute what Strang calls the *Fundamental Theorem of Linear Algebra, Part I* (see Strang [170]).

The two statements

$$\begin{aligned} \mathrm{Ker}\, f &= (\mathrm{Im}\, f^\top)^0 \\ \mathrm{Ker}\, f^\top &= (\mathrm{Im}\, f)^0 \end{aligned}$$

translate to

(1) The nullspace of A is the orthogonal of the row space of A.

(2) The left nullspace of A is the orthogonal of the column space of A.

The above statements constitute what Strang calls the *Fundamental Theorem of Linear Algebra, Part II* (see Strang [170]).

Since vectors are represented by column vectors and linear forms by row vectors (over a basis in E or F), a vector $x \in \mathbb{R}^n$ is orthogonal to a linear form y iff

$$yx = 0.$$

Then, a vector $x \in \mathbb{R}^n$ is orthogonal to the row space of A iff x is orthogonal to every row of A, namely $Ax = 0$, which is equivalent to the fact that x belong to the nullspace of A. Similarly, the column vector $y \in \mathbb{R}^m$ (representing a linear form over the dual basis of F^*) belongs to the nullspace of $A^\top$ iff $A^\top y = 0$, iff $y^\top A = 0$, which means that the linear form given by $y^\top$ (over the basis in F) is orthogonal to the column space of A.

Since (2) is equivalent to the fact that the column space of A is equal to the orthogonal of the left nullspace of A, we get the following criterion for the solvability of an equation of the form $Ax = b$:

The equation $Ax = b$ has a solution iff for all $y \in \mathbb{R}^m$, if $A^\top y = 0$, then $y^\top b = 0$.

Indeed, the condition on the right-hand side says that b is orthogonal to the left nullspace of A; that is, b belongs to the column space of A.

This criterion can be cheaper to check that checking directly that b is spanned by the columns of A. For example, if we consider the system

$$\begin{aligned} x_1 - x_2 &= b_1 \\ x_2 - x_3 &= b_2 \\ x_3 - x_1 &= b_3 \end{aligned}$$

which, in matrix form, is written $Ax = b$ as below:

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

we see that the rows of the matrix A add up to 0. In fact, it is easy to convince ourselves that the left nullspace of A is spanned by $y = (1, 1, 1)$, and so the system is solvable iff $y^\top b = 0$, namely

$$b_1 + b_2 + b_3 = 0.$$

Note that the above criterion can also be stated negatively as follows:

The equation $Ax = b$ has no solution iff there is some $y \in \mathbb{R}^m$ such that $A^\top y = 0$ and $y^\top b \neq 0$.

Since $A^\top y = 0$ iff $y^\top A = 0$, we can view $y^\top$ as a row vector representing a linear form, and $y^\top A = 0$ asserts that the linear form $y^\top$ vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form $y^\top$ defines the hyperplane H of equation $y^\top z = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a hyperplane H containing $A^1, \ldots, A^n$ and not containing b.

11.9 Summary

The main concepts and results of this chapter are listed below:

- The *dual space* E^* and *linear forms* (*covector*). The *bidual* E^{**}.
- The *bilinear pairing* $\langle -, - \rangle \colon E^* \times E \to K$ (the *canonical pairing*).
- *Evaluation at* v: $\mathrm{eval}_v \colon E^* \to K$.
- The map $\mathrm{eval}_E \colon E \to E^{**}$.
- *Othogonality* between a subspace V of E and a subspace U of E^*; the *orthogonal* V^0 and the *orthogonal* U^0.
- *Coordinate forms*.
- The *Duality theorem* (Theorem 11.4).
- The *dual basis* of a basis.
- The isomorphism $\mathrm{eval}_E \colon E \to E^{**}$ when $\dim(E)$ is finite.
- *Pairing* between two vector spaces; *nondegenerate pairing*; Proposition 11.6.
- Hyperplanes and linear forms.
- The *transpose* $f^\top \colon F^* \to E^*$ of a linear map $f \colon E \to F$.
- The fundamental identities:

  $$\mathrm{Ker}\, f^\top = (\mathrm{Im}\, f)^0 \quad \text{and} \quad \mathrm{Ker}\, f = (\mathrm{Im}\, f^\top)^0$$

  (Proposition 11.11).
- If F is finite-dimensional, then

  $$\mathrm{rk}(f) = \mathrm{rk}(f^\top).$$

  (Theorem 11.12).

- The matrix of the transpose map $f^\top$ is equal to the transpose of the matrix of the map f (Proposition 11.14).
- For any $m \times n$ matrix A,
$$\mathrm{rk}(A) = \mathrm{rk}(A^\top).$$
- Characterization of the rank of a matrix in terms of a maximal invertible submatrix (Proposition 11.16).
- The *four fundamental subspaces*:
$$\mathrm{Im}\, f,\ \mathrm{Im}\, f^\top,\ \mathrm{Ker}\, f,\ \mathrm{Ker}\, f^\top.$$
- The *column space*, the *nullspace*, the *row space*, and the *left nullspace* (of a matrix).
- Criterion for the solvability of an equation of the form $Ax = b$ in terms of the left nullspace.

11.10 Problems

Problem 11.1. Prove the following properties of transposition:

$$\begin{aligned} (f+g)^\top &= f^\top + g^\top \\ (g \circ f)^\top &= f^\top \circ g^\top \\ \mathrm{id}_E^\top &= \mathrm{id}_{E^*}. \end{aligned}$$

Problem 11.2. Let $(u_1, \ldots, u_{n-1})$ be $n-1$ linearly independent vectors $u_i \in \mathbb{C}^n$. Prove that the hyperlane H spanned by $(u_1, \ldots, u_{n-1})$ is the nullspace of the linear form

$$x \mapsto \det(u_1, \ldots, u_{n-1}, x), \quad x \in \mathbb{C}^n.$$

Prove that if A is the $n \times n$ matrix whose columns are $(u_1, \ldots, u_{n-1}, x)$, and if $c_i = (-1)^{i+n} \det(A_{in})$ is the cofactor of $a_{in} = x_i$ for $i = 1, \ldots, n$, then H is defined by the equation

$$c_1 x_1 + \cdots + c_n x_n = 0.$$

Problem 11.3. (1) Let $\varphi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the map defined by

$$\varphi((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = x_1 y_1 + \cdots + x_n y_n.$$

Prove that φ is a bilinear nondegenerate pairing. Deduce that $(\mathbb{R}^n)^*$ is isomorphic to $\mathbb{R}^n$.

Prove that $\varphi(x, x) = 0$ iff $x = 0$.

(2) Let $\varphi_L\colon \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$ be the map defined by

$$\varphi_L((x_1, x_2, x_3, x_4), (y_1, y_2, y_3, , y_4)) = x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4.$$

Prove that φ is a bilinear nondegenerate pairing.

Show that there exist nonzero vectors $x \in \mathbb{R}^4$ such that $\varphi_L(x, x) = 0$.

Remark: The vector space $\mathbb{R}^4$ equipped with the above bilinear form called the *Lorentz form* is called *Minkowski space*.

Problem 11.4. Given any two subspaces V_1, V_2 of a finite-dimensional vector space E, prove that

$$\begin{aligned}(V_1 + V_2)^0 &= V_1^0 \cap V_2^0 \\ (V_1 \cap V_2)^0 &= V_1^0 + V_2^0.\end{aligned}$$

Beware that in the second equation, V_1 and V_2 are subspaces of E, not E^*.

Hint. To prove the second equation, prove the inclusions $V_1^0 + V_2^0 \subseteq (V_1 \cap V_2)^0$ and $(V_1 \cap V_2)^0 \subseteq V_1^0 + V_2^0$. Proving the second inclusion is a little tricky. First, prove that we can pick a subspace W_1 of V_1 and a subspace W_2 of V_2 such that

1. V_1 is the direct sum $V_1 = (V_1 \cap V_2) \oplus W_1$.
2. V_2 is the direct sum $V_2 = (V_1 \cap V_2) \oplus W_2$.
3. $V_1 + V_2$ is the direct sum $V_1 + V_2 = (V_1 \cap V_2) \oplus W_1 \oplus W_2$.

Problem 11.5. (1) Let A be any $n \times n$ matrix such that the sum of the entries of every row of A is the same (say c_1), and the sum of entries of every column of A is the same (say c_2). Prove that $c_1 = c_2$.

(2) Prove that for any $n \geq 2$, the $2n - 2$ equations asserting that the sum of the entries of every row of A is the same, and the sum of entries of every column of A is the same are lineary independent. For example, when $n = 4$, we have the following 6 equations

$$\begin{aligned}
a_{11} + a_{12} + a_{13} + a_{14} - a_{21} - a_{22} - a_{23} - a_{24} &= 0 \\
a_{21} + a_{22} + a_{23} + a_{24} - a_{31} - a_{32} - a_{33} - a_{34} &= 0 \\
a_{31} + a_{32} + a_{33} + a_{34} - a_{41} - a_{42} - a_{43} - a_{44} &= 0 \\
a_{11} + a_{21} + a_{31} + a_{41} - a_{12} - a_{22} - a_{32} - a_{42} &= 0 \\
a_{12} + a_{22} + a_{32} + a_{42} - a_{13} - a_{23} - a_{33} - a_{43} &= 0 \\
a_{13} + a_{23} + a_{33} + a_{43} - a_{14} - a_{24} - a_{34} - a_{44} &= 0.
\end{aligned}$$

Hint. Group the equations as above; that is, first list the $n - 1$ equations relating the rows, and then list the $n - 1$ equations relating the columns. Prove that the first $n - 1$ equations are linearly independent, and that the last $n-1$ equations are also linearly independent. Then, find a relationship between the two groups of equations that will allow you to prove that they span subspace V^r and V^c such that $V^r \cap V^c = (0)$.

(3) Now consider *magic squares*. Such matrices satisfy the two conditions about the sum of the entries in each row and in each column to be the same number, and also the additional two constraints that the main descending and the main ascending diagonals add up to this common number. Traditionally, it is also required that the entries in a magic square are positive integers, but we will consider generalized magic square with arbitrary real entries. For example, in the case $n = 4$, we have the following system of 8 equations:

$$\begin{aligned}
a_{11} + a_{12} + a_{13} + a_{14} - a_{21} - a_{22} - a_{23} - a_{24} &= 0\\
a_{21} + a_{22} + a_{23} + a_{24} - a_{31} - a_{32} - a_{33} - a_{34} &= 0\\
a_{31} + a_{32} + a_{33} + a_{34} - a_{41} - a_{42} - a_{43} - a_{44} &= 0\\
a_{11} + a_{21} + a_{31} + a_{41} - a_{12} - a_{22} - a_{32} - a_{42} &= 0\\
a_{12} + a_{22} + a_{32} + a_{42} - a_{13} - a_{23} - a_{33} - a_{43} &= 0\\
a_{13} + a_{23} + a_{33} + a_{43} - a_{14} - a_{24} - a_{34} - a_{44} &= 0\\
a_{22} + a_{33} + a_{44} - a_{12} - a_{13} - a_{14} &= 0\\
a_{41} + a_{32} + a_{23} - a_{11} - a_{12} - a_{13} &= 0.
\end{aligned}$$

In general, the equation involving the descending diagonal is

$$a_{22} + a_{33} + \cdots + a_{nn} - a_{12} - a_{13} - \cdots - a_{1n} = 0 \tag{r}$$

and the equation involving the ascending diagonal is

$$a_{n1} + a_{n-12} + \cdots + a_{2n-1} - a_{11} - a_{12} - \cdots - a_{1n-1} = 0. \tag{c}$$

Prove that if $n \geq 3$, then the $2n$ equations asserting that a matrix is a generalized magic square are linearly independent.

Hint. Equations are really linear forms, so find some matrix annihilated by all equations except equation r, and some matrix annihilated by all equations except equation c.

Problem 11.6. Let $U_1, \ldots, U_p$ be some subspaces of a vector space E, and assume that they form a direct sum $U = U_1 \oplus \cdots \oplus U_p$. Let $j_i \colon U_i \to U_1 \oplus \cdots \oplus U_p$ be the canonical injections, and let $\pi_i \colon U_1^* \times \cdots \times U_p^* \to U_i^*$ be the canonical projections. Prove that there is an isomorphism f from $(U_1 \oplus \cdots \oplus U_p)^*$ to $U_1^* \times \cdots \times U_p^*$ such that

$$\pi_i \circ f = j_i^{\top}, \quad 1 \leq i \leq p.$$

Problem 11.7. Let U and V be two subspaces of a vector space E such that $E = U \oplus V$. Prove that

$$E^* = U^0 \oplus V^0.$$

Chapter 12

Euclidean Spaces

Rien n'est beau que le vrai.
—Hermann Minkowski

12.1 Inner Products, Euclidean Spaces

So far the framework of vector spaces allows us to deal with ratios of vectors and linear combinations, but there is no way to express the notion of angle or to talk about orthogonality of vectors. A Euclidean structure allows us to deal with *metric notions* such as angles, orthogonality, and length (or distance).

This chapter covers the bare bones of Euclidean geometry. Deeper aspects of Euclidean geometry are investigated in Chapter 13. One of our main goals is to give the basic properties of the transformations that preserve the Euclidean structure, rotations and reflections, since they play an important role in practice. Euclidean geometry is the study of properties invariant under certain affine maps called *rigid motions*. Rigid motions are the maps that preserve the distance between points.

We begin by defining inner products and Euclidean spaces. The Cauchy–Schwarz inequality and the Minkowski inequality are shown. We define orthogonality of vectors and of subspaces, orthogonal bases, and orthonormal bases. We prove that every finite-dimensional Euclidean space has orthonormal bases. The first proof uses duality and the second one the Gram–Schmidt orthogonalization procedure. The QR-decomposition for invertible matrices is shown as an application of the Gram–Schmidt procedure. Linear isometries (also called orthogonal transformations) are defined and studied briefly. We conclude with a short section in which some applications of Euclidean geometry are sketched. One of the most important applications, the method of least squares, is discussed in Chapter 23.

For a more detailed treatment of Euclidean geometry see Berger [11, 12], Snapper and Troyer [162], or any other book on geometry, such as Pedoe [136], Coxeter [44], Fresnel [65], Tisseron [175], or Cagnac, Ramis, and Commeau [32]. Serious readers should consult Emil

Artin's famous book [6], which contains an in-depth study of the orthogonal group, as well as other groups arising in geometry. It is still worth consulting some of the older classics, such as Hadamard [84, 85] and Rouché and de Comberousse [139]. The first edition of [84] was published in 1898 and finally reached its thirteenth edition in 1947! In this chapter it is assumed that all vector spaces are defined over the field $\mathbb{R}$ of real numbers unless specified otherwise (in a few cases, over the complex numbers $\mathbb{C}$).

First we define a Euclidean structure on a vector space. Technically, a Euclidean structure over a vector space E is provided by a symmetric bilinear form on the vector space satisfying some extra properties. Recall that a bilinear form $\varphi \colon E \times E \to \mathbb{R}$ is *definite* if for every $u \in E$, $u \neq 0$ implies that $\varphi(u, u) \neq 0$, and *positive* if for every $u \in E$, $\varphi(u, u) \geq 0$.

Definition 12.1. A *Euclidean space* is a real vector space E equipped with a symmetric bilinear form $\varphi \colon E \times E \to \mathbb{R}$ that is *positive definite*. More explicitly, $\varphi \colon E \times E \to \mathbb{R}$ satisfies the following axioms:

$$\begin{aligned} \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v), \\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2), \\ \varphi(\lambda u, v) &= \lambda\varphi(u, v), \\ \varphi(u, \lambda v) &= \lambda\varphi(u, v), \\ \varphi(u, v) &= \varphi(v, u), \\ u &\neq 0 \text{ implies that } \varphi(u, u) > 0. \end{aligned}$$

The real number $\varphi(u, v)$ is also called the *inner product (or scalar product) of u and v*. We also define the *quadratic form associated with φ* as the function $\Phi \colon E \to \mathbb{R}_+$ such that

$$\Phi(u) = \varphi(u, u),$$

for all $u \in E$.

Since φ is bilinear, we have $\varphi(0, 0) = 0$, and since it is positive definite, we have the stronger fact that

$$\varphi(u, u) = 0 \quad \text{iff} \quad u = 0,$$

that is, $\Phi(u) = 0$ iff $u = 0$.

Given an inner product $\varphi \colon E \times E \to \mathbb{R}$ on a vector space E, we also denote $\varphi(u, v)$ by

$$u \cdot v \quad \text{or} \quad \langle u, v\rangle \quad \text{or} \quad (u|v),$$

and $\sqrt{\Phi(u)}$ by $\|u\|$.

Example 12.1. The standard example of a Euclidean space is $\mathbb{R}^n$, under the inner product $\cdot$ defined such that

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

This Euclidean space is denoted by $\mathbb{E}^n$.

There are other examples.

Example 12.2. For instance, let E be a vector space of dimension 2, and let (e_1, e_2) be a basis of E. If $a > 0$ and $b^2 - ac < 0$, the bilinear form defined such that

$$\varphi(x_1e_1 + y_1e_2,\, x_2e_1 + y_2e_2) = ax_1x_2 + b(x_1y_2 + x_2y_1) + cy_1y_2$$

yields a Euclidean structure on E. In this case,

$$\Phi(xe_1 + ye_2) = ax^2 + 2bxy + cy^2.$$

Example 12.3. Let $\mathcal{C}[a, b]$ denote the set of continuous functions $f\colon [a, b] \to \mathbb{R}$. It is easily checked that $\mathcal{C}[a, b]$ is a vector space of infinite dimension. Given any two functions $f, g \in \mathcal{C}[a, b]$, let

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt.$$

We leave it as an easy exercise that $\langle -, -\rangle$ is indeed an inner product on $\mathcal{C}[a, b]$. In the case where $a = -\pi$ and $b = \pi$ (or $a = 0$ and $b = 2\pi$, this makes basically no difference), one should compute

$$\langle \sin px, \sin qx\rangle, \quad \langle \sin px, \cos qx\rangle, \quad \text{and} \quad \langle \cos px, \cos qx\rangle,$$

for all natural numbers $p, q \geq 1$. The outcome of these calculations is what makes Fourier analysis possible!

Example 12.4. Let $E = \mathrm{M}_n(\mathbb{R})$ be the vector space of real $n \times n$ matrices. If we view a matrix $A \in \mathrm{M}_n(\mathbb{R})$ as a "long" column vector obtained by concatenating together its columns, we can define the inner product of two matrices $A, B \in \mathrm{M}_n(\mathbb{R})$ as

$$\langle A, B\rangle = \sum_{i,j=1}^{n} a_{ij}b_{ij},$$

which can be conveniently written as

$$\langle A, B\rangle = \mathrm{tr}(A^\top B) = \mathrm{tr}(B^\top A).$$

Since this can be viewed as the Euclidean product on $\mathbb{R}^{n^2}$, it is an inner product on $\mathrm{M}_n(\mathbb{R})$. The corresponding norm

$$\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$$

is the Frobenius norm (see Section 9.2).

Let us observe that φ can be recovered from Φ.

Proposition 12.1. *We have*

$$\varphi(u,\, v) = \frac{1}{2}[\Phi(u+v) - \Phi(u) - \Phi(v)]$$

for all $u, v \in E$. We say that φ is the **polar form of** Φ.

Proof. By bilinearity and symmetry, we have

$$\begin{aligned}\Phi(u+v) &= \varphi(u+v,\, u+v)\\ &= \varphi(u,\, u+v) + \varphi(v,\, u+v)\\ &= \varphi(u,\, u) + 2\varphi(u,\, v) + \varphi(v,\, v)\\ &= \Phi(u) + 2\varphi(u,\, v) + \Phi(v).\end{aligned}$$

□

If E is finite-dimensional and if $\varphi\colon E \times E \to \mathbb{R}$ is a bilinear form on E, given any basis $(e_1, \ldots, e_n)$ of E, we can write $x = \sum_{i=1}^n x_i e_i$ and $y = \sum_{j=1}^n y_j e_j$, and we have

$$\varphi(x, y) = \varphi\left(\sum_{i=1}^n x_i e_i, \sum_{j=1}^n y_j e_j\right) = \sum_{i,j=1}^n x_i y_j \varphi(e_i, e_j).$$

If we let G be the matrix $G = (\varphi(e_i, e_j))$, and if x and y are the column vectors associated with $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, then we can write

$$\varphi(x, y) = x^\top G y = y^\top G^\top x.$$

Note that we are committing an abuse of notation since $x = \sum_{i=1}^n x_i e_i$ is a vector in E, but the column vector associated with $(x_1, \ldots, x_n)$ belongs to $\mathbb{R}^n$. To avoid this minor abuse, we could denote the column vector associated with $(x_1, \ldots, x_n)$ by $\mathbf{x}$ (and similarly $\mathbf{y}$ for the column vector associated with $(y_1, \ldots, y_n)$), in wich case the "correct" expression for $\varphi(x, y)$ is

$$\varphi(x, y) = \mathbf{x}^\top G \mathbf{y}.$$

However, in view of the isomorphism between E and $\mathbb{R}^n$, to keep notation as simple as possible, we will use x and y instead of $\mathbf{x}$ and $\mathbf{y}$.

Also observe that φ is symmetric iff $G = G^\top$, and φ is positive definite iff the matrix G is positive definite, that is,

$$x^\top G x > 0 \quad \text{for all } x \in \mathbb{R}^n,\ x \neq 0.$$

The matrix G associated with an inner product is called the *Gram matrix* of the inner product with respect to the basis $(e_1, \ldots, e_n)$.

Conversely, if A is a symmetric positive definite $n \times n$ matrix, it is easy to check that the bilinear form

$$\langle x, y \rangle = x^\top A y$$

is an inner product. If we make a change of basis from the basis $(e_1, \ldots, e_n)$ to the basis $(f_1, \ldots, f_n)$, and if the change of basis matrix is P (where the jth column of P consists of the coordinates of f_j over the basis $(e_1, \ldots, e_n)$), then with respect to coordinates x' and y' over the basis $(f_1, \ldots, f_n)$, we have

$$x^\top G y = x'^\top P^\top G P y',$$

so the matrix of our inner product over the basis $(f_1, \ldots, f_n)$ is $P^\top G P$. We summarize these facts in the following proposition.

Proposition 12.2. *Let E be a finite-dimensional vector space, and let $(e_1, \ldots, e_n)$ be a basis of E.*

1. *For any inner product $\langle -, - \rangle$ on E, if $G = (\langle e_i, e_j \rangle)$ is the Gram matrix of the inner product $\langle -, - \rangle$ w.r.t. the basis $(e_1, \ldots, e_n)$, then G is symmetric positive definite.*

2. *For any change of basis matrix P, the Gram matrix of $\langle -, - \rangle$ with respect to the new basis is $P^\top G P$.*

3. *If A is any $n \times n$ symmetric positive definite matrix, then*

$$\langle x, y \rangle = x^\top A y$$

is an inner product on E.

We will see later that a symmetric matrix is positive definite iff its eigenvalues are all positive.

One of the very important properties of an inner product φ is that the map $u \mapsto \sqrt{\Phi(u)}$ is a norm.

Proposition 12.3. *Let E be a Euclidean space with inner product φ, and let Φ be the corresponding quadratic form. For all $u, v \in E$, we have the* Cauchy–Schwarz inequality

$$\varphi(u, v)^2 \leq \Phi(u)\Phi(v),$$

the equality holding iff u and v are linearly dependent.

We also have the Minkowski inequality

$$\sqrt{\Phi(u + v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)},$$

the equality holding iff u and v are linearly dependent, where in addition if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some $\lambda > 0$.

Proof. For any vectors $u, v \in E$, we define the function $T\colon \mathbb{R} \to \mathbb{R}$ such that

$$T(\lambda) = \Phi(u + \lambda v),$$

for all $\lambda \in \mathbb{R}$. Using bilinearity and symmetry, we have

$$\begin{aligned}
\Phi(u + \lambda v) &= \varphi(u + \lambda v,\, u + \lambda v) \\
&= \varphi(u,\, u + \lambda v) + \lambda\varphi(v,\, u + \lambda v) \\
&= \varphi(u,\, u) + 2\lambda\varphi(u,\, v) + \lambda^2\varphi(v,\, v) \\
&= \Phi(u) + 2\lambda\varphi(u,\, v) + \lambda^2\Phi(v).
\end{aligned}$$

Since φ is positive definite, Φ is nonnegative, and thus $T(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$. If $\Phi(v) = 0$, then $v = 0$, and we also have $\varphi(u,\, v) = 0$. In this case, the Cauchy–Schwarz inequality is trivial, and $v = 0$ and u are linearly dependent.

Now assume $\Phi(v) > 0$. Since $T(\lambda) \geq 0$, the quadratic equation

$$\lambda^2\Phi(v) + 2\lambda\varphi(u,\, v) + \Phi(u) = 0$$

cannot have distinct real roots, which means that its discriminant

$$\Delta = 4(\varphi(u,\, v)^2 - \Phi(u)\Phi(v))$$

is null or negative, which is precisely the Cauchy–Schwarz inequality

$$\varphi(u, v)^2 \leq \Phi(u)\Phi(v).$$

Let us now consider the case where we have the equality

$$\varphi(u, v)^2 = \Phi(u)\Phi(v).$$

There are two cases. If $\Phi(v) = 0$, then $v = 0$ and u and v are linearly dependent. If $\Phi(v) \neq 0$, then the above quadratic equation has a double root λ_0, and we have $\Phi(u + \lambda_0 v) = 0$. Since φ is positive definite, $\Phi(u + \lambda_0 v) = 0$ implies that $u + \lambda_0 v = 0$, which shows that u and v are linearly dependent. Conversely, it is easy to check that we have equality when u and v are linearly dependent.

The Minkowski inequality

$$\sqrt{\Phi(u + v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

is equivalent to

$$\Phi(u + v) \leq \Phi(u) + \Phi(v) + 2\sqrt{\Phi(u)\Phi(v)}.$$

However, we have shown that

$$2\varphi(u,\, v) = \Phi(u + v) - \Phi(u) - \Phi(v),$$

and so the above inequality is equivalent to

$$\varphi(u,\, v) \leq \sqrt{\Phi(u)\Phi(v)},$$

which is trivial when $\varphi(u,\, v) \leq 0$, and follows from the Cauchy–Schwarz inequality when $\varphi(u,\, v) \geq 0$. Thus, the Minkowski inequality holds. Finally assume that $u \neq 0$ and $v \neq 0$, and that

$$\sqrt{\Phi(u+v)} = \sqrt{\Phi(u)} + \sqrt{\Phi(v)}.$$

When this is the case, we have

$$\varphi(u,\, v) = \sqrt{\Phi(u)\Phi(v)},$$

and we know from the discussion of the Cauchy–Schwarz inequality that the equality holds iff u and v are linearly dependent. The Minkowski inequality is an equality when u or v is null. Otherwise, if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some $\lambda \neq 0$, and since

$$\varphi(u,\, v) = \lambda\varphi(v,\, v) = \sqrt{\Phi(u)\Phi(v)},$$

by positivity, we must have $\lambda > 0$. □

Note that the Cauchy–Schwarz inequality can also be written as

$$|\varphi(u, v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Remark: It is easy to prove that the Cauchy–Schwarz and the Minkowski inequalities still hold for a symmetric bilinear form that is positive, but not necessarily definite (i.e., $\varphi(u, v) \geq 0$ for all $u, v \in E$). However, u and v need not be linearly dependent when the equality holds.

The Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

shows that the map $u \mapsto \sqrt{\Phi(u)}$ satisfies the convexity inequality (also known as triangle inequality), condition (N3) of Definition 9.1, and since φ is bilinear and positive definite, it also satisfies conditions (N1) and (N2) of Definition 9.1, and thus it is a *norm* on E. The norm induced by φ is called the *Euclidean norm induced by* φ.

The Cauchy–Schwarz inequality can be written as

$$|u \cdot v| \leq \|u\|\|v\|,$$

and the Minkowski inequality as

$$\|u + v\| \leq \|u\| + \|v\|.$$

If u and v are nonzero vectors then the Cauchy–Schwarz inequality implies that

$$-1 \leq \frac{u \cdot v}{\|u\| \, \|v\|} \leq +1.$$

Then there is a unique $\theta \in [0, \pi]$ such that

$$\cos \theta = \frac{u \cdot v}{\|u\| \, \|v\|}.$$

We have $u = v$ iff $\theta = 0$ and $u = -v$ iff $\theta = \pi$. For $0 < \theta < \pi$, the vectors u and v are linearly independent and there is an orientation of the plane spanned by u and v such that θ is the angle between u and v. See Problem 12.8 for the precise notion of orientation. If u is a unit vector (which means that $\|u\| = 1$), then the vector

$$(\|v\| \cos \theta) u = (u \cdot v) u = (v \cdot u) u$$

is called the *orthogonal projection* of v onto the space spanned by u.

Remark: One might wonder if every norm on a vector space is induced by some Euclidean inner product. In general this is false, but remarkably, there is a simple necessary and sufficient condition, which is that the norm must satisfy the *parallelogram law*:

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

See Figure 12.1.

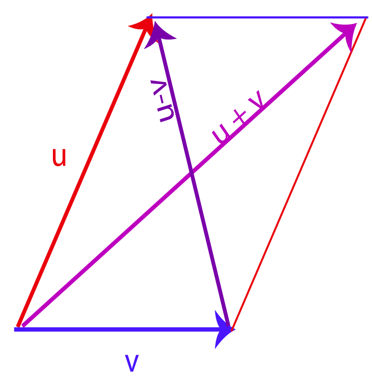


Figure 12.1: The parallelogram law states that the sum of the lengths of the diagonals of the parallelogram determined by vectors u and v equals the sum of all the sides.

If $\langle -, - \rangle$ is an inner product, then we have

$$\begin{aligned}
\|u + v\|^2 &= \|u\|^2 + \|v\|^2 + 2\langle u, v \rangle \\
\|u - v\|^2 &= \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle,
\end{aligned}$$

and by adding and subtracting these identities, we get the parallelogram law and the equation

$$\langle u, v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2),$$

which allows us to recover $\langle -, -\rangle$ from the norm.

Conversely, if $\| \ \|$ is a norm satisfying the parallelogram law, and if it comes from an inner product, then this inner product must be given by

$$\langle u, v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2).$$

We need to prove that the above form is indeed symmetric and bilinear.

Symmetry holds because $\|u-v\| = \|-(u-v)\| = \|v-u\|$. Let us prove additivity in the variable u. By the parallelogram law, we have

$$2(\|x+z\|^2 + \|y\|^2) = \|x+y+z\|^2 + \|x-y+z\|^2$$

which yields

$$\begin{aligned}
\|x+y+z\|^2 &= 2(\|x+z\|^2 + \|y\|^2) - \|x-y+z\|^2 \\
\|x+y+z\|^2 &= 2(\|y+z\|^2 + \|x\|^2) - \|y-x+z\|^2,
\end{aligned}$$

where the second formula is obtained by swapping x and y. Then by adding up these equations, we get

$$\|x+y+z\|^2 = \|x\|^2 + \|y\|^2 + \|x+z\|^2 + \|y+z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|y-x+z\|^2.$$

Replacing z by $-z$ in the above equation, we get

$$\|x+y-z\|^2 = \|x\|^2 + \|y\|^2 + \|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|x-y-z\|^2 - \frac{1}{2}\|y-x-z\|^2,$$

Since $\|x-y+z\| = \|-(x-y+z)\| = \|y-x-z\|$ and $\|y-x+z\| = \|-(y-x+z)\| = \|x-y-z\|$, by subtracting the last two equations, we get

$$\begin{aligned}
\langle x+y, z\rangle &= \frac{1}{4}(\|x+y+z\|^2 - \|x+y-z\|^2) \\
&= \frac{1}{4}(\|x+z\|^2 - \|x-z\|^2) + \frac{1}{4}(\|y+z\|^2 - \|y-z\|^2) \\
&= \langle x, z\rangle + \langle y, z\rangle,
\end{aligned}$$

as desired.

Proving that

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{R}$$

is a little tricky. The strategy is to prove the identity for $\lambda \in \mathbb{Z}$, then to promote it to $\mathbb{Q}$, and then to $\mathbb{R}$ by continuity.

Since

$$\begin{aligned}\langle -u, v\rangle &= \frac{1}{4}(\|-u+v\|^2 - \|-u-v\|^2)\\ &= \frac{1}{4}(\|u-v\|^2 - \|u+v\|^2)\\ &= -\langle u, v\rangle,\end{aligned}$$

the property holds for $\lambda = -1$. By linearity and by induction, for any $n \in \mathbb{N}$ with $n \geq 1$, writing $n = n - 1 + 1$, we get

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{N},$$

and since the above also holds for $\lambda = -1$, it holds for all $\lambda \in \mathbb{Z}$. For $\lambda = p/q$ with $p, q \in \mathbb{Z}$ and $q \neq 0$, we have

$$q\langle (p/q)u, v\rangle = \langle pu, v\rangle = p\langle u, v\rangle,$$

which shows that

$$\langle (p/q)u, v\rangle = (p/q)\langle u, v\rangle,$$

and thus

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \text{for all } \lambda \in \mathbb{Q}.$$

To finish the proof, we use the fact that a norm is a continuous map $x \mapsto \|x\|$. Then, the continuous function $t \mapsto \frac{1}{t}\langle tu, v\rangle$ defined on $\mathbb{R} - \{0\}$ agrees with $\langle u, v\rangle$ on $\mathbb{Q} - \{0\}$, so it is equal to $\langle u, v\rangle$ on $\mathbb{R} - \{0\}$. The case $\lambda = 0$ is trivial, so we are done.

We now define orthogonality.

12.2 Orthogonality and Duality in Euclidean Spaces

An inner product on a vector space gives the ability to define the notion of orthogonality. Families of nonnull pairwise orthogonal vectors must be linearly independent. They are called orthogonal families. In a vector space of finite dimension it is always possible to find orthogonal bases. This is very useful theoretically and practically. Indeed, in an orthogonal basis, finding the coordinates of a vector is very cheap: It takes an inner product. Fourier series make crucial use of this fact. When E has finite dimension, we prove that the inner product on E induces a natural isomorphism between E and its dual space E^*. This allows us to define the adjoint of a linear map in an intrinsic fashion (i.e., independently of bases). It is also possible to orthonormalize any basis (certainly when the dimension is finite). We give two proofs, one using duality, the other more constructive using the Gram–Schmidt orthonormalization procedure.

Definition 12.2. Given a Euclidean space E, any two vectors $u, v \in E$ are *orthogonal, or perpendicular*, if $u \cdot v = 0$. Given a family $(u_i)_{i\in I}$ of vectors in E, we say that $(u_i)_{i\in I}$ is *orthogonal* if $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$. We say that the family $(u_i)_{i\in I}$ is *orthonormal* if $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$, and $\|u_i\| = u_i \cdot u_i = 1$, for all $i \in I$. For any subset F of E, the set

$$F^{\perp} = \{v \in E \mid u \cdot v = 0, \text{ for all } u \in F\},$$

of all vectors orthogonal to all vectors in F, is called the *orthogonal complement of F*.

Since inner products are positive definite, observe that for any vector $u \in E$, we have

$$u \cdot v = 0 \quad \text{for all } v \in E \quad \text{iff} \quad u = 0.$$

It is immediately verified that the orthogonal complement $F^{\perp}$ of F is a subspace of E.

Example 12.5. Going back to Example 12.3 and to the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$$

on the vector space $\mathcal{C}[-\pi, \pi]$, it is easily checked that

$$\langle \sin px, \sin qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 1, \end{cases}$$

$$\langle \cos px, \cos qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 0, \end{cases}$$

and

$$\langle \sin px, \cos qx\rangle = 0,$$

for all $p \geq 1$ and $q \geq 0$, and of course, $\langle 1, 1\rangle = \int_{-\pi}^{\pi} dx = 2\pi$.

As a consequence, the family $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ is orthogonal. It is not orthonormal, but becomes so if we divide every trigonometric function by $\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$.

Proposition 12.4. *Given a Euclidean space E, for any family $(u_i)_{i\in I}$ of nonnull vectors in E, if $(u_i)_{i\in I}$ is orthogonal, then it is linearly independent.*

Proof. Assume there is a linear dependence

$$\sum_{j\in J} \lambda_j u_j = 0$$

for some $\lambda_j \in \mathbb{R}$ and some finite subset J of I. By taking the inner product with u_i for any $i \in J$, and using the the bilinearity of the inner product and the fact that $u_i \cdot u_j = 0$ whenever $i \neq j$, we get

$$\begin{aligned} 0 = u_i \cdot 0 = u_i \cdot \left(\sum_{j \in J} \lambda_j u_j \right) \\ = \sum_{j \in J} \lambda_j (u_i \cdot u_j) = \lambda_i (u_i \cdot u_i), \end{aligned}$$

so

$$\lambda_i (u_i \cdot u_i) = 0, \qquad \text{for all } i \in J,$$

and since $u_i \neq 0$ and an inner product is positive definite, $u_i \cdot u_i \neq 0$, so we obtain

$$\lambda_i = 0, \qquad \text{for all } i \in J,$$

which shows that the family $(u_i)_{i \in I}$ is linearly independent. $\square$

We leave the following simple result as an exercise.

Proposition 12.5. *Given a Euclidean space E, any two vectors $u, v \in E$ are orthogonal iff*

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

See Figure 12.2 for a geometrical interpretation.

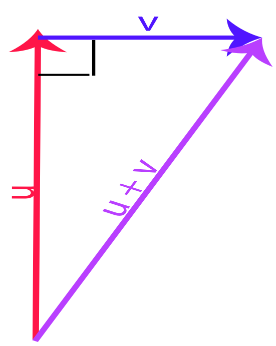


Figure 12.2: The sum of the lengths of the two sides of a right triangle is equal to the length of the hypotenuse; i.e. the Pythagorean theorem.

One of the most useful features of orthonormal bases is that they afford a very simple method for computing the coordinates of a vector over any basis vector. Indeed, assume that $(e_1, \ldots, e_m)$ is an orthonormal basis. For any vector

$$x = x_1 e_1 + \cdots + x_m e_m,$$

if we compute the inner product $x \cdot e_i$, we get

$$x \cdot e_i = x_1 e_1 \cdot e_i + \cdots + x_i e_i \cdot e_i + \cdots + x_m e_m \cdot e_i = x_i,$$

since

$$e_i \cdot e_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

is the property characterizing an orthonormal family. Thus,

$$x_i = x \cdot e_i,$$

which means that $x_i e_i = (x \cdot e_i) e_i$ is the orthogonal projection of x onto the subspace generated by the basis vector e_i. See Figure 12.3. If the basis is orthogonal but not necessarily

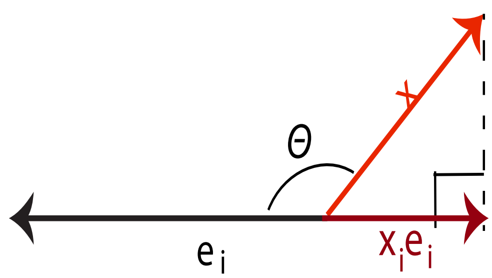


Figure 12.3: The orthogonal projection of the red vector x onto the black basis vector e_i is the maroon vector $x_i e_i$. Observe that $x \cdot e_i = \|x\| \cos\theta$.

orthonormal, then

$$x_i = \frac{x \cdot e_i}{e_i \cdot e_i} = \frac{x \cdot e_i}{\|e_i\|^2}.$$

All this is true even for an infinite orthonormal (or orthogonal) basis $(e_i)_{i \in I}$.

However, remember that every vector x is expressed as a linear combination

$$x = \sum_{i \in I} x_i e_i$$

where the family of scalars $(x_i)_{i \in I}$ has **finite support**, which means that $x_i = 0$ for all $i \in I - J$, where J is a finite set. Thus, even though the family $(\sin px)_{p \geq 1} \cup (\cos qx)_{q \geq 0}$ is orthogonal (it is not orthonormal, but becomes so if we divide every trigonometric function by

$\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$; we won't because it looks messy!), the fact that a function $f \in \mathcal{C}^0[-\pi, \pi]$ can be written as a Fourier series as

$$f(x) = a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx)$$

does not mean that $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ is a basis of this vector space of functions, because in general, the families (a_k) and (b_k) **do not** have finite support! In order for this infinite linear combination to make sense, it is necessary to prove that the partial sums

$$a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx)$$

of the series converge to a limit when n goes to infinity. This requires a topology on the space.

A very important property of Euclidean spaces of finite dimension is that the inner product induces a canonical bijection (i.e., independent of the choice of bases) between the vector space E and its dual E^*. The reason is that an inner product $\cdot\colon E \times E \to \mathbb{R}$ defines a nondegenerate pairing, as defined in Definition 11.4. Indeed, if $u \cdot v = 0$ for all $v \in E$ then $u = 0$, and similarly if $u \cdot v = 0$ for all $u \in E$ then $v = 0$ (since an inner product is positive definite and symmetric). By Proposition 11.6, there is a canonical isomorphism between E and E^*. We feel that the reader will appreciate if we exhibit this mapping explicitly and reprove that it is an isomorphism.

The mapping from E to E^* is defined as follows.

Definition 12.3. For any vector $u \in E$, let $\varphi_u\colon E \to \mathbb{R}$ be the map defined such that

$$\varphi_u(v) = u \cdot v, \quad \text{for all } v \in E.$$

Since the inner product is bilinear, the map φ_u is a linear form in E^*. Thus, we have a map $\flat\colon E \to E^*$, defined such that

$$\flat(u) = \varphi_u.$$

Theorem 12.6. *Given a Euclidean space E, the map $\flat\colon E \to E^*$ defined such that*

$$\flat(u) = \varphi_u$$

is linear and injective. When E is also of finite dimension, the map $\flat\colon E \to E^*$ is a canonical isomorphism.

Proof. That $\flat\colon E \to E^*$ is a linear map follows immediately from the fact that the inner product is bilinear. If $\varphi_u = \varphi_v$, then $\varphi_u(w) = \varphi_v(w)$ for all $w \in E$, which by definition of φ_u means that $u \cdot w = v \cdot w$ for all $w \in E$, which by bilinearity is equivalent to

$$(v - u) \cdot w = 0$$

for all $w \in E$, which implies that $u = v$, since the inner product is positive definite. Thus, $\flat\colon E \to E^*$ is injective. Finally, when E is of finite dimension n, we know that E^* is also of dimension n, and then $\flat\colon E \to E^*$ is bijective. $\square$

The inverse of the isomorphism $\flat\colon E \to E^*$ is denoted by $\sharp\colon E^* \to E$.

As a consequence of Theorem 12.6 we have the following corollary.

Corollary 12.7. *If E is a Euclidean space of finite dimension, every linear form $f \in E^*$ corresponds to a unique $u \in E$ such that*

$$f(v) = u \cdot v, \quad \textit{for every } v \in E.$$

In particular, if f is not the zero form, the kernel of f, which is a hyperplane H, is precisely the set of vectors that are orthogonal to u.

Remarks:

(1) The "musical map" $\flat\colon E \to E^*$ is not surjective when E has infinite dimension. The result can be salvaged by restricting our attention to continuous linear maps, and by assuming that the vector space E is a *Hilbert space* (i.e., E is a complete normed vector space w.r.t. the Euclidean norm). This is the famous "little" Riesz theorem (or Riesz representation theorem).

(2) Theorem 12.6 still holds if the inner product on E is replaced by a nondegenerate symmetric bilinear form φ. We say that a symmetric bilinear form $\varphi\colon E \times E \to \mathbb{R}$ is *nondegenerate* if for every $u \in E$,

$$\text{if} \quad \varphi(u, v) = 0 \quad \text{for all } v \in E, \quad \text{then} \quad u = 0.$$

For example, the symmetric bilinear form on $\mathbb{R}^4$ (the Lorentz form) defined such that

$$\varphi((x_1, x_2, x_3, x_4), (y_1, y_2, y_3, y_4)) = x_1y_1 + x_2y_2 + x_3y_3 - x_4y_4$$

is nondegenerate. However, there are nonnull vectors $u \in \mathbb{R}^4$ such that $\varphi(u, u) = 0$, which is impossible in a Euclidean space. Such vectors are called *isotropic*.

Example 12.6. Consider $\mathbb{R}^n$ with its usual Euclidean inner product. Given any differentiable function $f\colon U \to \mathbb{R}$, where U is some open subset of $\mathbb{R}^n$, by definition, for any $x \in U$, the *total derivative* df_x of f at x is the linear form defined so that for all $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$df_x(u) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)\, u_i.$$

The unique vector $v \in \mathbb{R}^n$ such that

$$v \cdot u = df_x(u) \quad \text{for all } u \in \mathbb{R}^n$$

is the transpose of the *Jacobian matrix* of f at x, the $1 \times n$ matrix

$$\begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

This is the *gradient* $\mathrm{grad}(f)_x$ of f at x, given by

$$\mathrm{grad}(f)_x = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

Example 12.7. Given any two vectors $u, v \in \mathbb{R}^3$, let $c(u, v)$ be the linear form given by

$$c(u, v)(w) = \det(u, v, w) \quad \text{for all } w \in \mathbb{R}^3.$$

Since

$$\begin{aligned} \det(u, v, w) = \begin{vmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{vmatrix} &= w_1 \begin{vmatrix} u_2 & v_2 \\ u_3 & v_3 \end{vmatrix} - w_2 \begin{vmatrix} u_1 & v_1 \\ u_3 & v_3 \end{vmatrix} + w_3 \begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix} \\ &= w_1(u_2 v_3 - u_3 v_2) + w_2(u_3 v_1 - u_1 v_3) + w_3(u_1 v_2 - u_2 v_1), \end{aligned}$$

we see that the unique vector $z \in \mathbb{R}^3$ such that

$$z \cdot w = c(u, v)(w) = \det(u, v, w) \quad \text{for all } w \in \mathbb{R}^3$$

is the vector

$$z = \begin{pmatrix} u_2 v_3 - u_3 v_2 \\ u_3 v_1 - u_1 v_3 \\ u_1 v_2 - u_2 v_1 \end{pmatrix}.$$

This is just the *cross-product* $u \times v$ of u and v. Since $\det(u, v, u) = \det(u, v, v) = 0$, we see that $u \times v$ is orthogonal to both u and v. The above allows us to generalize the cross-product to $\mathbb{R}^n$. Given any $n-1$ vectors $u_1, \ldots, u_{n-1} \in \mathbb{R}^n$, the cross-product $u_1 \times \cdots \times u_{n-1}$ is the unique vector in $\mathbb{R}^n$ such that

$$(u_1 \times \cdots \times u_{n-1}) \cdot w = \det(u_1, \ldots, u_{n-1}, w) \quad \text{for all } w \in \mathbb{R}^n.$$

Example 12.8. Consider the vector space $\mathrm{M}_n(\mathbb{R})$ of real $n \times n$ matrices with the inner product

$$\langle A, B\rangle = \mathrm{tr}(A^\top B).$$

Let $s\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ be the function given by

$$s(A) = \sum_{i,j=1}^{n} a_{ij},$$

where $A = (a_{ij})$. It is immediately verified that s is a linear form. It is easy to check that the unique matrix Z such that

$$\langle Z, A\rangle = s(A) \quad \text{for all } A \in \mathrm{M}_n(\mathbb{R})$$

is the matrix $Z = \mathbf{ones}(n, n)$ whose entries are all equal to 1.

12.3 Adjoint of a Linear Map

The existence of the isomorphism $\flat\colon E \to E^*$ is crucial to the existence of adjoint maps. The importance of adjoint maps stems from the fact that the linear maps arising in physical problems are often self-adjoint, which means that $f = f^*$. Moreover, self-adjoint maps can be diagonalized over orthonormal bases of eigenvectors. This is the key to the solution of many problems in mechanics and engineering in general (see Strang [169]).

Let E be a Euclidean space of finite dimension n, and let $f\colon E \to E$ be a linear map. For every $u \in E$, the map

$$v \mapsto u \cdot f(v)$$

is clearly a linear form in E^*, and by Theorem 12.6, there is a unique vector in E denoted by $f^*(u)$ such that

$$f^*(u) \cdot v = u \cdot f(v),$$

for every $v \in E$. The following simple proposition shows that the map f^* is linear.

Proposition 12.8. *Given a Euclidean space E of finite dimension, for every linear map $f\colon E \to E$, there is a unique linear map $f^*\colon E \to E$ such that*

$$f^*(u) \cdot v = u \cdot f(v), \quad \textit{for all } u, v \in E.$$

Proof. Given $u_1, u_2 \in E$, since the inner product is bilinear, we have

$$(u_1 + u_2) \cdot f(v) = u_1 \cdot f(v) + u_2 \cdot f(v),$$

for all $v \in E$, and

$$(f^*(u_1) + f^*(u_2)) \cdot v = f^*(u_1) \cdot v + f^*(u_2) \cdot v,$$

for all $v \in E$, and since by assumption,

$$f^*(u_1) \cdot v = u_1 \cdot f(v) \quad \text{and} \quad f^*(u_2) \cdot v = u_2 \cdot f(v),$$

for all $v \in E$. Thus we get

$$(f^*(u_1) + f^*(u_2)) \cdot v = (u_1 + u_2) \cdot f(v) = f^*(u_1 + u_2) \cdot v,$$

for all $v \in E$. Since our inner product is positive definite, this implies that

$$f^*(u_1 + u_2) = f^*(u_1) + f^*(u_2).$$

Similarly,

$$(\lambda u) \cdot f(v) = \lambda(u \cdot f(v)),$$

for all $v \in E$, and

$$(\lambda f^*(u)) \cdot v = \lambda(f^*(u) \cdot v),$$

for all $v \in E$, and since by assumption,

$$f^*(u) \cdot v = u \cdot f(v),$$

for all $v \in E$, we get

$$(\lambda f^*(u)) \cdot v = \lambda(u \cdot f(v)) = (\lambda u) \cdot f(v) = f^*(\lambda u) \cdot v$$

for all $v \in E$. Since $\flat$ is bijective, this implies that

$$f^*(\lambda u) = \lambda f^*(u).$$

Thus, f^* is indeed a linear map, and it is unique since $\flat$ is a bijection. $\square$

Definition 12.4. Given a Euclidean space E of finite dimension, for every linear map $f\colon E \to E$, the unique linear map $f^*\colon E \to E$ such that

$$f^*(u) \cdot v = u \cdot f(v), \quad \text{for all } u, v \in E$$

given by Proposition 12.8 is called the *adjoint of f (w.r.t. to the inner product)*. Linear maps $f\colon E \to E$ such that $f = f^*$ are called *self-adjoint* maps.

Self-adjoint linear maps play a very important role because they have real eigenvalues, and because orthonormal bases arise from their eigenvectors. Furthermore, many physical problems lead to self-adjoint linear maps (in the form of symmetric matrices).

Remark: Proposition 12.8 still holds if the inner product on E is replaced by a nondegenerate symmetric bilinear form φ.

Linear maps such that $f^{-1} = f^*$, or equivalently

$$f^* \circ f = f \circ f^* = \mathrm{id},$$

also play an important role. They are *linear isometries*, or *isometries*. Rotations are special kinds of isometries. Another important class of linear maps are the linear maps satisfying the property

$$f^* \circ f = f \circ f^*,$$

called *normal linear maps*. We will see later on that normal maps can always be diagonalized over orthonormal bases of eigenvectors, but this will require using a Hermitian inner product (over $\mathbb{C}$).

Given two Euclidean spaces E and F, where the inner product on E is denoted by $\langle -, - \rangle_1$ and the inner product on F is denoted by $\langle -, - \rangle_2$, given any linear map $f \colon E \to F$, it is immediately verified that the proof of Proposition 12.8 can be adapted to show that there is a unique linear map $f^* \colon F \to E$ such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the *adjoint of* f.

The following properties immediately follow from the definition of the adjoint map:

(1) For any linear map $f \colon E \to F$, we have

$$f^{**} = f.$$

(2) For any two linear maps $f, g \colon E \to F$ and any scalar $\lambda \in \mathbb{R}$:

$$\begin{aligned} (f + g)^* &= f^* + g^* \\ (\lambda f)^* &= \lambda f^*. \end{aligned}$$

(3) If E, F, G are Euclidean spaces with respective inner products $\langle -, - \rangle_1, \langle -, - \rangle_2$, and $\langle -, - \rangle_3$, and if $f \colon E \to F$ and $g \colon F \to G$ are two linear maps, then

$$(g \circ f)^* = f^* \circ g^*.$$

Remark: Given any basis for E and any basis for F, it is possible to characterize the matrix of the adjoint f^* of f in terms of the matrix of f and the Gram matrices defining the inner products; see Problem 12.5. We will do so with respect to orthonormal bases in Proposition 12.14(2). Also, since inner products are symmetric, the adjoint f^* of f is also characterized by

$$f(u) \cdot v = u \cdot f^*(v),$$

for all $u, v \in E$.

12.4 Existence and Construction of Orthonormal Bases

We can also use Theorem 12.6 to show that any Euclidean space of finite dimension has an orthonormal basis.

Proposition 12.9. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, there is an orthonormal basis $(u_1, \ldots, u_n)$ for E.*

Proof. We proceed by induction on n. When $n = 1$, take any nonnull vector $v \in E$, which exists since we assumed E nontrivial, and let

$$u = \frac{v}{\|v\|}.$$

If $n \geq 2$, again take any nonnull vector $v \in E$, and let

$$u_1 = \frac{v}{\|v\|}.$$

Consider the linear form φ_{u_1} associated with u_1. Since $u_1 \neq 0$, by Theorem 12.6, the linear form φ_{u_1} is nonnull, and its kernel is a hyperplane H. Since $\varphi_{u_1}(w) = 0$ iff $u_1 \cdot w = 0$, the hyperplane H is the orthogonal complement of $\{u_1\}$. Furthermore, since $u_1 \neq 0$ and the inner product is positive definite, $u_1 \cdot u_1 \neq 0$, and thus, $u_1 \notin H$, which implies that $E = H \oplus \mathbb{R}u_1$. However, since E is of finite dimension n, the hyperplane H has dimension $n-1$, and by the induction hypothesis, we can find an orthonormal basis $(u_2, \ldots, u_n)$ for H. Now because H and the one dimensional space $\mathbb{R}u_1$ are orthogonal and $E = H \oplus \mathbb{R}u_1$, it is clear that $(u_1, \ldots, u_n)$ is an orthonormal basis for E. $\square$

As a consequence of Proposition 12.9, given any Euclidean space of finite dimension n, if $(e_1, \ldots, e_n)$ is an orthonormal basis for E, then for any two vectors $u = u_1e_1 + \cdots + u_ne_n$ and $v = v_1e_1 + \cdots + v_ne_n$, the inner product $u \cdot v$ is expressed as

$$u \cdot v = (u_1e_1 + \cdots + u_ne_n) \cdot (v_1e_1 + \cdots + v_ne_n) = \sum_{i=1}^{n} u_iv_i,$$

and the norm $\|u\|$ as

$$\|u\| = \|u_1e_1 + \cdots + u_ne_n\| = \left(\sum_{i=1}^{n} u_i^2\right)^{1/2}.$$

The fact that a Euclidean space always has an orthonormal basis implies that any Gram matrix G can be written as

$$G = Q^\top Q,$$

for some invertible matrix Q. Indeed, we know that in a change of basis matrix, a Gram matrix G becomes $G' = P^\top G P$. If the basis corresponding to G' is orthonormal, then $G' = I$, so $G = (P^{-1})^\top P^{-1}$.

There is a more constructive way of proving Proposition 12.9, using a procedure known as the *Gram–Schmidt orthonormalization procedure*. Among other things, the Gram–Schmidt orthonormalization procedure yields the *QR-decomposition for matrices*, an important tool in numerical methods.

Proposition 12.10. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, from any basis $(e_1, \ldots, e_n)$ for E we can construct an orthonormal basis $(u_1, \ldots, u_n)$ for E, with the property that for every k, $1 \leq k \leq n$, the families $(e_1, \ldots, e_k)$ and $(u_1, \ldots, u_k)$ generate the same subspace.*

Proof. We proceed by induction on n. For $n = 1$, let

$$u_1 = \frac{e_1}{\|e_1\|}.$$

For $n \geq 2$, we also let

$$u_1 = \frac{e_1}{\|e_1\|},$$

and assuming that $(u_1, \ldots, u_k)$ is an orthonormal system that generates the same subspace as $(e_1, \ldots, e_k)$, for every k with $1 \leq k < n$, we note that the vector

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} (e_{k+1} \cdot u_i)\, u_i$$

is nonnull, since otherwise, because $(u_1, \ldots, u_k)$ and $(e_1, \ldots, e_k)$ generate the same subspace, $(e_1, \ldots, e_{k+1})$ would be linearly dependent, which is absurd, since $(e_1, \ldots, e_n)$ is a basis. Thus, the norm of the vector u'_{k+1} being nonzero, we use the following construction of the vectors u_k and u'_k:

$$u'_1 = e_1, \qquad u_1 = \frac{u'_1}{\|u'_1\|},$$

and for the inductive step

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} (e_{k+1} \cdot u_i)\, u_i, \qquad u_{k+1} = \frac{u'_{k+1}}{\|u'_{k+1}\|},$$

where $1 \leq k \leq n - 1$. It is clear that $\|u_{k+1}\| = 1$, and since $(u_1, \ldots, u_k)$ is an orthonormal system, we have

$$u'_{k+1} \cdot u_i = e_{k+1} \cdot u_i - (e_{k+1} \cdot u_i) u_i \cdot u_i = e_{k+1} \cdot u_i - e_{k+1} \cdot u_i = 0,$$

for all i with $1 \leq i \leq k$. This shows that the family $(u_1, \ldots, u_{k+1})$ is orthonormal, and since $(u_1, \ldots, u_k)$ and $(e_1, \ldots, e_k)$ generates the same subspace, it is clear from the definition of u_{k+1} that $(u_1, \ldots, u_{k+1})$ and $(e_1, \ldots, e_{k+1})$ generate the same subspace. This completes the induction step and the proof of the proposition. $\square$

Note that u'_{k+1} is obtained by subtracting from e_{k+1} the projection of e_{k+1} itself onto the orthonormal vectors $u_1, \ldots, u_k$ that have already been computed. Then u'_{k+1} is normalized.

Example 12.9. For a specific example of this procedure, let $E = \mathbb{R}^3$ with the standard Euclidean norm. Take the basis

$$e_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad e_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad e_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Then

$$u_1 = 1/\sqrt{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

and

$$u'_2 = e_2 - (e_2 \cdot u_1)u_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2/3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1/3 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

This implies that

$$u_2 = 1/\sqrt{6} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix},$$

and that

$$u'_3 = e_3 - (e_3 \cdot u_1)u_1 - (e_3 \cdot u_2)u_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - 2/3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 1/6 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = 1/2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

To complete the orthonormal basis, normalize u'_3 to obtain

$$u_3 = 1/\sqrt{2} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

An illustration of this example is provided by Figure 12.4.

Remarks:

(1) The QR-decomposition can now be obtained very easily, but we postpone this until Section 12.8.

(2) The proof of Proposition 12.10 also works for a countably infinite basis for E, producing a countably infinite orthonormal basis.

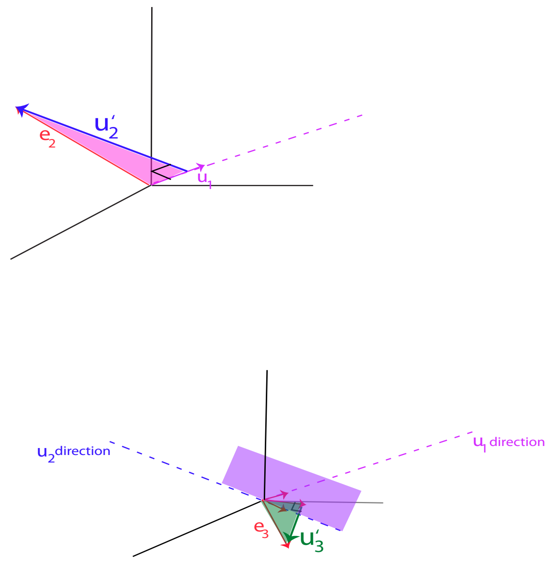


Figure 12.4: The top figure shows the construction of the blue u_2' as perpendicular to the orthogonal projection of e_2 onto u_1, while the bottom figure shows the construction of the green u_3' as normal to the plane determined by u_1 and u_2.

It should also be said that the Gram–Schmidt orthonormalization procedure that we have presented is not very stable numerically, and instead, one should use the *modified Gram–Schmidt method*. To compute u_{k+1}', instead of projecting e_{k+1} onto $u_1, \ldots, u_k$ in a single step, it is better to perform k projections. We compute $u_1^{k+1}, u_2^{k+1}, \ldots, u_k^{k+1}$ as follows:

$$\begin{aligned} u_1^{k+1} &= e_{k+1} - (e_{k+1} \cdot u_1)\, u_1, \\ u_{i+1}^{k+1} &= u_i^{k+1} - (u_i^{k+1} \cdot u_{i+1})\, u_{i+1}, \end{aligned}$$

where $1 \leq i \leq k-1$. It is easily shown that $u_{k+1}' = u_k^{k+1}$.

Example 12.10. Let us apply the modified Gram–Schmidt method to the (e_1, e_2, e_3) basis of Example 12.9. The only change is the computation of u_3'. For the modified Gram–Schmidt procedure, we first calculate

$$u_1^3 = e_3 - (e_3 \cdot u_1)u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - 2/3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1/3 \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

Then

$$u_2^3 = u_1^3 - (u_1^3 \cdot u_2)u_2 = 1/3 \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} + 1/6 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = 1/2 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

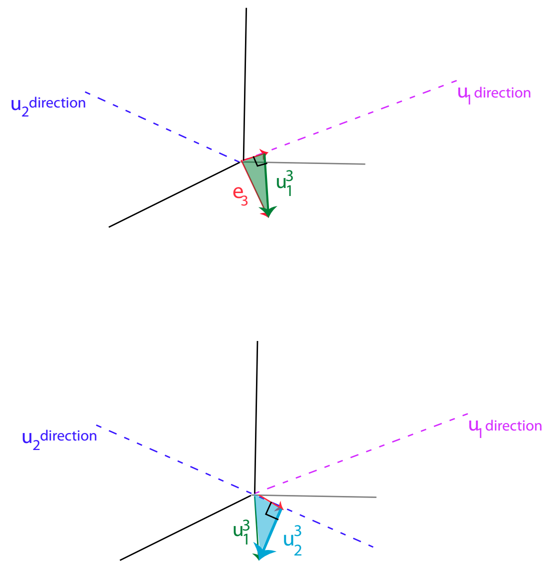


Figure 12.5: The top figure shows the construction of the blue u_1^3 as perpendicular to the orthogonal projection of e_3 onto u_1, while the bottom figure shows the construction of the sky blue u_2^3 as perpendicular to the orthogonal projection of u_1^3 onto u_2.

and observe that $u_2^3 = u_3'$. See Figure 12.5.

The following `Matlab` program implements the modified Gram–Schmidt procedure.

```
function q = gramschmidt4(e)
n = size(e,1);
for i = 1:n
    q(:,i) = e(:,i);
    for j = 1:i-1
        r = q(:,j)'*q(:,i);
        q(:,i) = q(:,i) - r*q(:,j);
    end
    r = sqrt(q(:,i)'*q(:,i));
    q(:,i) = q(:,i)/r;
end
end
```

If we apply the above function to the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

the ouput is the matrix

$$\begin{pmatrix} 0.5774 & 0.4082 & 0.7071 \\ 0.5774 & -0.8165 & -0.0000 \\ 0.5774 & 0.4082 & -0.7071 \end{pmatrix},$$

which matches the result of Example 12.9.

Example 12.11. If we consider polynomials and the inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)dt,$$

applying the Gram–Schmidt orthonormalization procedure to the polynomials

$$1, x, x^2, \ldots, x^n, \ldots,$$

which form a basis of the polynomials in one variable with real coefficients, we get a family of orthonormal polynomials $Q_n(x)$ related to the *Legendre polynomials*.

The Legendre polynomials $P_n(x)$ have many nice properties. They are orthogonal, but their norm is not always 1. The Legendre polynomials $P_n(x)$ can be defined as follows. Letting f_n be the function

$$f_n(x) = (x^2 - 1)^n,$$

we define $P_n(x)$ as follows:

$$P_0(x) = 1, \quad \text{and} \quad P_n(x) = \frac{1}{2^n n!} f_n^{(n)}(x),$$

where $f_n^{(n)}$ is the nth derivative of f_n.

They can also be defined inductively as follows:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_{n+1}(x) &= \frac{2n+1}{n+1} xP_n(x) - \frac{n}{n+1} P_{n-1}(x). \end{aligned}$$

Here is an explicit summation for $P_n(x)$:

$$P_n(x) = \frac{1}{2^n} \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n}{k} \binom{2n-2k}{n} x^{n-2k}.$$

The polynomials Q_n are related to the Legendre polynomials P_n as follows:

$$Q_n(x) = \sqrt{\frac{2n+1}{2}} P_n(x).$$

Example 12.12. Consider polynomials over $[-1, 1]$, with the symmetric bilinear form

$$\langle f, g\rangle = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} f(t)g(t)dt.$$

We leave it as an exercise to prove that the above defines an inner product. It can be shown that the polynomials $T_n(x)$ given by

$$T_n(x) = \cos(n \arccos x), \quad n \geq 0,$$

(equivalently, with $x = \cos\theta$, we have $T_n(\cos\theta) = \cos(n\theta)$) are orthogonal with respect to the above inner product. These polynomials are the *Chebyshev polynomials*. Their norm is not equal to 1. Instead, we have

$$\langle T_n, T_n\rangle = \begin{cases} \frac{\pi}{2} & \text{if } n > 0, \\ \pi & \text{if } n = 0. \end{cases}$$

Using the identity $(\cos\theta + i\sin\theta)^n = \cos n\theta + i \sin n\theta$ and the binomial formula, we obtain the following expression for $T_n(x)$:

$$T_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k} (x^2-1)^k x^{n-2k}.$$

The Chebyshev polynomials are defined inductively as follows:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \quad n \geq 1. \end{aligned}$$

Using these recurrence equations, we can show that

$$T_n(x) = \frac{(x - \sqrt{x^2-1})^n + (x + \sqrt{x^2-1})^n}{2}.$$

The polynomial T_n has n distinct roots in the interval $[-1, 1]$. The Chebyshev polynomials play an important role in approximation theory. They are used as an approximation to a best polynomial approximation of a continuous function under the sup-norm (∞-norm).

The inner products of the last two examples are special cases of an inner product of the form

$$\langle f, g\rangle = \int_{-1}^{1} W(t)f(t)g(t)dt,$$

where $W(t)$ is a *weight function*. If W is a continuous function such that $W(x) > 0$ on $(-1, 1)$, then the above bilinear form is indeed positive definite. Families of orthogonal polynomials used in approximation theory and in physics arise by a suitable choice of the weight function W. Besides the previous two examples, the *Hermite polynomials* correspond to $W(x) = e^{-x^2}$, the *Laguerre polynomials* to $W(x) = e^{-x}$, and the *Jacobi polynomials* to $W(x) = (1-x)^\alpha(1+x)^\beta$, with $\alpha, \beta > -1$. Comprehensive treatments of orthogonal polynomials can be found in Lebedev [114], Sansone [144], and Andrews, Askey and Roy [3].

We can also prove the following proposition regarding orthogonal spaces.

Proposition 12.11. *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, for any subspace F of dimension k, the orthogonal complement $F^\perp$ of F has dimension $n-k$, and $E = F \oplus F^\perp$. Furthermore, we have $F^{\perp\perp} = F$.*

Proof. From Proposition 12.9, the subspace F has some orthonormal basis $(u_1, \ldots, u_k)$. This linearly independent family $(u_1, \ldots, u_k)$ can be extended to a basis $(u_1, \ldots, u_k, v_{k+1}, \ldots, v_n)$, and by Proposition 12.10, it can be converted to an orthonormal basis $(u_1, \ldots, u_n)$, which contains $(u_1, \ldots, u_k)$ as an orthonormal basis of F. Now any vector $w = w_1u_1 + \cdots + w_nu_n \in E$ is orthogonal to F iff $w \cdot u_i = 0$, for every i, where $1 \leq i \leq k$, iff $w_i = 0$ for every i, where $1 \leq i \leq k$. Clearly, this shows that $(u_{k+1}, \ldots, u_n)$ is a basis of $F^\perp$, and thus $E = F \oplus F^\perp$, and $F^\perp$ has dimension $n-k$. Similarly, any vector $w = w_1u_1 + \cdots + w_nu_n \in E$ is orthogonal to $F^\perp$ iff $w \cdot u_i = 0$, for every i, where $k+1 \leq i \leq n$, iff $w_i = 0$ for every i, where $k+1 \leq i \leq n$. Thus, $(u_1, \ldots, u_k)$ is a basis of $F^{\perp\perp}$, and $F^{\perp\perp} = F$. $\square$

12.5 Linear Isometries (Orthogonal Transformations)

In this section we consider linear maps between Euclidean spaces that preserve the Euclidean norm. These transformations, sometimes called *rigid motions*, play an important role in geometry.

Definition 12.5. Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, a function $f\colon E \to F$ is *an orthogonal transformation, or a linear isometry*, if it is linear and

$$\|f(u)\| = \|u\|, \quad \text{for all } u \in E.$$

Remarks:

(1) A linear isometry is often defined as a linear map such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$. Since the map f is linear, the two definitions are equivalent. The second definition just focuses on preserving the distance between vectors.

(2) Sometimes, a linear map satisfying the condition of Definition 12.5 is called a *metric map*, and a linear isometry is defined as a *bijective* metric map.

An isometry (without the word linear) is sometimes defined as a function $f\colon E \to F$ (not necessarily linear) such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$, i.e., as a function that preserves the distance. This requirement turns out to be very strong. Indeed, the next proposition shows that all these definitions are equivalent when E and F are of finite dimension, and for functions such that $f(0) = 0$.

Proposition 12.12. *Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, the following properties are equivalent:*

(1) f is a linear map and $\|f(u)\| = \|u\|$, for all $u \in E$;

(2) $\|f(v) - f(u)\| = \|v - u\|$, for all $u, v \in E$, and $f(0) = 0$;

(3) $f(u) \cdot f(v) = u \cdot v$, for all $u, v \in E$.

Furthermore, such a map is bijective.

Proof. Clearly, (1) implies (2), since in (1) it is assumed that f is linear.

Assume that (2) holds. In fact, we shall prove a slightly stronger result. We prove that if

$$\|f(v) - f(u)\| = \|v - u\|$$

for all $u, v \in E$, then for any vector $\tau \in E$, the function $g\colon E \to F$ defined such that

$$g(u) = f(\tau + u) - f(\tau)$$

for all $u \in E$ is a map satisfying Condition (2), and that (2) implies (3). Clearly, $g(0) = f(\tau) - f(\tau) = 0$.

Note that from the hypothesis

$$\|f(v) - f(u)\| = \|v - u\|$$

for all $u, v \in E$, we conclude that

$$\begin{aligned}
\|g(v) - g(u)\| &= \|f(\tau + v) - f(\tau) - (f(\tau + u) - f(\tau))\|, \\
&= \|f(\tau + v) - f(\tau + u)\|, \\
&= \|\tau + v - (\tau + u)\|, \\
&= \|v - u\|,
\end{aligned}$$

for all $u, v \in E$. Since $g(0) = 0$, by setting $u = 0$ in

$$\|g(v) - g(u)\| = \|v - u\|,$$

we get

$$\|g(v)\| = \|v\|$$

for all $v \in E$. In other words, g preserves both the distance and the norm.

To prove that g preserves the inner product, we use the simple fact that

$$2u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2$$

for all $u, v \in E$. Then since g preserves distance and norm, we have

$$\begin{aligned} 2g(u) \cdot g(v) &= \|g(u)\|^2 + \|g(v)\|^2 - \|g(u) - g(v)\|^2 \\ &= \|u\|^2 + \|v\|^2 - \|u - v\|^2 \\ &= 2u \cdot v, \end{aligned}$$

and thus $g(u) \cdot g(v) = u \cdot v$, for all $u, v \in E$, which is (3). In particular, if $f(0) = 0$, by letting $\tau = 0$, we have $g = f$, and f preserves the scalar product, i.e., (3) holds.

Now assume that (3) holds. Since E is of finite dimension, we can pick an orthonormal basis $(e_1, \ldots, e_n)$ for E. Since f preserves inner products, $(f(e_1), \ldots, f(e_n))$ is also orthonormal, and since F also has dimension n, it is a basis of F. Then note that since $(e_1, \ldots, e_n)$ and $(f(e_1), \ldots, f(e_n))$ are orthonormal bases, for any $u \in E$ we have

$$u = \sum_{i=1}^{n} (u \cdot e_i) e_i = \sum_{i=1}^{n} u_i e_i$$

and

$$f(u) = \sum_{i=1}^{n} (f(u) \cdot f(e_i)) f(e_i),$$

and since f preserves inner products, this shows that

$$f(u) = \sum_{i=1}^{n} (f(u) \cdot f(e_i)) f(e_i) = \sum_{i=1}^{n} (u \cdot e_i) f(e_i) = \sum_{i=1}^{n} u_i f(e_i),$$

which proves that f is linear. Obviously, f preserves the Euclidean norm, and (3) implies (1).

Finally, if $f(u) = f(v)$, then by linearity $f(v - u) = 0$, so that $\|f(v - u)\| = 0$, and since f preserves norms, we must have $\|v - u\| = 0$, and thus $u = v$. Thus, f is injective, and since E and F have the same finite dimension, f is bijective. $\square$

Remarks:

(i) The dimension assumption is needed only to prove that (3) implies (1) when f is not known to be linear, and to prove that f is surjective, but the proof shows that (1) implies that f is injective.

(ii) The implication that (3) implies (1) holds if we also assume that f is surjective, even if E has infinite dimension.

In (2), when f does not satisfy the condition $f(0) = 0$, the proof shows that f is an affine map. Indeed, taking any vector τ as an origin, the map g is linear, and

$$f(\tau + u) = f(\tau) + g(u) \quad \text{for all } u \in E.$$

By Proposition 24.7, this shows that f is affine with associated linear map g.

This fact is worth recording as the following proposition.

Proposition 12.13. *Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, if*

$$\|f(v) - f(u)\| = \|v - u\| \quad \textit{for all } u, v \in E,$$

then f is an affine map, and its associated linear map g is an isometry.

In view of Proposition 12.12, we usually abbreviate "linear isometry" as "isometry," unless we wish to emphasize that we are dealing with a map between vector spaces.

We are now going to take a closer look at the isometries $f\colon E \to E$ of a Euclidean space of finite dimension.

12.6 The Orthogonal Group, Orthogonal Matrices

In this section we explore some of the basic properties of the orthogonal group and of orthogonal matrices.

Proposition 12.14. *Let E be any Euclidean space of finite dimension n, and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) The linear map $f\colon E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, if the matrix of f is A, then the matrix of f^* is the transpose $A^\top$ of A, and f is an isometry iff A satisfies the identities

$$A\,A^\top = A^\top A = I_n,$$

where I_n denotes the identity matrix of order n, iff the columns of A form an orthonormal basis of $\mathbb{R}^n$, iff the rows of A form an orthonormal basis of $\mathbb{R}^n$.

Proof. (1) The linear map $f\colon E \to E$ is an isometry iff

$$f(u) \cdot f(v) = u \cdot v,$$

for all $u, v \in E$, iff

$$f^*(f(u)) \cdot v = f(u) \cdot f(v) = u \cdot v$$

for all $u, v \in E$, which implies

$$(f^*(f(u)) - u) \cdot v = 0$$

for all $u, v \in E$. Since the inner product is positive definite, we must have

$$f^*(f(u)) - u = 0$$

for all $u \in E$, that is,

$$f^* \circ f = \mathrm{id}.$$

But an endomorphism f of a finite-dimensional vector space that has a left inverse is an isomorphism, so $f \circ f^* = \mathrm{id}$. The converse is established by doing the above steps backward.

(2) If $(e_1, \ldots, e_n)$ is an orthonormal basis for E, let $A = (a_{i\,j})$ be the matrix of f, and let $B = (b_{i\,j})$ be the matrix of f^*. Since f^* is characterized by

$$f^*(u) \cdot v = u \cdot f(v)$$

for all $u, v \in E$, using the fact that if $w = w_1 e_1 + \cdots + w_n e_n$ we have $w_k = w \cdot e_k$ for all k, $1 \leq k \leq n$, letting $u = e_i$ and $v = e_j$, we get

$$b_{j\,i} = f^*(e_i) \cdot e_j = e_i \cdot f(e_j) = a_{i\,j},$$

for all i, j, $1 \leq i, j \leq n$. Thus, $B = A^\top$. Now if X and Y are arbitrary matrices over the basis $(e_1, \ldots, e_n)$, denoting as usual the jth column of X by X^j, and similarly for Y, a simple calculation shows that

$$X^\top Y = (X^i \cdot Y^j)_{1 \leq i, j \leq n}.$$

Then it is immediately verified that if $X = Y = A$, then

$$A^\top A = A\,A^\top = I_n$$

iff the column vectors $(A^1, \ldots, A^n)$ form an orthonormal basis. Thus, from (1), we see that (2) is clear (also because the rows of A are the columns of $A^\top$). $\square$

Proposition 12.14 shows that the inverse of an isometry f is its adjoint f^*. Recall that the set of all real $n \times n$ matrices is denoted by $\mathrm{M}_n(\mathbb{R})$. Proposition 12.14 also motivates the following definition.

Definition 12.6. A real $n \times n$ matrix is an *orthogonal matrix* if

$$A\,A^\top = A^\top A = I_n.$$

Remark: It is easy to show that the conditions $A A^\top = I_n$, $A^\top A = I_n$, and $A^{-1} = A^\top$, are equivalent.

Given any two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, if P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, since the columns of P are the coordinates of the vectors v_j with respect to the basis $(u_1, \ldots, u_n)$, if $v_{j_1} = \sum_{i_1=1}^n p_{i_1 j_1} u_{i_1}$ and $v_{j_2} = \sum_{i_2=1}^n p_{i_2 j_2} u_{i_2}$, since $(u_1, \ldots, u_n)$ is orthonormal,

$$v_{j_1} \cdot v_{j_2} = \sum_{i_1=1}^n \sum_{i_2=1}^n p_{i_1 j_1} p_{i_2 j_2} (u_{i_1} \cdot u_{i_2}) = \sum_{i=1}^n p_{i j_1} p_{i j_2},$$

and since $(v_1, \ldots, v_n)$ is orthonormal, $v_{j_1} \cdot v_{j_2} = \delta_{j_1\, j_2}$, so the columns of P are orthonormal, and by Proposition 12.14 (2), the matrix P is orthogonal.

The proof of Proposition 12.12 (3) also shows that if f is an isometry, then the image of an orthonormal basis $(u_1, \ldots, u_n)$ is an orthonormal basis. Students often ask why ortho*gon*al matrices are not called ortho*norm*al matrices, since their columns (and rows) are orthonormal bases! I have no good answer, but isometries do preserve orthogonality, and orthogonal matrices correspond to isometries.

Recall that the determinant $\det(f)$ of a linear map $f \colon E \to E$ is independent of the choice of a basis in E. Also, for every matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have $\det(A) = \det(A^\top)$, and for any two $n \times n$ matrices A and B, we have $\det(AB) = \det(A)\det(B)$. Then if f is an isometry, and A is its matrix with respect to any orthonormal basis, $A A^\top = A^\top A = I_n$ implies that $\det(A)^2 = 1$, that is, either $\det(A) = 1$, or $\det(A) = -1$. It is also clear that the isometries of a Euclidean space of dimension n form a group, and that the isometries of determinant $+1$ form a subgroup. This leads to the following definition.

Definition 12.7. Given a Euclidean space E of dimension n, the set of isometries $f \colon E \to E$ forms a subgroup of $\mathbf{GL}(E)$ denoted by $\mathbf{O}(E)$, or $\mathbf{O}(n)$ when $E = \mathbb{R}^n$, called the *orthogonal group (of E)*. For every isometry f, we have $\det(f) = \pm 1$, where $\det(f)$ denotes the determinant of f. The isometries such that $\det(f) = 1$ are called *rotations, or proper isometries, or proper orthogonal transformations*, and they form a subgroup of the special linear group $\mathbf{SL}(E)$ (and of $\mathbf{O}(E)$), denoted by $\mathbf{SO}(E)$, or $\mathbf{SO}(n)$ when $E = \mathbb{R}^n$, called the *special orthogonal group (of E)*. The isometries such that $\det(f) = -1$ are called *improper isometries, or improper orthogonal transformations, or flip transformations*.

12.7 The Rodrigues Formula

When $n = 3$ and A is a skew symmetric matrix, it is possible to work out an explicit formula for e^A. For any 3×3 real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

if we let $\theta = \sqrt{a^2 + b^2 + c^2}$ and

$$B = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix},$$

then we have the following result known as *Rodrigues' formula* (1840). The (real) vector space of $n \times n$ skew symmetric matrices is denoted by $\mathfrak{so}(n)$.

Proposition 12.15. *The exponential map* $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ *is given by*

$$e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

or, equivalently, by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2$$

if $\theta \neq 0$, *with* $e^{0_3} = I_3$.

Proof sketch. First observe that

$$A^2 = -\theta^2 I_3 + B,$$

since

$$\begin{aligned}
A^2 &= \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} = \begin{pmatrix} -c^2 - b^2 & ba & ca \\ ab & -c^2 - a^2 & cb \\ ac & cb & -b^2 - a^2 \end{pmatrix} \\
&= \begin{pmatrix} -a^2 - b^2 - c^2 & 0 & 0 \\ 0 & -a^2 - b^2 - c^2 & 0 \\ 0 & 0 & -a^2 - b^2 - c^2 \end{pmatrix} + \begin{pmatrix} a^2 & ba & ca \\ ab & b^2 & cb \\ ac & cb & c^2 \end{pmatrix} \\
&= -\theta^2 I_3 + B,
\end{aligned}$$

and that

$$AB = BA = 0.$$

From the above, deduce that

$$A^3 = -\theta^2 A,$$

and for any $k \geq 0$,

$$\begin{aligned}
A^{4k+1} &= \theta^{4k} A, \\
A^{4k+2} &= \theta^{4k} A^2, \\
A^{4k+3} &= -\theta^{4k+2} A, \\
A^{4k+4} &= -\theta^{4k+2} A^2.
\end{aligned}$$

Then prove the desired result by writing the power series for e^A and regrouping terms so that the power series for $\cos\theta$ and $\sin\theta$ show up. In particular

$$\begin{aligned}
e^A = I_3 + \sum_{p\geq 1}\frac{A^p}{p!} &= I_3 + \sum_{p\geq 0}\frac{A^{2p+1}}{(2p+1)!} + \sum_{p\geq 1}\frac{A^{2p}}{(2p)!} \\
&= I_3 + \sum_{p\geq 0}\frac{(-1)^p\theta^{2p}}{(2p+1)!}A + \sum_{p\geq 1}\frac{(-1)^{p-1}\theta^{2(p-1)}}{(2p)!}A^2 \\
&= I_3 + \frac{A}{\theta}\sum_{p\geq 0}\frac{(-1)^p\theta^{2p+1}}{(2p+1)!} - \frac{A^2}{\theta^2}\sum_{p\geq 1}\frac{(-1)^p\theta^{2p}}{(2p)!} \\
&= I_3 + \frac{\sin\theta}{\theta}A - \frac{A^2}{\theta^2}\sum_{p\geq 0}\frac{(-1)^p\theta^{2p}}{(2p)!} + \frac{A^2}{\theta^2} \\
&= I_3 + \frac{\sin\theta}{\theta}A + \frac{(1-\cos\theta)}{\theta^2}A^2,
\end{aligned}$$

as claimed. ☐

The above formulae are the well-known formulae expressing a rotation of axis specified by the vector (a, b, c) and angle θ.

The Rodrigues formula can used to show that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective.

Given any rotation matrix $R \in \mathbf{SO}(3)$, we have the following cases:

(1) The case $R = I$ is trivial.

(2) If $R \neq I$ and $\mathrm{tr}(R) \neq -1$, then

$$\exp^{-1}(R) = \left\{ \frac{\theta}{2\sin\theta}(R - R^T) \;\middle|\; 1 + 2\cos\theta = \mathrm{tr}(R) \right\}.$$

(Recall that $\mathrm{tr}(R) = r_{1\,1} + r_{2\,2} + r_{3\,3}$, the *trace* of the matrix R).

Then there is a unique skew-symmetric B with corresponding θ satisfying $0 < \theta < \pi$ such that $e^B = R$.

(3) If $R \neq I$ and $\mathrm{tr}(R) = -1$, then R is a rotation by the angle π and things are more complicated, but a matrix B can be found. We leave this part as a good exercise: see Problem 17.8.

The computation of a logarithm of a rotation in $\mathbf{SO}(3)$ as sketched above has applications in kinematics, robotics, and motion interpolation.

As an immediate corollary of the Gram–Schmidt orthonormalization procedure, we obtain the QR-decomposition for invertible matrices.

12.8 QR-Decomposition for Invertible Matrices

Now that we have the definition of an orthogonal matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decomposition for matrices.

Definition 12.8. Given any real $n \times n$ matrix A, a *QR-decomposition* of A is any pair of $n \times n$ matrices (Q, R), where Q is an orthogonal matrix and R is an upper triangular matrix such that $A = QR$.

Note that if A is not invertible, then some diagonal entry in R must be zero.

Proposition 12.16. *Given any real $n \times n$ matrix A, if A is invertible, then there is an orthogonal matrix Q and an upper triangular matrix R with positive diagonal entries such that $A = QR$.*

Proof. We can view the columns of A as vectors $A^1, \ldots, A^n$ in $\mathbb{E}^n$. If A is invertible, then they are linearly independent, and we can apply Proposition 12.10 to produce an orthonormal basis using the Gram–Schmidt orthonormalization procedure. Recall that we construct vectors Q^k and Q'^k as follows:

$$Q'^1 = A^1, \qquad Q^1 = \frac{Q'^1}{\|Q'^1\|},$$

and for the inductive step

$$Q'^{k+1} = A^{k+1} - \sum_{i=1}^{k} (A^{k+1} \cdot Q^i)\, Q^i, \qquad Q^{k+1} = \frac{Q'^{k+1}}{\|Q'^{k+1}\|},$$

where $1 \leq k \leq n-1$. If we express the vectors A^k in terms of the Q^i and Q'^i, we get the triangular system

$$\begin{aligned}
A^1 &= \|Q'^1\| Q^1, \\
&\vdots \\
A^j &= (A^j \cdot Q^1)\, Q^1 + \cdots + (A^j \cdot Q^i)\, Q^i + \cdots + (A^j \cdot Q^{j-1})\, Q^{j-1} + \|Q'^j\| Q^j, \\
&\vdots \\
A^n &= (A^n \cdot Q^1)\, Q^1 + \cdots + (A^n \cdot Q^{n-1})\, Q^{n-1} + \|Q'^n\| Q^n.
\end{aligned}$$

Letting $r_{k\,k} = \|Q'^k\|$, and $r_{i\,j} = A^j \cdot Q^i$ (the reversal of i and j on the right-hand side *is* intentional!), where $1 \leq k \leq n$, $2 \leq j \leq n$, and $1 \leq i \leq j-1$, and letting $q_{i\,j}$ be the ith component of Q^j, we note that $a_{i\,j}$, the ith component of A^j, is given by

$$a_{i\,j} = r_{1\,j} q_{i\,1} + \cdots + r_{i\,j} q_{i\,i} + \cdots + r_{j\,j} q_{i\,j} = q_{i\,1} r_{1\,j} + \cdots + q_{i\,i} r_{i\,j} + \cdots + q_{i\,j} r_{j\,j}.$$

If we let $Q = (q_{i\,j})$, the matrix whose columns are the components of the Q^j, and $R = (r_{i\,j})$, the above equations show that $A = QR$, where R is upper triangular. The diagonal entries $r_{k\,k} = \|Q'^k\| = A^k \cdot Q^k$ are indeed positive. $\square$

The reader should try the above procedure on some concrete examples for 2×2 and 3×3 matrices.

Remarks:

(1) Because the diagonal entries of R are positive, it can be shown that Q and R are unique. More generally, if A is invertible and if $A = Q_1 R_1 = Q_2 R_2$ are two QR-decompositions for A, then
$$R_1 R_2^{-1} = Q_1^\top Q_2.$$
The matrix $Q_1^\top Q_2$ is orthogonal and it is easy to see that $R_1 R_2^{-1}$ is upper triangular. But an upper triangular matrix which is orthogonal must be a diagonal matrix D with diagonal entries ± 1, so $Q_2 = Q_1 D$ and $R_1 = D R_2$.

(2) The QR-decomposition holds even when A is not invertible. In this case, R has some zero on the diagonal. However, a different proof is needed. We will give a nice proof using Householder matrices (see Proposition 13.4, and also Strang [169, 170], Golub and Van Loan [80], Trefethen and Bau [176], Demmel [48], Kincaid and Cheney [102], or Ciarlet [41]).

For better numerical stability, it is preferable to use the modified Gram–Schmidt method to implement the QR-factorization method. Here is a `Matlab` program implementing QR-factorization using modified Gram–Schmidt.

```
function [Q,R] = qrv4(A)
n = size(A,1);
for i = 1:n
    Q(:,i) = A(:,i);
    for j = 1:i-1
        R(j,i) = Q(:,j)'*Q(:,i);
        Q(:,i) = Q(:,i) - R(j,i)*Q(:,j);
    end
    R(i,i) = sqrt(Q(:,i)'*Q(:,i));
    Q(:,i) = Q(:,i)/R(i,i);
end
end
```

Example 12.13. Consider the matrix
$$A = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

To determine the QR-decomposition of A, we first use the Gram-Schmidt orthonormalization procedure to calculate $Q = (Q^1 Q^2 Q^3)$. By definition

$$A^1 = Q'^1 = Q^1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and since $A^2 = \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}$, we discover that

$$Q'^2 = A^2 - (A^2 \cdot Q^1)Q^1 = \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix}.$$

Hence, $Q^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. Finally,

$$Q'^3 = A_3 - (A^3 \cdot Q^1)Q^1 - (A^3 \cdot Q^2)Q^2 = \begin{pmatrix} 5 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix},$$

which implies that $Q^3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. According to Proposition 12.16, in order to determine R we need to calculate

$$\begin{aligned} r_{11} &= \left\|Q'^1\right\| = 1 & r_{12} &= A^2 \cdot Q^1 = 1 & r_{13} &= A^3 \cdot Q^1 = 1 \\ & & r_{22} &= \left\|Q'^2\right\| = 4 & r_{23} &= A_3 \cdot Q^2 = 1 \\ & & & & r_{33} &= \left\|Q'^3\right\| = 5. \end{aligned}$$

In summary, we have found that the QR-decomposition of $A = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ is

$$Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad R = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 1 \\ 0 & 0 & 5 \end{pmatrix}.$$

Example 12.14. Another example of QR-decomposition is

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 1/\sqrt{2} & \sqrt{2} \\ 0 & 1/\sqrt{2} & \sqrt{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Example 12.15. If we apply the above `Matlab` function to the matrix

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix},$$

we obtain

$$Q = \begin{pmatrix} 0.9701 & -0.2339 & 0.0619 & -0.0166 & 0.0046 \\ 0.2425 & 0.9354 & -0.2477 & 0.0663 & -0.0184 \\ 0 & 0.2650 & 0.9291 & -0.2486 & 0.0691 \\ 0 & 0 & 0.2677 & 0.9283 & -0.2581 \\ 0 & 0 & 0 & 0.2679 & 0.9634 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 4.1231 & 1.9403 & 0.2425 & 0 & 0 \\ 0 & 3.7730 & 1.9956 & 0.2650 & 0 \\ 0 & 0 & 3.7361 & 1.9997 & 0.2677 \\ 0 & 0 & & 073.7324 & 2.0000 \\ 0 & 0 & 0 & 0 & 3.5956 \end{pmatrix}.$$

Remark: The `Matlab` function `qr`, called by `[Q, R] = qr(A)`, does not necessarily return an upper-triangular matrix whose diagonal entries are positive.

The QR-decomposition yields a rather efficient and numerically stable method for solving systems of linear equations. Indeed, given a system $Ax = b$, where A is an $n \times n$ invertible matrix, writing $A = QR$, since Q is orthogonal, we get

$$Rx = Q^{\top} b,$$

and since R is upper triangular, we can solve it by Gaussian elimination, by solving for the last variable x_n first, substituting its value into the system, then solving for x_{n-1}, etc. The QR-decomposition is also very useful in solving least squares problems (we will come back to this in Chapter 23), and for finding eigenvalues; see Chapter 18. It can be easily adapted to the case where A is a rectangular $m \times n$ matrix with independent columns (thus, $n \leq m$). In this case, Q is not quite orthogonal. It is an $m \times n$ matrix whose columns are orthogonal, and R is an invertible $n \times n$ upper triangular matrix with positive diagonal entries. For more on QR, see Strang [169, 170], Golub and Van Loan [80], Demmel [48], Trefethen and Bau [176], or Serre [156].

A somewhat surprising consequence of the QR-decomposition is a famous determinantal inequality due to Hadamard.

Proposition 12.17. *(Hadamard) For any real $n \times n$ matrix $A = (a_{ij})$, we have*

$$|\det(A)| \leq \prod_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} \quad \textit{and} \quad |\det(A)| \leq \prod_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}^2 \right)^{1/2}.$$

Moreover, equality holds iff either A has orthogonal rows in the left inequality or orthogonal columns in the right inequality.

Proof. If $\det(A) = 0$, then the inequality is trivial. In addition, if the righthand side is also 0, then either some column or some row is zero. If $\det(A) \neq 0$, then we can factor A as $A = QR$, with Q is orthogonal and $R = (r_{ij})$ upper triangular with positive diagonal entries. Then since Q is orthogonal $\det(Q) = \pm 1$, so

$$|\det(A)| = |\det(Q)|\, |\det(R)| = \prod_{j=1} r_{jj}.$$

Now as Q is orthogonal, it preserves the Euclidean norm, so

$$\sum_{i=1}^{n} a_{ij}^2 = \left\| A^j \right\|_2^2 = \left\| QR^j \right\|_2^2 = \left\| R^j \right\|_2^2 = \sum_{i=1}^{n} r_{ij}^2 \geq r_{jj}^2,$$

which implies that

$$|\det(A)| = \prod_{j=1}^{n} r_{jj} \leq \prod_{j=1}^{n} \left\| R^j \right\|_2 = \prod_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}^2 \right)^{1/2}.$$

The other inequality is obtained by replacing A by $A^\top$. Finally, if $\det(A) \neq 0$ and equality holds, then we must have

$$r_{jj} = \left\| A^j \right\|_2, \quad 1 \leq j \leq n,$$

which can only occur if A has orthogonal columns. $\square$

Another version of Hadamard's inequality applies to symmetric positive semidefinite matrices.

Proposition 12.18. *(Hadamard) For any real $n \times n$ matrix $A = (a_{ij})$, if A is symmetric positive semidefinite, then we have*

$$\det(A) \leq \prod_{i=1}^{n} a_{ii}.$$

Moreover, if A is positive definite, then equality holds iff A is a diagonal matrix.

Proof. If $\det(A) = 0$, the inequality is trivial. Otherwise, A is positive definite, and by Theorem 8.10 (the Cholesky Factorization), there is a unique upper triangular matrix B with positive diagonal entries such that

$$A = B^\top B.$$

Thus, $\det(A) = \det(B^\top B) = \det(B^\top)\det(B) = \det(B)^2$. If we apply the Hadamard inequality (Proposition 12.17) to B, we obtain

$$\det(B) \leq \prod_{j=1}^{n} \left(\sum_{i=1}^{n} b_{ij}^2 \right)^{1/2}. \tag{$*$}$$

However, the diagonal entries a_{jj} of $A = B^\top B$ are precisely the square norms $\|B^j\|_2^2 = \sum_{i=1}^n b_{ij}^2$, so by squaring $(*)$, we obtain

$$\det(A) = \det(B)^2 \leq \prod_{j=1}^{n} \left(\sum_{i=1}^{n} b_{ij}^2 \right) = \prod_{j=1}^{n} a_{jj}.$$

If $\det(A) \neq 0$ and equality holds, then B must have orthogonal columns, which implies that B is a diagonal matrix, and so is A. $\square$

We derived the second Hadamard inequality (Proposition 12.18) from the first (Proposition 12.17). We leave it as an exercise to prove that the first Hadamard inequality can be deduced from the second Hadamard inequality.

12.9 Some Applications of Euclidean Geometry

Euclidean geometry has applications in computational geometry, in particular Voronoi diagrams and Delaunay triangulations. In turn, Voronoi diagrams have applications in motion planning (see O'Rourke [133]).

Euclidean geometry also has applications to matrix analysis. Recall that a real $n \times n$ matrix A is *symmetric* if it is equal to its transpose $A^\top$. One of the most important properties of symmetric matrices is that they have real eigenvalues and that they can be diagonalized by an orthogonal matrix (see Chapter 17). This means that for every symmetric matrix A, there is a diagonal matrix D and an orthogonal matrix P such that

$$A = PDP^\top.$$

Even though it is not always possible to diagonalize an arbitrary matrix, there are various decompositions involving orthogonal matrices that are of great practical interest. For example, for every real matrix A, there is the *QR-decomposition*, which says that a real matrix A can be expressed as

$$A = QR,$$

where Q is orthogonal and R is an upper triangular matrix. This can be obtained from the Gram–Schmidt orthonormalization procedure, as we saw in Section 12.8, or better, using Householder matrices, as shown in Section 13.2. There is also the *polar decomposition*, which says that a real matrix A can be expressed as

$$A = QS,$$

where Q is orthogonal and S is symmetric positive semidefinite (which means that the eigenvalues of S are nonnegative). Such a decomposition is important in continuum mechanics and in robotics, since it separates stretching from rotation. Finally, there is the wonderful *singular value decomposition*, abbreviated as SVD, which says that a real matrix A can be expressed as

$$A = VDU^{\top},$$

where U and V are orthogonal and D is a diagonal matrix with nonnegative entries (see Chapter 22). This decomposition leads to the notion of *pseudo-inverse*, which has many applications in engineering (least squares solutions, etc). For an excellent presentation of all these notions, we highly recommend Strang [170, 169], Golub and Van Loan [80], Demmel [48], Serre [156], and Trefethen and Bau [176].

The method of least squares, invented by Gauss and Legendre around 1800, is another great application of Euclidean geometry. Roughly speaking, the method is used to solve inconsistent linear systems $Ax = b$, where the number of equations is greater than the number of variables. Since this is generally impossible, the method of least squares consists in finding a solution x minimizing the Euclidean norm $\|Ax - b\|^2$, that is, the sum of the squares of the "errors." It turns out that there is always a unique solution x^+ of smallest norm minimizing $\|Ax - b\|^2$, and that it is a solution of the square system

$$A^{\top}Ax = A^{\top}b,$$

called the system of *normal equations*. The solution x^+ can be found either by using the QR-decomposition in terms of Householder transformations, or by using the notion of pseudo-inverse of a matrix. The pseudo-inverse can be computed using the SVD decomposition. Least squares methods are used extensively in computer vision. More details on the method of least squares and pseudo-inverses can be found in Chapter 23.

12.10 Summary

The main concepts and results of this chapter are listed below:

- Bilinear forms; *positive definite* bilinear forms.
- *Inner products*, *scalar products*, *Euclidean spaces*.
- *Quadratic form* associated with a bilinear form.

- The Euclidean space $\mathbb{E}^n$.
- The *polar form* of a quadratic form.
- *Gram matrix* associated with an inner product.
- The *Cauchy–Schwarz inequality*; the *Minkowski inequality*.
- The *parallelogram law*.
- *Orthogonality*, *orthogonal complement* $F^{\perp}$; *orthonormal family*.
- The *musical isomorphisms* $\flat \colon E \to E^*$ and $\sharp \colon E^* \to E$ (when E is finite-dimensional); Theorem 12.6.
- The *adjoint* of a linear map (with respect to an inner product).
- Existence of an orthonormal basis in a finite-dimensional Euclidean space (Proposition 12.9).
- The *Gram–Schmidt orthonormalization procedure* (Proposition 12.10).
- The *Legendre* and the *Chebyshev* polynomials.
- *Linear isometries* (*orthogonal transformations*, *rigid motions*).
- The *orthogonal group*, *orthogonal matrices*.
- The matrix representing the adjoint f^* of a linear map f is the transpose of the matrix representing f.
- The *orthogonal group* $\mathbf{O}(n)$ and the *special orthogonal group* $\mathbf{SO}(n)$.
- *QR-decomposition* for invertible matrices.
- The *Hadamard inequality* for arbitrary real matrices.
- The *Hadamard inequality* for symmetric positive semidefinite matrices.
- The *Rodrigues formula* for rotations in $\mathbf{SO}(3)$.

12.11 Problems

Problem 12.1. E be a vector space of dimension 2, and let (e_1, e_2) be a basis of E. Prove that if $a > 0$ and $b^2 - ac < 0$, then the bilinear form defined such that

$$\varphi(x_1e_1 + y_1e_2,\, x_2e_1 + y_2e_2) = ax_1x_2 + b(x_1y_2 + x_2y_1) + cy_1y_2$$

is a Euclidean inner product.

Problem 12.2. Let $\mathcal{C}[a,b]$ denote the set of continuous functions $f\colon [a,b] \to \mathbb{R}$. Given any two functions $f, g \in \mathcal{C}[a,b]$, let

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt.$$

Prove that the above bilinear form is indeed a Euclidean inner product.

Problem 12.3. Consider the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$$

of Problem 12.2 on the vector space $\mathcal{C}[-\pi, \pi]$. Prove that

$$\langle \sin px, \sin qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 1, \end{cases}$$

$$\langle \cos px, \cos qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 0, \end{cases}$$

$$\langle \sin px, \cos qx\rangle = 0,$$

for all $p \geq 1$ and $q \geq 0$, and $\langle 1, 1\rangle = \int_{-\pi}^{\pi} dx = 2\pi$.

Problem 12.4. Prove that the following matrix is orthogonal and skew-symmetric:

$$M = \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \\ -1 & -1 & 1 & 0 \end{pmatrix}.$$

Problem 12.5. Let E and F be two finite Euclidean spaces, let $(u_1, \ldots, u_n)$ be a basis of E, and let $(v_1, \ldots, v_m)$ be a basis of F. For any linear map $f\colon E \to F$, if A is the matrix of f w.r.t. the basis $(u_1, \ldots, u_n)$ and B is the matrix of f^* w.r.t. the basis $(v_1, \ldots, v_m)$, if G_1 is the Gram matrix of the inner product on E (w.r.t. $(u_1, \ldots, u_n)$) and if G_2 is the Gram matrix of the inner product on F (w.r.t. $(v_1, \ldots, v_m)$), then

$$B = G_1^{-1} A^\top G_2.$$

Problem 12.6. Let A be an invertible matrix. Prove that if $A = Q_1R_1 = Q_2R_2$ are two QR-decompositions of A and if the diagonal entries of R_1 and R_2 are positive, then $Q_1 = Q_2$ and $R_1 = R_2$.

Problem 12.7. Prove that the first Hadamard inequality can be deduced from the second Hadamard inequality.

Problem 12.8. Let E be a real vector space of finite dimension, $n \geq 1$. Say that two bases, $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, of E have the *same orientation* iff $\det(P) > 0$, where P the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, namely, the matrix whose jth columns consist of the coordinates of v_j over the basis $(u_1, \ldots, u_n)$.

(1) Prove that having the same orientation is an equivalence relation with two equivalence classes.

An *orientation* of a vector space, E, is the choice of any fixed basis, say $(e_1, \ldots, e_n)$, of E. Any other basis, $(v_1, \ldots, v_n)$, has the *same orientation* as $(e_1, \ldots, e_n)$ (and is said to be *positive* or *direct*) iff $\det(P) > 0$, else it is said to have the *opposite orientation* of $(e_1, \ldots, e_n)$ (or to be *negative* or *indirect*), where P is the change of basis matrix from $(e_1, \ldots, e_n)$ to $(v_1, \ldots, v_n)$. An *oriented* vector space is a vector space with some chosen orientation (a positive basis).

(2) Let $B_1 = (u_1, \ldots, u_n)$ and $B_2 = (v_1, \ldots, v_n)$ be two orthonormal bases. For any sequence of vectors, $(w_1, \ldots, w_n)$, in E, let $\det_{B_1}(w_1, \ldots, w_n)$ be the determinant of the matrix whose columns are the coordinates of the w_j's over the basis B_1 and similarly for $\det_{B_2}(w_1, \ldots, w_n)$.

Prove that if B_1 and B_2 have the same orientation, then

$$\det_{B_1}(w_1, \ldots, w_n) = \det_{B_2}(w_1, \ldots, w_n).$$

Given any oriented vector space, E, for any sequence of vectors, $(w_1, \ldots, w_n)$, in E, the common value, $\det_B(w_1, \ldots, w_n)$, for all positive orthonormal bases, B, of E is denoted

$$\lambda_E(w_1, \ldots, w_n)$$

and called a *volume form* of $(w_1, \ldots, w_n)$.

(3) Given any Euclidean oriented vector space, E, of dimension n for any $n-1$ vectors, $w_1, \ldots, w_{n-1}$, in E, check that the map

$$x \mapsto \lambda_E(w_1, \ldots, w_{n-1}, x)$$

is a linear form. Then prove that there is a unique vector, denoted $w_1 \times \cdots \times w_{n-1}$, such that

$$\lambda_E(w_1, \ldots, w_{n-1}, x) = (w_1 \times \cdots \times w_{n-1}) \cdot x,$$

for all $x \in E$. The vector $w_1 \times \cdots \times w_{n-1}$ is called the *cross-product* of $(w_1, \ldots, w_{n-1})$. It is a generalization of the cross-product in $\mathbb{R}^3$ (when $n = 3$).

Problem 12.9. Given p vectors $(u_1, \ldots, u_p)$ in a Euclidean space E of dimension $n \geq p$, the *Gram determinant (or Gramian)* of the vectors $(u_1, \ldots, u_p)$ is the determinant

$$\mathrm{Gram}(u_1, \ldots, u_p) = \begin{vmatrix} \|u_1\|^2 & \langle u_1, u_2 \rangle & \ldots & \langle u_1, u_p \rangle \\ \langle u_2, u_1 \rangle & \|u_2\|^2 & \ldots & \langle u_2, u_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_p, u_1 \rangle & \langle u_p, u_2 \rangle & \ldots & \|u_p\|^2 \end{vmatrix}.$$

(1) Prove that

$$\mathrm{Gram}(u_1, \ldots, u_n) = \lambda_E(u_1, \ldots, u_n)^2.$$

Hint. If $(e_1, \ldots, e_n)$ is an orthonormal basis and A is the matrix of the vectors $(u_1, \ldots, u_n)$ over this basis,

$$\det(A)^2 = \det(A^\top A) = \det(A^i \cdot A^j),$$

where A^i denotes the ith column of the matrix A, and $(A^i \cdot A^j)$ denotes the $n \times n$ matrix with entries $A^i \cdot A^j$.

(2) Prove that

$$\|u_1 \times \cdots \times u_{n-1}\|^2 = \mathrm{Gram}(u_1, \ldots, u_{n-1}).$$

Hint. Letting $w = u_1 \times \cdots \times u_{n-1}$, observe that

$$\lambda_E(u_1, \ldots, u_{n-1}, w) = \langle w, w \rangle = \|w\|^2,$$

and show that

$$\begin{aligned} \|w\|^4 &= \lambda_E(u_1, \ldots, u_{n-1}, w)^2 = \mathrm{Gram}(u_1, \ldots, u_{n-1}, w) \\ &= \mathrm{Gram}(u_1, \ldots, u_{n-1})\|w\|^2. \end{aligned}$$

Problem 12.10. Let $\varphi \colon E \times E \to \mathbb{R}$ be a bilinear form on a real vector space E of finite dimension n. Given any basis $(e_1, \ldots, e_n)$ of E, let $A = (a_{i\,j})$ be the matrix defined such that

$$a_{i\,j} = \varphi(e_i, e_j),$$

$1 \leq i, j \leq n$. We call A *the matrix of* φ *w.r.t. the basis* $(e_1, \ldots, e_n)$.

(1) For any two vectors x and y, if X and Y denote the column vectors of coordinates of x and y w.r.t. the basis $(e_1, \ldots, e_n)$, prove that

$$\varphi(x, y) = X^\top A Y.$$

(2) Recall that A is a *symmetric* matrix if $A = A^\top$. Prove that φ is symmetric if A is a symmetric matrix.

(3) If $(f_1, \ldots, f_n)$ is another basis of E and P is the change of basis matrix from $(e_1, \ldots, e_n)$ to $(f_1, \ldots, f_n)$, prove that the matrix of φ w.r.t. the basis $(f_1, \ldots, f_n)$ is

$$P^\top A P.$$

The common rank of all matrices representing φ is called the *rank* of φ.

Problem 12.11. Let $\varphi\colon E \times E \to \mathbb{R}$ be a symmetric bilinear form on a real vector space E of finite dimension n. Two vectors x and y are said to be *conjugate or orthogonal w.r.t.* φ if $\varphi(x, y) = 0$. The main purpose of this problem is to prove that there is a basis of vectors that are pairwise conjugate w.r.t. φ.

(1) Prove that if $\varphi(x, x) = 0$ for all $x \in E$, then φ is identically null on E.

Otherwise, we can assume that there is some vector $x \in E$ such that $\varphi(x, x) \neq 0$.

Use induction to prove that there is a basis of vectors $(u_1, \ldots, u_n)$ that are pairwise conjugate w.r.t. φ.

Hint. For the induction step, proceed as follows. Let $(u_1, e_2, \ldots, e_n)$ be a basis of E, with $\varphi(u_1, u_1) \neq 0$. Prove that there are scalars $\lambda_2, \ldots, \lambda_n$ such that each of the vectors

$$v_i = e_i + \lambda_i u_1$$

is conjugate to u_1 w.r.t. φ, where $2 \leq i \leq n$, and that $(u_1, v_2, \ldots, v_n)$ is a basis.

(2) Let $(e_1, \ldots, e_n)$ be a basis of vectors that are pairwise conjugate w.r.t. φ and assume that they are ordered such that

$$\varphi(e_i, e_i) = \begin{cases} \theta_i \neq 0 & \text{if } 1 \leq i \leq r, \\ 0 & \text{if } r + 1 \leq i \leq n, \end{cases}$$

where r is the rank of φ. Show that the matrix of φ w.r.t. $(e_1, \ldots, e_n)$ is a diagonal matrix, and that

$$\varphi(x, y) = \sum_{i=1}^{r} \theta_i x_i y_i,$$

where $x = \sum_{i=1}^n x_i e_i$ and $y = \sum_{i=1}^n y_i e_i$.

Prove that for every symmetric matrix A, there is an invertible matrix P such that

$$P^\top A P = D,$$

where D is a diagonal matrix.

(3) Prove that there is an integer p, $0 \leq p \leq r$ (where r is the rank of φ), such that $\varphi(u_i, u_i) > 0$ for exactly p vectors of every basis $(u_1, \ldots, u_n)$ of vectors that are pairwise conjugate w.r.t. φ (*Sylvester's inertia theorem*).

Proceed as follows. Assume that in the basis $(u_1, \ldots, u_n)$, for any $x \in E$, we have

$$\varphi(x, x) = \alpha_1 x_1^2 + \cdots + \alpha_p x_p^2 - \alpha_{p+1} x_{p+1}^2 - \cdots - \alpha_r x_r^2,$$

where $x = \sum_{i=1}^n x_i u_i$, and that in the basis $(v_1, \ldots, v_n)$, for any $x \in E$, we have

$$\varphi(x, x) = \beta_1 y_1^2 + \cdots + \beta_q y_q^2 - \beta_{q+1} y_{q+1}^2 - \cdots - \beta_r y_r^2,$$

where $x = \sum_{i=1}^n y_i v_i$, with $\alpha_i > 0$, $\beta_i > 0$, $1 \leq i \leq r$.

Assume that $p > q$ and derive a contradiction. First consider x in the subspace F spanned by

$$(u_1, \ldots, u_p, u_{r+1}, \ldots, u_n),$$

and observe that $\varphi(x, x) \geq 0$ if $x \neq 0$. Next consider x in the subspace G spanned by

$$(v_{q+1}, \ldots, v_r),$$

and observe that $\varphi(x, x) < 0$ if $x \neq 0$. Prove that $F \cap G$ is nontrivial (i.e., contains some nonnull vector), and derive a contradiction. This implies that $p \leq q$. Finish the proof.

The pair $(p, r - p)$ is called the *signature* of φ.

(4) A symmetric bilinear form φ is *definite* if for every $x \in E$, if $\varphi(x, x) = 0$, then $x = 0$.

Prove that a symmetric bilinear form is definite iff its signature is either $(n, 0)$ or $(0, n)$. In other words, a symmetric definite bilinear form has rank n and is either positive or negative.

Problem 12.12. Consider the $n \times n$ matrices $R^{i,j}$ defined for all i, j with $1 \leq i < j \leq n$ and $n \geq 3$, such that the only nonzero entries are

$$\begin{aligned}
R^{i,j}(i, j) &= -1 \\
R^{i,j}(i, i) &= 0 \\
R^{i,j}(j, i) &= 1 \\
R^{i,j}(j, j) &= 0 \\
R^{i,j}(k, k) &= 1, \quad 1 \leq k \leq n, k \neq i, j.
\end{aligned}$$

For example,

$$R^{i,j} = \begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & 0 & 0 & \cdots & 0 & -1 & & & \\
& & & 0 & 1 & \cdots & 0 & 0 & & & \\
& & & \vdots & \vdots & \ddots & \vdots & \vdots & & & \\
& & & 0 & 0 & \cdots & 1 & 0 & & & \\
& & & 1 & 0 & \cdots & 0 & 0 & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}.$$

(1) Prove that the $R^{i,j}$ are rotation matrices. Use the matrices R^{ij} to form a basis of the $n \times n$ skew-symmetric matrices.

(2) Consider the $n \times n$ symmetric matrices $S^{i,j}$ defined for all i, j with $1 \leq i < j \leq n$ and $n \geq 3$, such that the only nonzero entries are

$$\begin{aligned}
S^{i,j}(i,j) &= 1 \\
S^{i,j}(i,i) &= 0 \\
S^{i,j}(j,i) &= 1 \\
S^{i,j}(j,j) &= 0 \\
S^{i,j}(k,k) &= 1, \quad 1 \leq k \leq n, k \neq i, j,
\end{aligned}$$

and if $i + 2 \leq j$ then $S^{i,j}(i+1, i+1) = -1$, else if $i > 1$ and $j = i+1$ then $S^{i,j}(1,1) = -1$, and if $i = 1$ and $j = 2$, then $S^{i,j}(3,3) = -1$.

For example,

$$S^{i,j} = \begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & 0 & 0 & \cdots & 0 & 1 & & & \\
& & & 0 & -1 & \cdots & 0 & 0 & & & \\
& & & \vdots & \vdots & \ddots & \vdots & \vdots & & & \\
& & & 0 & 0 & \cdots & 1 & 0 & & & \\
& & & 1 & 0 & \cdots & 0 & 0 & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}.$$

Note that $S^{i,j}$ has a single diagonal entry equal to -1. Prove that the $S^{i,j}$ are rotations matrices.

Use Problem 3.15 together with the $S^{i,j}$ to form a basis of the $n \times n$ symmetric matrices.

(3) Prove that if $n \geq 3$, the set of all linear combinations of matrices in $\mathbf{SO}(n)$ is the space $\mathrm{M}_n(\mathbb{R})$ of all $n \times n$ matrices.

Prove that if $n \geq 3$ and if a matrix $A \in \mathrm{M}_n(\mathbb{R})$ commutes with all rotations matrices, then A commutes with all matrices in $\mathrm{M}_n(\mathbb{R})$.

What happens for $n = 2$?

Problem 12.13. (1) Let H be the affine hyperplane in $\mathbb{R}^n$ given by the equation

$$a_1 x_1 + \cdots + a_n x_n = c,$$

with $a_i \neq 0$ for some $i, 1 \leq i \leq n$. The linear hyperplane H_0 parallel to H is given by the equation

$$a_1 x_1 + \cdots + a_n x_n = 0,$$

and we say that a vector $y \in \mathbb{R}^n$ is *orthogonal* (or *perpendicular*) to H iff y is orthogonal to H_0. Let h be the intersection of H with the line through the origin and perpendicular to H. Prove that the coordinates of h are given by

$$\frac{c}{a_1^2 + \cdots + a_n^2}(a_1, \ldots, a_n).$$

(2) For any point $p \in H$, prove that $\|h\| \leq \|p\|$. Thus, it is natural to define the *distance* $d(O, H)$ from the origin O to the hyperplane H as $d(O, H) = \|h\|$. Prove that

$$d(O, H) = \frac{|c|}{(a_1^2 + \cdots + a_n^2)^{\frac{1}{2}}}.$$

(3) Let S be a finite set of $n \geq 3$ points in the plane ($\mathbb{R}^2$). Prove that if for every pair of distinct points $p_i, p_j \in S$, there is a third point $p_k \in S$ (distinct from p_i and p_j) such that p_i, p_j, p_k belong to the same (affine) line, then all points in S belong to a common (affine) line.

Hint. Proceed by contradiction and use a minimality argument. This is either ∞-hard or relatively easy, depending how you proceed!

Problem 12.14. (The space of closed polygons in $\mathbb{R}^2$, after Hausmann and Knutson)

An *open polygon* P in the plane is a sequence $P = (v_1, \ldots, v_{n+1})$ of points $v_i \in \mathbb{R}^2$ called *vertices* (with $n \geq 1$). A *closed polygon*, for short a *polygon*, is an open polygon $P = (v_1, \ldots, v_{n+1})$ such that $v_{n+1} = v_1$. The sequence of *edge vectors* $(e_1, \ldots, e_n)$ associated with the open (or closed) polygon $P = (v_1, \ldots, v_{n+1})$ is defined by

$$e_i = v_{i+1} - v_i, \quad i = 1, \ldots, n.$$

Thus, a closed or open polygon is also defined by a pair $(v_1, (e_1, \ldots, e_n))$, with the vertices given by

$$v_{i+1} = v_i + e_i, \quad i = 1, \ldots, n.$$

Observe that a polygon $(v_1, (e_1, \ldots, e_n))$ is closed iff

$$e_1 + \cdots + e_n = 0.$$

Since every polygon $(v_1, (e_1, \ldots, e_n))$ can be translated by $-v_1$, so that $v_1 = (0, 0)$, we may assume that our polygons are specified by a sequence of edge vectors.

Recall that the plane $\mathbb{R}^2$ is isomorphic to $\mathbb{C}$, via the isomorphism

$$(x, y) \mapsto x + iy.$$

We will represent each edge vector e_k by the square of a complex number $w_k = a_k + ib_k$. Thus, every sequence of complex numbers $(w_1, \ldots, w_n)$ defines a polygon (namely, $(w_1^2, \ldots, w_n^2)$).

This representation is many-to-one: the sequences $(\pm w_1, \ldots, \pm w_n)$ describe the same polygon. To every sequence of complex numbers $(w_1, \ldots, w_n)$, we associate the pair of vectors (a, b), with $a, b \in \mathbb{R}^n$, such that if $w_k = a_k + ib_k$, then

$$a = (a_1, \ldots, a_n), \quad b = (b_1, \ldots, b_n).$$

The mapping

$$(w_1, \ldots, w_n) \mapsto (a, b)$$

is clearly a bijection, so we can also represent polygons by pairs of vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$.

(1) Prove that a polygon P represented by a pair of vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$ is closed iff $a \cdot b = 0$ and $\|a\|_2 = \|b\|_2$.

(2) Given a polygon P represented by a pair of vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, the length $l(P)$ of the polygon P is defined by $l(P) = |w_1|^2 + \cdots + |w_n|^2$, with $w_k = a_k + ib_k$. Prove that

$$l(P) = \|a\|_2^2 + \|b\|_2^2 .$$

Deduce from (a) and (b) that every closed polygon of length 2 with n edges is represented by a $n \times 2$ matrix A such that $A^\top A = I$.

Remark: The space of all a $n \times 2$ real matrices A such that $A^\top A = I$ is a space known as the *Stiefel manifold* $S(2, n)$.

(3) Recall that in $\mathbb{R}^2$, the rotation of angle θ specified by the matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is expressed in terms of complex numbers by the map

$$z \mapsto ze^{i\theta}.$$

Let P be a polygon represented by a pair of vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$. Prove that the polygon $R_\theta(P)$ obtained by applying the rotation R_θ to every vertex $w_k^2 = (a_k + ib_k)^2$ of P is specified by the pair of vectors

$$(\cos(\theta/2)a - \sin(\theta/2)b,\ \sin(\theta/2)a + \cos(\theta/2)b) = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{pmatrix} \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}.$$

(4) The reflection ρ_x about the x-axis corresponds to the map

$$z \mapsto \overline{z},$$

whose matrix is,

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Prove that the polygon $\rho_x(P)$ obtained by applying the reflection ρ_x to every vertex $w_k^2 = (a_k + ib_k)^2$ of P is specified by the pair of vectors

$$(a, -b) = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

(5) Let $Q \in \mathbf{O}(2)$ be any isometry such that $\det(Q) = -1$ (a reflection). Prove that there is a rotation $R_{-\theta} \in \mathbf{SO}(2)$ such that

$$Q = \rho_x \circ R_{-\theta}.$$

Prove that the isometry Q, which is given by the matrix

$$Q = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix},$$

is the reflection about the line corresponding to the angle $\theta/2$ (the line of equation $y = \tan(\theta/2)x$).

Prove that the polygon $Q(P)$ obtained by applying the reflection $Q = \rho_x \circ R_{-\theta}$ to every vertex $w_k^2 = (a_k + ib_k)^2$ of P, is specified by the pair of vectors

$$(\cos(\theta/2)a + \sin(\theta/2)b,\ \sin(\theta/2)a - \cos(\theta/2)b) = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{pmatrix} \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ \sin(\theta/2) & -\cos(\theta/2) \end{pmatrix}.$$

(6) Define an equivalence relation $\sim$ on $S(2, n)$ such that if $A_1, A_2 \in S(2, n)$ are any $n \times 2$ matrices such that $A_1^\top A_1 = A_2^\top A_2 = I$, then

$$A_1 \sim A_2 \quad \text{iff} \quad A_2 = A_1 Q \quad \text{for some } Q \in \mathbf{O}(2).$$

Prove that the quotient $G(2, n) = S(2, n)/\sim$ is in bijection with the set of all 2-dimensional subspaces (the planes) of $\mathbb{R}^n$. The space $G(2, n)$ is called a *Grassmannian manifold*.

Prove that up to translations and isometries in $\mathbf{O}(2)$ (rotations and reflections), the n-sided closed polygons of length 2 are represented by planes in $G(2, n)$.

Problem 12.15. (1) Find two symmetric matrices, A and B, such that AB is not symmetric.

(2) Find two matrices A and B such that

$$e^A e^B \neq e^{A+B}.$$

Hint. Try

$$A = \pi \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \pi \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

and use the Rodrigues formula.

(3) Find some square matrices A, B such that $AB \neq BA$, yet

$$e^A e^B = e^{A+B}.$$

Hint. Look for 2×2 matrices with zero trace and use Problem 9.15.

Problem 12.16. Given a field K and any nonempty set I, let $K^{(I)}$ be the subset of the cartesian product K^I consisting of all functions $\lambda\colon I \to K$ with *finite support*, which means that $\lambda(i) = 0$ for all but finitely many $i \in I$. We usually denote the function defined by λ as $(\lambda_i)_{i\in I}$, and call is a *family indexed by* I. We define addition and multiplication by a scalar as follows:

$$(\lambda_i)_{i\in I} + (\mu_i)_{i\in I} = (\lambda_i + \mu_i)_{i\in I},$$

and

$$\alpha \cdot (\mu_i)_{i\in I} = (\alpha\mu_i)_{i\in I}.$$

(1) Check that $K^{(I)}$ is a vector space.

(2) If I is any nonempty subset, for any $i \in I$, we denote by e_i the family $(e_j)_{j\in I}$ defined so that

$$e_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i. \end{cases}$$

Prove that the family $(e_i)_{i\in I}$ is linearly independent and spans $K^{(I)}$, so that it is a basis of $K^{(I)}$ called the *canonical basis* of $K^{(I)}$. When I is finite, say of cardinality n, then prove that $K^{(I)}$ is isomorphic to K^n.

(3) The function $\iota\colon I \to K^{(I)}$, such that $\iota(i) = e_i$ for every $i \in I$, is clearly an injection.

For any other vector space F, for any function $f\colon I \to F$, prove that there is a *unique linear map* $\overline{f}\colon K^{(I)} \to F$, such that

$$f = \overline{f} \circ \iota,$$

as in the following commutative diagram:

$$\begin{tikzcd} I \arrow[r, "\iota"] \arrow[dr, "f"'] & K^{(I)} \arrow[d, "\overline{f}"] \\ & F \end{tikzcd}.$$

We call the vector space $K^{(I)}$ the vector space *freely generated* by the set I.

Problem 12.17. (Some pitfalls of infinite dimension) Let E be the vector space freely generated by the set of natural numbers, $\mathbb{N} = \{0, 1, 2, \ldots\}$, and let $(e_0, e_1, e_2, \ldots, e_n, \ldots)$ be its canonical basis. We define the function φ such that

$$\varphi(e_i, e_j) = \begin{cases} \delta_{ij} & \text{if } i, j \geq 1, \\ 1 & \text{if } i = j = 0, \\ 1/2^j & \text{if } i = 0, j \geq 1, \\ 1/2^i & \text{if } i \geq 1, j = 0, \end{cases}$$

and we extend φ by bilinearity to a function $\varphi\colon E \times E \to K$. This means that if $u = \sum_{i\in\mathbb{N}} \lambda_i e_i$ and $v = \sum_{j\in\mathbb{N}} \mu_j e_j$, then

$$\varphi\left(\sum_{i\in\mathbb{N}} \lambda_i e_i, \sum_{j\in\mathbb{N}} \mu_j e_j\right) = \sum_{i,j\in\mathbb{N}} \lambda_i \mu_j \varphi(e_i, e_j),$$

but remember that $\lambda_i \neq 0$ and $\mu_j \neq 0$ *only for finitely many indices* i, j.

(1) Prove that φ is positive definite, so that it is an inner product on E.

What would happen if we changed $1/2^j$ to 1 (or any constant)?

(2) Let H be the subspace of E spanned by the family $(e_i)_{i\geq 1}$, a hyperplane in E. Find $H^\perp$ and $H^{\perp\perp}$, and prove that

$$H \neq H^{\perp\perp}.$$

(3) Let U be the subspace of E spanned by the family $(e_{2i})_{i\geq 1}$, and let V be the subspace of E spanned by the family $(e_{2i-1})_{i\geq 1}$. Prove that

$$\begin{aligned} U^\perp &= V \\ V^\perp &= U \\ U^{\perp\perp} &= U \\ V^{\perp\perp} &= V, \end{aligned}$$

yet

$$(U \cap V)^\perp \neq U^\perp + V^\perp$$

and

$$(U + V)^{\perp\perp} \neq U + V.$$

If W is the subspace spanned by e_0 and e_1, prove that

$$(W \cap H)^\perp \neq W^\perp + H^\perp.$$

(4) Consider the dual space E^* of E, and let $(e_i^*)_{i\in\mathbb{N}}$ be the family of dual forms of the basis $(e_i)_{i\in N}$. Check that the family $(e_i^*)_{i\in\mathbb{N}}$ is linearly independent.

(5) Let $f \in E^*$ be the linear form defined by

$$f(e_i) = 1 \quad \text{for all } i \in \mathbb{N}.$$

Prove that f is not in the subspace spanned by the e_i^*. If F is the subspace of E^* spanned by the e_i^* and f, find F^0 and F^{00}, and prove that

$$F \neq F^{00}.$$

Chapter 13

QR-Decomposition for Arbitrary Matrices

13.1 Orthogonal Reflections

Hyperplane reflections are represented by matrices called Householder matrices. These matrices play an important role in numerical methods, for instance for solving systems of linear equations, solving least squares problems, for computing eigenvalues, and for transforming a symmetric matrix into a tridiagonal matrix. We prove a simple geometric lemma that immediately yields the QR-decomposition of arbitrary matrices in terms of Householder matrices.

Orthogonal symmetries are a very important example of isometries. First let us review the definition of projections, introduced in Section 6.1, just after Proposition 6.7. Given a vector space E, let F and G be subspaces of E that form a direct sum $E = F \oplus G$. Since every $u \in E$ can be written uniquely as $u = v + w$, where $v \in F$ and $w \in G$, we can define the two *projections* $p_F \colon E \to F$ and $p_G \colon E \to G$ such that $p_F(u) = v$ and $p_G(u) = w$. In Section 6.1 we used the notation π_1 and π_2, but in this section it is more convenient to use p_F and p_G.

It is immediately verified that p_G and p_F are linear maps, and that

$$p_F^2 = p_F,\ p_G^2 = p_G,\ p_F \circ p_G = p_G \circ p_F = 0, \quad \text{and} \quad p_F + p_G = \mathrm{id}.$$

.

Definition 13.1. Given a vector space E, for any two subspaces F and G that form a direct sum $E = F \oplus G$, the *symmetry (or reflection) with respect to F and parallel to G* is the linear map $s \colon E \to E$ defined such that

$$s(u) = 2p_F(u) - u,$$

for every $u \in E$.

Because $p_F + p_G = \mathrm{id}$, note that we also have

$$s(u) = p_F(u) - p_G(u)$$

and

$$s(u) = u - 2p_G(u),$$

$s^2 = \mathrm{id}$, s is the identity on F, and $s = -\mathrm{id}$ on G.

We now assume that E is a Euclidean space of *finite* dimension.

Definition 13.2. Let E be a Euclidean space of finite dimension n. For any two subspaces F and G, if F and G form a direct sum $E = F \oplus G$ and F and G are orthogonal, i.e., $F = G^{\perp}$, the *orthogonal symmetry (or reflection) with respect to F and parallel to G* is the linear map $s\colon E \to E$ defined such that

$$s(u) = 2p_F(u) - u = p_F(u) - p_G(u),$$

for every $u \in E$. When F is a hyperplane, we call s a *hyperplane symmetry with respect to F (or reflection about F)*, and when G is a plane (and thus $\dim(F) = n - 2$), we call s a *flip about F*.

A reflection about a hyperplane F is shown in Figure 13.1.

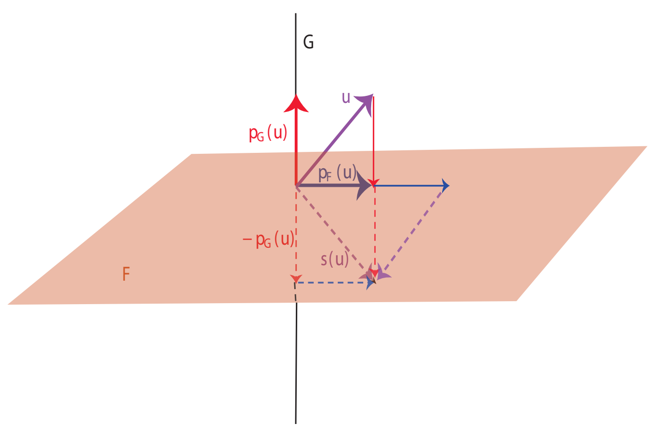


Figure 13.1: A reflection about the peach hyperplane F. Note that u is purple, $p_F(u)$ is blue and $p_G(u)$ is red.

For any two vectors $u, v \in E$, it is easily verified using the bilinearity of the inner product that

$$\|u + v\|^2 - \|u - v\|^2 = 4(u \cdot v). \qquad (*)$$

In particular, if $u \cdot v = 0$, then $\|u + v\| = \|u - v\|$. Then since

$$u = p_F(u) + p_G(u)$$

and

$$s(u) = p_F(u) - p_G(u),$$

and since F and G are orthogonal, it follows that

$$p_F(u) \cdot p_G(v) = 0,$$

and thus by $(*)$

$$\|s(u)\| = \|p_F(u) - p_G(u)\| = \|p_F(u) + p_G(u)\| = \|u\|,$$

so that s *is an isometry.*

Using Proposition 12.10, it is possible to find an orthonormal basis $(e_1, \ldots, e_n)$ of E consisting of an orthonormal basis of F and an orthonormal basis of G. Assume that F has dimension p, so that G has dimension $n - p$. With respect to the orthonormal basis $(e_1, \ldots, e_n)$, the symmetry s has a matrix of the form

$$\begin{pmatrix} I_p & 0 \\ 0 & -I_{n-p} \end{pmatrix}.$$

Thus, $\det(s) = (-1)^{n-p}$, and s is a rotation iff $n - p$ is even. In particular, when F is a hyperplane H, we have $p = n - 1$ and $n - p = 1$, so that s is an improper orthogonal transformation. When $F = \{0\}$, we have $s = -\mathrm{id}$, which is called the *symmetry with respect to the origin*. The symmetry with respect to the origin is a rotation iff n is even, and an improper orthogonal transformation iff n is odd. When n is odd, since $s \circ s = \mathrm{id}$ and $\det(s) = (-1)^n = -1$, we observe that every improper orthogonal transformation f is the composition $f = (f \circ s) \circ s$ of the rotation $f \circ s$ with s, the symmetry with respect to the origin. When G is a plane, $p = n - 2$, and $\det(s) = (-1)^2 = 1$, so that a flip about F is a rotation. In particular, when $n = 3$, F is a line, and a flip about the line F is indeed a rotation of measure π as illustrated by Figure 13.2.

Remark: Given any two orthogonal subspaces F, G forming a direct sum $E = F \oplus G$, let f be the symmetry with respect to F and parallel to G, and let g be the symmetry with respect to G and parallel to F. We leave as an exercise to show that

$$f \circ g = g \circ f = -\mathrm{id}.$$

When $F = H$ is a hyperplane, we can give an explicit formula for $s(u)$ in terms of any nonnull vector w orthogonal to H. Indeed, from

$$u = p_H(u) + p_G(u),$$

since $p_G(u) \in G$ and G is spanned by w, which is orthogonal to H, we have

$$p_G(u) = \lambda w$$

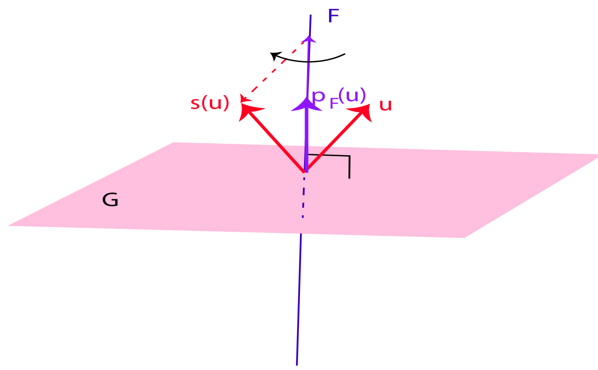


Figure 13.2: A flip in $\mathbb{R}^3$ is a rotation of π about the F axis.

for some $\lambda \in \mathbb{R}$, and we get

$$u \cdot w = \lambda \|w\|^2,$$

and thus

$$p_G(u) = \frac{(u \cdot w)}{\|w\|^2}\, w.$$

Since

$$s(u) = u - 2p_G(u),$$

we get

$$s(u) = u - 2\,\frac{(u \cdot w)}{\|w\|^2}\, w.$$

Since the above formula is important, we record it in the following proposition.

Proposition 13.1. *Let E be a finite-dimensional Euclidean space and let H be a hyperplane in E. For any nonzero vector w orthogonal to H, the hyperplane reflection s about H is given by*

$$s(u) = u - 2\,\frac{(u \cdot w)}{\|w\|^2}\, w, \quad u \in E.$$

Such reflections are represented by matrices called *Householder matrices*, which play an important role in numerical matrix analysis (see Kincaid and Cheney [102] or Ciarlet [41]).

Definition 13.3. A *Householder matrix* is a matrix of the form

$$H = I_n - 2\,\frac{WW^\top}{\|W\|^2} = I_n - 2\,\frac{WW^\top}{W^\top W},$$

where $W \in \mathbb{R}^n$ is a nonzero vector.

Householder matrices are symmetric and orthogonal. It is easily checked that over an orthonormal basis $(e_1, \ldots, e_n)$, a hyperplane reflection about a hyperplane H orthogonal to a nonzero vector w is represented by the matrix

$$H = I_n - 2\,\frac{WW^\top}{\|W\|^2},$$

where W is the column vector of the coordinates of w over the basis $(e_1, \ldots, e_n)$. Since

$$p_G(u) = \frac{(u \cdot w)}{\|w\|^2}\, w,$$

the matrix representing p_G is

$$\frac{WW^\top}{W^\top W},$$

and since $p_H + p_G = \mathrm{id}$, the matrix representing p_H is

$$I_n - \frac{WW^\top}{W^\top W}.$$

These formulae can be used to derive a formula for a rotation of $\mathbb{R}^3$, given the direction w of its axis of rotation and given the angle θ of rotation.

The following fact is the key to the proof that every isometry can be decomposed as a product of reflections.

Proposition 13.2. *Let E be any nontrivial Euclidean space. For any two vectors $u, v \in E$, if $\|u\| = \|v\|$, then there is a hyperplane H such that the reflection s about H maps u to v, and if $u \neq v$, then this reflection is unique. See Figure 13.3.*

Proof. If $u = v$, then any hyperplane containing u does the job. Otherwise, we must have $H = \{v - u\}^\perp$, and by the above formula,

$$s(u) = u - 2\,\frac{(u \cdot (v-u))}{\|(v-u)\|^2}\,(v-u) = u + \frac{2\|u\|^2 - 2u \cdot v}{\|(v-u)\|^2}\,(v-u),$$

and since

$$\|(v-u)\|^2 = \|u\|^2 + \|v\|^2 - 2u \cdot v$$

and $\|u\| = \|v\|$, we have

$$\|(v-u)\|^2 = 2\|u\|^2 - 2u \cdot v,$$

and thus, $s(u) = v$. $\square$

If E is a complex vector space and the inner product is Hermitian, Proposition 13.2 is false. The problem is that the vector $v - u$ does not work unless the inner product $u \cdot v$ is real! The proposition can be salvaged enough to yield the QR-decomposition in terms of Householder transformations; see Section 14.5.

We now show that hyperplane reflections can be used to obtain another proof of the QR-decomposition.

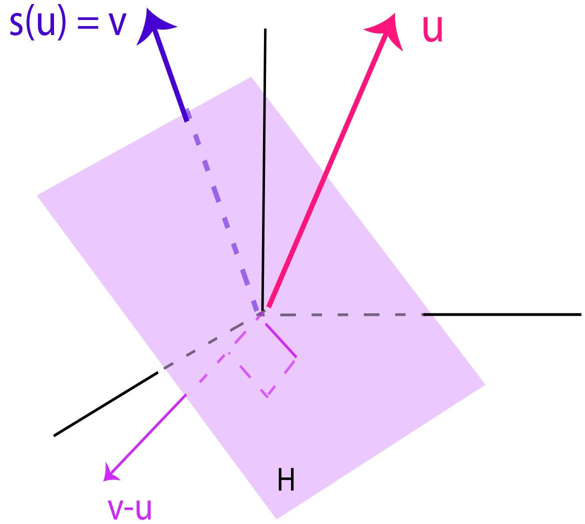


Figure 13.3: In $\mathbb{R}^3$, the (hyper)plane perpendicular to $v - u$ reflects u onto v.

13.2 QR-Decomposition Using Householder Matrices

First we state the result geometrically. When translated in terms of Householder matrices, we obtain the fact advertised earlier that every matrix (not necessarily invertible) has a QR-decomposition.

Proposition 13.3. *Let E be a nontrivial Euclidean space of dimension n. For any orthonormal basis $(e_1, \ldots, e_n)$ and for any n-tuple of vectors $(v_1, \ldots, v_n)$, there is a sequence of n isometries $h_1, \ldots, h_n$ such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by*

$$r_j = h_n \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq n$. Equivalently, the matrix R whose columns are the components of the r_j over the basis $(e_1, \ldots, e_n)$ is an upper triangular matrix. Furthermore, the h_i can be chosen so that the diagonal entries of R are nonnegative.

Proof. We proceed by induction on n. For $n = 1$, we have $v_1 = \lambda e_1$ for some $\lambda \in \mathbb{R}$. If $\lambda \geq 0$, we let $h_1 = \mathrm{id}$, else if $\lambda < 0$, we let $h_1 = -\mathrm{id}$, the reflection about the origin.

For $n \geq 2$, we first have to find h_1. Let

$$r_{1,1} = \|v_1\|.$$

If $v_1 = r_{1,1}e_1$, we let $h_1 = \mathrm{id}$. Otherwise, there is a unique hyperplane reflection h_1 such that

$$h_1(v_1) = r_{1,1}\, e_1,$$

defined such that

$$h_1(u) = u - 2\,\frac{(u \cdot w_1)}{\|w_1\|^2}\, w_1$$

for all $u \in E$, where

$$w_1 = r_{1,1}\, e_1 - v_1.$$

The map h_1 is the reflection about the hyperplane H_1 orthogonal to the vector $w_1 = r_{1,1}\, e_1 - v_1$. See Figure 13.4. Letting

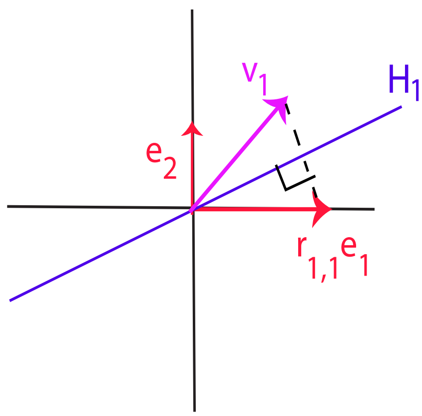


Figure 13.4: The construction of h_1 in Proposition 13.3.

$$r_1 = h_1(v_1) = r_{1,1}\, e_1,$$

it is obvious that r_1 belongs to the subspace spanned by e_1, and $r_{1,1} = \|v_1\|$ is nonnegative.

Next assume that we have found k linear maps $h_1, \ldots, h_k$, hyperplane reflections or the identity, where $1 \leq k \leq n-1$, such that if $(r_1, \ldots, r_k)$ are the vectors given by

$$r_j = h_k \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq k$. See Figure 13.5. The vectors $(e_1, \ldots, e_k)$ form a basis for the subspace denoted by U'_k, the vectors $(e_{k+1}, \ldots, e_n)$ form a basis for the subspace denoted by U''_k, the subspaces U'_k and U''_k are orthogonal, and $E = U'_k \oplus U''_k$. Let

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}).$$

We can write

$$u_{k+1} = u'_{k+1} + u''_{k+1},$$

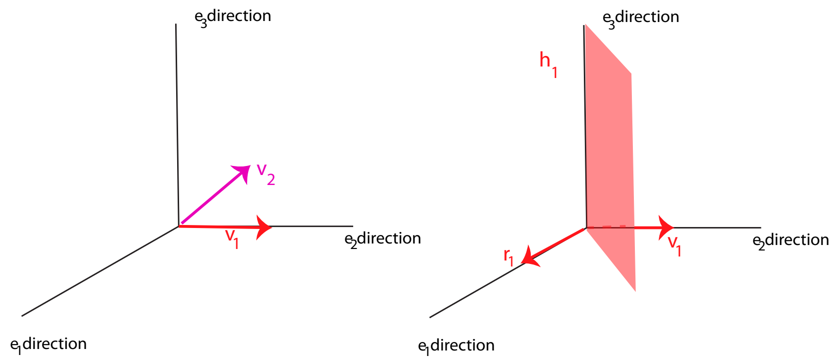


Figure 13.5: The construction of $r_1 = h_1(v_1)$ in Proposition 13.3.

where $u'_{k+1} \in U'_k$ and $u''_{k+1} \in U''_k$. See Figure 13.6. Let

$$r_{k+1,k+1} = \|u''_{k+1}\|.$$

If $u''_{k+1} = r_{k+1,k+1}\, e_{k+1}$, we let $h_{k+1} = \mathrm{id}$. Otherwise, there is a unique hyperplane reflection h_{k+1} such that

$$h_{k+1}(u''_{k+1}) = r_{k+1,k+1}\, e_{k+1},$$

defined such that

$$h_{k+1}(u) = u - 2\,\frac{(u \cdot w_{k+1})}{\|w_{k+1}\|^2}\, w_{k+1}$$

for all $u \in E$, where

$$w_{k+1} = r_{k+1,k+1}\, e_{k+1} - u''_{k+1}.$$

The map h_{k+1} is the reflection about the hyperplane H_{k+1} orthogonal to the vector $w_{k+1} = r_{k+1,k+1}\, e_{k+1} - u''_{k+1}$. However, since $u''_{k+1}, e_{k+1} \in U''_k$ and U'_k is orthogonal to U''_k, the subspace U'_k is contained in H_{k+1}, and thus, the vectors $(r_1, \ldots, r_k)$ and u'_{k+1}, which belong to U'_k, are invariant under h_{k+1}. This proves that

$$h_{k+1}(u_{k+1}) = h_{k+1}(u'_{k+1}) + h_{k+1}(u''_{k+1}) = u'_{k+1} + r_{k+1,k+1}\, e_{k+1}$$

is a linear combination of $(e_1, \ldots, e_{k+1})$. Letting

$$r_{k+1} = h_{k+1}(u_{k+1}) = u'_{k+1} + r_{k+1,k+1}\, e_{k+1},$$

since $u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1})$, the vector

$$r_{k+1} = h_{k+1} \circ \cdots \circ h_2 \circ h_1(v_{k+1})$$

is a linear combination of $(e_1, \ldots, e_{k+1})$. See Figure 13.7. The coefficient of r_{k+1} over e_{k+1} is $r_{k+1,k+1} = \|u''_{k+1}\|$, which is nonnegative. This concludes the induction step, and thus the proof. $\square$

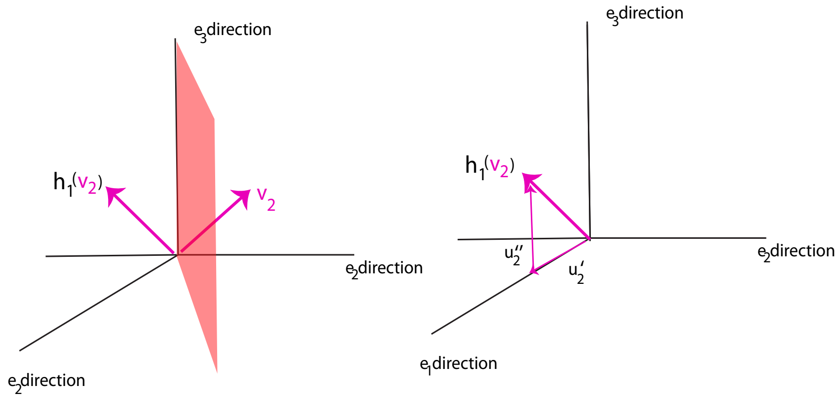


Figure 13.6: The construction of $u_2 = h_1(v_2)$ and its decomposition as $u_2 = u_2' + u_2''$.

Remarks:

(1) Since every h_i is a hyperplane reflection or the identity,

$$\rho = h_n \circ \cdots \circ h_2 \circ h_1$$

is an isometry.

(2) If we allow negative diagonal entries in R, the last isometry h_n may be omitted.

(3) Instead of picking $r_{k,k} = \|u_k''\|$, which means that

$$w_k = r_{k,k}\, e_k - u_k'',$$

where $1 \leq k \leq n$, it might be preferable to pick $r_{k,k} = -\|u_k''\|$ if this makes $\|w_k\|^2$ larger, in which case

$$w_k = r_{k,k}\, e_k + u_k''.$$

Indeed, since the definition of h_k involves division by $\|w_k\|^2$, it is desirable to avoid division by very small numbers.

(4) The method also applies to any m-tuple of vectors $(v_1, \ldots, v_m)$, with $m \leq n$. Then R is an upper triangular $m \times m$ matrix and Q is an $n \times m$ matrix with orthogonal columns ($Q^\top Q = I_m$). We leave the minor adjustments to the method as an exercise to the reader

Proposition 13.3 directly yields the QR-decomposition in terms of Householder transformations (see Strang [169, 170], Golub and Van Loan [80], Trefethen and Bau [176], Kincaid and Cheney [102], or Ciarlet [41]).

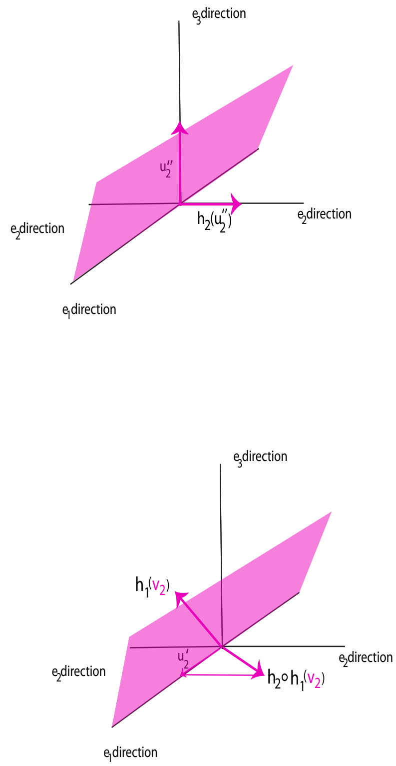


Figure 13.7: The construction of h_2 and $r_2 = h_2 \circ h_1(v_2)$ in Proposition 13.3.

Theorem 13.4. *For every real $n \times n$ matrix A, there is a sequence $H_1, \ldots, H_n$ of matrices, where each H_i is either a Householder matrix or the identity, and an upper triangular matrix R such that*

$$R = H_n \cdots H_2 H_1 A.$$

As a corollary, there is a pair of matrices Q, R, where Q is orthogonal and R is upper triangular, such that $A = QR$ (a QR-decomposition of A). Furthermore, R can be chosen so that its diagonal entries are nonnegative.

Proof. The jth column of A can be viewed as a vector v_j over the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{E}^n$ (where $(e_j)_i = 1$ if $i = j$, and 0 otherwise, $1 \leq i, j \leq n$). Applying Proposition 13.3 to $(v_1, \ldots, v_n)$, there is a sequence of n isometries $h_1, \ldots, h_n$ such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by

$$r_j = h_n \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $1 \leq j \leq n$. Letting R be the matrix whose columns are the vectors r_j, and H_i the matrix associated with h_i, it is clear that

$$R = H_n \cdots H_2 H_1 A,$$

where R is upper triangular and every H_i is either a Householder matrix or the identity. However, $h_i \circ h_i = \mathrm{id}$ for all i, $1 \leq i \leq n$, and so

$$v_j = h_1 \circ h_2 \circ \cdots \circ h_n(r_j)$$

for all j, $1 \leq j \leq n$. But $\rho = h_1 \circ h_2 \circ \cdots \circ h_n$ is an isometry represented by the orthogonal matrix $Q = H_1 H_2 \cdots H_n$. It is clear that $A = QR$, where R is upper triangular. As we noted in Proposition 13.3, the diagonal entries of R can be chosen to be nonnegative. $\square$

Remarks:

(1) Letting

$$A_{k+1} = H_k \cdots H_2 H_1 A,$$

with $A_1 = A$, $1 \leq k \leq n$, the proof of Proposition 13.3 can be interpreted in terms of the computation of the sequence of matrices $A_1, \ldots, A_{n+1} = R$. The matrix A_{k+1} has the shape

$$A_{k+1} = \begin{pmatrix} \times & \times & \times & u_1^{k+1} & \times & \times & \times & \times \\ 0 & \times & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \times & u_k^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+2}^{k+1} & \times & \times & \times & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & u_{n-1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_n^{k+1} & \times & \times & \times & \times \end{pmatrix},$$

where the $(k+1)$th column of the matrix is the vector

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}),$$

and thus

$$u'_{k+1} = \left(u_1^{k+1}, \ldots, u_k^{k+1}\right)$$

and

$$u''_{k+1} = \left(u_{k+1}^{k+1}, u_{k+2}^{k+1}, \ldots, u_n^{k+1}\right).$$

If the last $n-k-1$ entries in column $k+1$ are all zero, there is nothing to do, and we let $H_{k+1} = I$. Otherwise, we kill these $n-k-1$ entries by multiplying A_{k+1} on the left by the Householder matrix H_{k+1} sending

$$\left(0, \ldots, 0, u_{k+1}^{k+1}, \ldots, u_n^{k+1}\right) \quad \text{to} \quad (0, \ldots, 0, r_{k+1,k+1}, 0, \ldots, 0),$$

where $r_{k+1,k+1} = \|(u_{k+1}^{k+1}, \ldots, u_n^{k+1})\|$.

(2) If A is invertible and the diagonal entries of R are positive, it can be shown that Q and R are unique.

(3) If we allow negative diagonal entries in R, the matrix H_n may be omitted ($H_n = I$).

(4) The method allows the computation of the determinant of A. We have

$$\det(A) = (-1)^m r_{1,1} \cdots r_{n,n},$$

where m is the number of Householder matrices (not the identity) among the H_i.

(5) The "condition number" of the matrix A is preserved (see Strang [170], Golub and Van Loan [80], Trefethen and Bau [176], Kincaid and Cheney [102], or Ciarlet [41]). This is very good for numerical stability.

(6) The method also applies to a rectangular $m \times n$ matrix. If $m \geq n$, then R is an $n \times n$ upper triangular matrix and Q is an $m \times n$ matrix such that $Q^\top Q = I_n$.

The following `Matlab` functions implement the QR-factorization method of a real square (possibly singular) matrix A using Householder reflections

The main function `houseqr` computes the upper triangular matrix R obtained by applying Householder reflections to A. It makes use of the function `house`, which computes a unit vector u such that given a vector $x \in \mathbb{R}^p$, the Householder transformation $P = I - 2uu^\top$ sets to zero all entries in x but the first entry x_1. It only applies if $\|x(2:p)\|_1 = |x_2| + \cdots + |x_p| > 0$. Since computations are done in floating point, we use a tolerance factor tol, and if $\|x(2:p)\|_1 \leq tol$, then we return $u = 0$, which indicates that the corresponding Householder transformation is the identity. To make sure that $\|Px\|$ is as large as possible, we pick $uu = x + \mathrm{sign}(x_1)\,\|x\|_2\, e_1$, where $\mathrm{sign}(z) = 1$ if $z \geq 0$ and $\mathrm{sign}(z) = -1$ if $z < 0$. Note that as a result, diagonal entries in R may be negative. We will take care of this issue later.

```
function s = signe(x)
%  if x >= 0, then signe(x) = 1
%  else if x < 0 then signe(x) = -1
%

if x < 0
   s = -1;
else
    s = 1;
end
end
```

```
function [uu, u] = house(x)
% This constructs the unnormalized  vector uu
% defining the Householder reflection that
% zeros all but the first entries in x.
% u is the normalized vector uu/||uu||
%

tol = 2*10^(-15);   % tolerance
uu = x;
p = size(x,1);
% computes l^1-norm of x(2:p,1)
n1 = sum(abs(x(2:p,1)));
if n1 <= tol
   u = zeros(p,1);  uu = u;
else
   l = sqrt(x'*x);  % l^2 norm of x
   uu(1) = x(1) + signe(x(1))*l;
   u = uu/sqrt(uu'*uu);
end
end
```

The Householder transformations are recorded in an array u of $n-1$ vectors. There are more efficient implementations, but for the sake of clarity we present the following version.

```
function [R,  u] = houseqr(A)
%  This function computes the upper triangular R in the QR factorization
%  of A using Householder reflections, and an implicit representation
%  of Q as  a sequence of n - 1 vectors u_i representing Householder
%  reflections

n = size(A, 1);
R = A;
u = zeros(n,n-1);
for i = 1:n-1
    [~, u(i:n,i)] = house(R(i:n,i));
    if u(i:n,i) == zeros(n - i + 1,1)
       R(i+1:n,i) = zeros(n - i,1);
    else
       R(i:n,i:n) = R(i:n,i:n) - 2*u(i:n,i)*(u(i:n,i)'*R(i:n,i:n));
    end
end
end
```

If only R is desired, then `houseqr` does the job. In order to obtain R, we need to compose the Householder transformations. We present a simple method which is not the most efficient (there is a way to avoid multiplying explicity the Householder matrices).

The function `buildhouse` creates a Householder reflection from a vector v.

```
function P = buildhouse(v,i)
% This function builds a Householder reflection
%   [I 0 ]
%   [0 PP]
%   from a Householder reflection
%   PP = I - 2uu*uu'
%   where uu = v(i:n)
%   If uu = 0 then P - I
%

n = size(v,1);
if v(i:n) == zeros(n - i + 1,1)
   P = eye(n);
else
   PP = eye(n - i + 1) - 2*v(i:n)*v(i:n)';
   P = [eye(i-1) zeros(i-1, n - i + 1); zeros(n - i + 1, i - 1) PP];
end
end
```

The function `buildQ` builds the matrix Q in the QR-decomposition of A.

```
function Q = buildQ(u)
% Builds the matrix Q in the QR decomposition
% of an nxn  matrix A using Householder matrices,
% where u is a representation of the n - 1
% Householder reflection by a list u of vectors produced by
% houseqr

n = size(u,1);
Q = buildhouse(u(:,1),1);
for i = 2:n-1
  Q = Q*buildhouse(u(:,i),i);
end
end
```

The function `buildhouseQR` computes a QR-factorization of A. At the end, if some entries on the diagonal of R are negative, it creates a diagonal orthogonal matrix P such that PR has nonnegative diagonal entries, so that $A = (QP)(PR)$ is the desired QR-factorization of A.

```
function [Q,R] = buildhouseQR(A)
%
%   Computes the QR decomposition of a square
%   matrix A (possibly singular) using Householder reflections

n = size(A,1);
[R,u] = houseqr(A);
Q = buildQ(u);
% Produces a matrix R whose diagonal entries are
% nonnegative
P = eye(n);
for i = 1:n
   if R(i,i) < 0
      P(i,i) = -1;
   end
end
Q = Q*P; R = P*R;
end
```

Example 13.1. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}.$$

Running the function `buildhouseQR`, we get

$$Q = \begin{pmatrix} 0.1826 & 0.8165 & 0.4001 & 0.3741 \\ 0.3651 & 0.4082 & -0.2546 & -0.7970 \\ 0.5477 & -0.0000 & -0.6910 & 0.4717 \\ 0.7303 & -0.4082 & 0.5455 & -0.0488 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 5.4772 & 7.3030 & 9.1287 & 10.9545 \\ 0 & 0.8165 & 1.6330 & 2.4495 \\ 0 & -0.0000 & 0.0000 & 0.0000 \\ 0 & -0.0000 & 0 & 0.0000 \end{pmatrix}.$$

Observe that A has rank 2. The reader should check that $A = QR$.

Remark: Curiously, running `Matlab` built-in function `qr`, the same R is obtained (up to column signs) but a different Q is obtained (the last two columns are different).

13.3 Summary

The main concepts and results of this chapter are listed below:

- *Symmetry (or reflection) with respect to F and parallel to G.*
- *Orthogonal symmetry (or reflection) with respect to F and parallel to G; reflections, flips.*
- Hyperplane reflections and *Householder matrices*.
- A key fact about reflections (Proposition 13.2).
- *QR-decomposition in terms of Householder transformations* (Theorem 13.4).

13.4 Problems

Problem 13.1. (1) Given a unit vector $(-\sin\theta, \cos\theta)$, prove that the Householder matrix determined by the vector $(-\sin\theta, \cos\theta)$ is

$$\begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}.$$

Give a geometric interpretation (i.e., why the choice $(-\sin\theta, \cos\theta)$?).

(2) Given any matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

Prove that there is a Householder matrix H such that AH is lower triangular, i.e.,

$$AH = \begin{pmatrix} a' & 0 \\ c' & d' \end{pmatrix}$$

for some $a', c', d' \in \mathbb{R}$.

Problem 13.2. Given a Euclidean space E of dimension n, if h is a reflection about some hyperplane orthogonal to a nonzero vector u and f is any isometry, prove that $f \circ h \circ f^{-1}$ is the reflection about the hyperplane orthogonal to $f(u)$.

Problem 13.3. (1) Given a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

prove that there are Householder matrices G, H such that

$$GAH = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix} = D,$$

where D is a diagonal matrix, iff the following equations hold:

$$(b+c)\cos(\theta+\varphi) = (a-d)\sin(\theta+\varphi),$$
$$(c-b)\cos(\theta-\varphi) = (a+d)\sin(\theta-\varphi).$$

(2) Discuss the solvability of the system. Consider the following cases:

Case 1: $a-d=a+d=0$.

Case 2a: $a-d=b+c=0$, $a+d\neq 0$.

Case 2b: $a-d=0$, $b+c\neq 0$, $a+d\neq 0$.

Case 3a: $a+d=c-b=0$, $a-d\neq 0$.

Case 3b: $a+d=0$, $c-b\neq 0$, $a-d\neq 0$.

Case 4: $a+d\neq 0$, $a-d\neq 0$. Show that the solution in this case is

$$\theta = \frac{1}{2}\left[\arctan\left(\frac{b+c}{a-d}\right) + \arctan\left(\frac{c-b}{a+d}\right)\right],$$
$$\varphi = \frac{1}{2}\left[\arctan\left(\frac{b+c}{a-d}\right) - \arctan\left(\frac{c-b}{a+d}\right)\right].$$

If $b=0$, show that the discussion is simpler: basically, consider $c=0$ or $c\neq 0$.

(3) Expressing everything in terms of $u=\cot\theta$ and $v=\cot\varphi$, show that the equations in (2) become

$$(b+c)(uv-1) = (u+v)(a-d),$$
$$(c-b)(uv+1) = (-u+v)(a+d).$$

Problem 13.4. Let A be an $n\times n$ real invertible matrix.

(1) Prove that $A^\top A$ is symmetric positive definite.

(2) Use the Cholesky factorization $A^\top A = R^\top R$ with R upper triangular with positive diagonal entries to prove that $Q=AR^{-1}$ is orthogonal, so that $A=QR$ is the QR-factorization of A.

Problem 13.5. Modify the function `houseqr` so that it applies to an $m\times n$ matrix with $m\geq n$, to produce an $m\times n$ upper-triangular matrix whose last $m-n$ rows are zeros.

Problem 13.6. The purpose of this problem is to prove that given any self-adjoint linear map $f\colon E\to E$ (i.e., such that $f^*=f$), where E is a Euclidean space of dimension $n\geq 3$, given an orthonormal basis $(e_1,\ldots,e_n)$, there are $n-2$ isometries h_i, hyperplane reflections or the identity, such that the matrix of

$$h_{n-2}\circ\cdots\circ h_1\circ f\circ h_1\circ\cdots\circ h_{n-2}$$

is a symmetric tridiagonal matrix.

(1) Prove that for any isometry $f\colon E \to E$, we have $f = f^* = f^{-1}$ iff $f \circ f = \mathrm{id}$.

Prove that if f and h are self-adjoint linear maps ($f^* = f$ and $h^* = h$), then $h \circ f \circ h$ is a self-adjoint linear map.

(2) Let V_k be the subspace spanned by $(e_{k+1}, \ldots, e_n)$. Proceed by induction. For the base case, proceed as follows.

Let

$$f(e_1) = a_1^0 e_1 + \cdots + a_n^0 e_n,$$

and let

$$r_{1,\,2} = \left\| a_2^0 e_2 + \cdots + a_n^0 e_n \right\|.$$

Find an isometry h_1 (reflection or id) such that

$$h_1(f(e_1) - a_1^0 e_1) = r_{1,\,2}\, e_2.$$

Observe that

$$w_1 = r_{1,\,2}\, e_2 + a_1^0 e_1 - f(e_1) \in V_1,$$

and prove that $h_1(e_1) = e_1$, so that

$$h_1 \circ f \circ h_1(e_1) = a_1^0 e_1 + r_{1,\,2}\, e_2.$$

Let $f_1 = h_1 \circ f \circ h_1$.

Assuming by induction that

$$f_k = h_k \circ \cdots \circ h_1 \circ f \circ h_1 \circ \cdots \circ h_k$$

has a tridiagonal matrix up to the kth row and column, $1 \leq k \leq n-3$, let

$$f_k(e_{k+1}) = a_k^k e_k + a_{k+1}^k e_{k+1} + \cdots + a_n^k e_n,$$

and let

$$r_{k+1,\,k+2} = \left\| a_{k+2}^k e_{k+2} + \cdots + a_n^k e_n \right\|.$$

Find an isometry h_{k+1} (reflection or id) such that

$$h_{k+1}(f_k(e_{k+1}) - a_k^k e_k - a_{k+1}^k e_{k+1}) = r_{k+1,\,k+2}\, e_{k+2}.$$

Observe that

$$w_{k+1} = r_{k+1,\,k+2}\, e_{k+2} + a_k^k e_k + a_{k+1}^k e_{k+1} - f_k(e_{k+1}) \in V_{k+1},$$

and prove that $h_{k+1}(e_k) = e_k$ and $h_{k+1}(e_{k+1}) = e_{k+1}$, so that

$$h_{k+1} \circ f_k \circ h_{k+1}(e_{k+1}) = a_k^k e_k + a_{k+1}^k e_{k+1} + r_{k+1,\,k+2}\, e_{k+2}.$$

Let $f_{k+1} = h_{k+1} \circ f_k \circ h_{k+1}$, and finish the proof.

(3) Prove that given any symmetric $n \times n$-matrix A, there are $n-2$ matrices $H_1, \ldots, H_{n-2}$, Householder matrices or the identity, such that

$$B = H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2}$$

is a symmetric tridiagonal matrix.

(4) Write a computer program implementing the above method.

Problem 13.7. Recall from Problem 6.6 that an $n \times n$ matrix H is *upper Hessenberg* if $h_{jk} = 0$ for all (j, k) such that $j - k \geq 0$. Adapt the proof of Problem 13.6 to prove that given any $n \times n$-matrix A, there are $n - 2 \geq 1$ matrices $H_1, \ldots, H_{n-2}$, Householder matrices or the identity, such that

$$B = H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2}$$

is upper Hessenberg.

Problem 13.8. The purpose of this problem is to prove that given any linear map $f\colon E \to E$, where E is a Euclidean space of dimension $n \geq 2$, given an orthonormal basis $(e_1, \ldots, e_n)$, there are isometries g_i, h_i, hyperplane reflections or the identity, such that the matrix of

$$g_n \circ \cdots \circ g_1 \circ f \circ h_1 \circ \cdots \circ h_n$$

is a lower bidiagonal matrix, which means that the nonzero entries (if any) are on the main descending diagonal and on the diagonal below it.

(1) Let U'_k be the subspace spanned by $(e_1, \ldots, e_k)$ and U''_k be the subspace spanned by $(e_{k+1}, \ldots, e_n)$, $1 \leq k \leq n - 1$. Proceed by induction For the base case, proceed as follows.

Let $v_1 = f^*(e_1)$ and $r_{1,1} = \|v_1\|$. Find an isometry h_1 (reflection or id) such that

$$h_1(f^*(e_1)) = r_{1,1} e_1.$$

Observe that $h_1(f^*(e_1)) \in U'_1$, so that

$$\langle h_1(f^*(e_1)), e_j \rangle = 0$$

for all $j, 2 \leq j \leq n$, and conclude that

$$\langle e_1, f \circ h_1(e_j) \rangle = 0$$

for all $j, 2 \leq j \leq n$.

Next let

$$u_1 = f \circ h_1(e_1) = u'_1 + u''_1,$$

where $u'_1 \in U'_1$ and $u''_1 \in U''_1$, and let $r_{2,1} = \|u''_1\|$. Find an isometry g_1 (reflection or id) such that

$$g_1(u''_1) = r_{2,1} e_2.$$

Show that $g_1(e_1) = e_1$,

$$g_1 \circ f \circ h_1(e_1) = u'_1 + r_{2,\,1}e_2,$$

and that

$$\langle e_1, g_1 \circ f \circ h_1(e_j)\rangle = 0$$

for all $j, 2 \leq j \leq n$. At the end of this stage, show that $g_1 \circ f \circ h_1$ has a matrix such that all entries on its first row except perhaps the first are zero, and that all entries on the first column, except perhaps the first two, are zero.

Assume by induction that some isometries $g_1, \ldots, g_k$ and $h_1, \ldots, h_k$ have been found, either reflections or the identity, and such that

$$f_k = g_k \circ \cdots \circ g_1 \circ f \circ h_1 \cdots \circ h_k$$

has a matrix which is lower bidiagonal up to and including row and column k, where $1 \leq k \leq n-2$.

Let

$$v_{k+1} = f_k^*(e_{k+1}) = v'_{k+1} + v''_{k+1},$$

where $v'_{k+1} \in U'_k$ and $v''_{k+1} \in U''_k$, and let $r_{k+1,\,k+1} = \left\|v''_{k+1}\right\|$. Find an isometry h_{k+1} (reflection or id) such that

$$h_{k+1}(v''_{k+1}) = r_{k+1,\,k+1}e_{k+1}.$$

Show that if h_{k+1} is a reflection, then $U'_k \subseteq H_{k+1}$, where H_{k+1} is the hyperplane defining the reflection h_{k+1}. Deduce that $h_{k+1}(v'_{k+1}) = v'_{k+1}$, and that

$$h_{k+1}(f_k^*(e_{k+1})) = v'_{k+1} + r_{k+1,\,k+1}e_{k+1}.$$

Observe that $h_{k+1}(f_k^*(e_{k+1})) \in U'_{k+1}$, so that

$$\langle h_{k+1}(f_k^*(e_{k+1})), e_j\rangle = 0$$

for all j, $k+2 \leq j \leq n$, and thus,

$$\langle e_{k+1}, f_k \circ h_{k+1}(e_j)\rangle = 0$$

for all j, $k+2 \leq j \leq n$.

Next let

$$u_{k+1} = f_k \circ h_{k+1}(e_{k+1}) = u'_{k+1} + u''_{k+1},$$

where $u'_{k+1} \in U'_{k+1}$ and $u''_{k+1} \in U''_{k+1}$, and let $r_{k+2,\,k+1} = \left\|u''_{k+1}\right\|$. Find an isometry g_{k+1} (reflection or id) such that

$$g_{k+1}(u''_{k+1}) = r_{k+2,\,k+1}e_{k+2}.$$

Show that if g_{k+1} is a reflection, then $U'_{k+1} \subseteq G_{k+1}$, where G_{k+1} is the hyperplane defining the reflection g_{k+1}. Deduce that $g_{k+1}(e_i) = e_i$ for all i, $1 \leq i \leq k+1$, and that

$$g_{k+1} \circ f_k \circ h_{k+1}(e_{k+1}) = u'_{k+1} + r_{k+2,\,k+1}e_{k+2}.$$

Since by induction hypothesis,

$$\langle e_i, f_k \circ h_{k+1}(e_j)\rangle = 0$$

for all i, j, $1 \leq i \leq k+1$, $k+2 \leq j \leq n$, and since $g_{k+1}(e_i) = e_i$ for all i, $1 \leq i \leq k+1$, conclude that

$$\langle e_i, g_{k+1} \circ f_k \circ h_{k+1}(e_j)\rangle = 0$$

for all i, j, $1 \leq i \leq k+1$, $k+2 \leq j \leq n$. Finish the proof.

Chapter 14

Hermitian Spaces

14.1 Sesquilinear and Hermitian Forms, Pre-Hilbert Spaces and Hermitian Spaces

In this chapter we generalize the basic results of Euclidean geometry presented in Chapter 12 to vector spaces over the complex numbers. Such a generalization is inevitable and not simply a luxury. For example, linear maps may not have real eigenvalues, but they always have complex eigenvalues. Furthermore, some very important classes of linear maps can be diagonalized if they are extended to the complexification of a real vector space. This is the case for orthogonal matrices and, more generally, normal matrices. Also, complex vector spaces are often the natural framework in physics or engineering, and they are more convenient for dealing with Fourier series. However, some complications arise due to complex conjugation.

Recall that for any complex number $z \in \mathbb{C}$, if $z = x + iy$ where $x, y \in \mathbb{R}$, we let $\Re z = x$, the real part of z, and $\Im z = y$, the imaginary part of z. We also denote the conjugate of $z = x + iy$ by $\overline{z} = x - iy$, and the absolute value (or length, or modulus) of z by $|z|$. Recall that $|z|^2 = z\overline{z} = x^2 + y^2$.

There are many natural situations where a map $\varphi\colon E \times E \to \mathbb{C}$ is linear in its first argument and only semilinear in its second argument, which means that $\varphi(u, \mu v) = \overline{\mu}\varphi(u, v)$, as opposed to $\varphi(u, \mu v) = \mu\varphi(u, v)$. For example, the natural inner product to deal with functions $f \colon \mathbb{R} \to \mathbb{C}$, especially Fourier series, is

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

which is semilinear (but not linear) in g. Thus, when generalizing a result from the real case of a Euclidean space to the complex case, we always have to check very carefully that our proofs do not rely on linearity in the second argument. Otherwise, we need to revise our proofs, and sometimes the result is simply wrong!

Before defining the natural generalization of an inner product, it is convenient to define semilinear maps.

Definition 14.1. Given two vector spaces E and F over the complex field $\mathbb{C}$, a function $f\colon E \to F$ is *semilinear* if

$$\begin{aligned} f(u+v) &= f(u) + f(v), \\ f(\lambda u) &= \overline{\lambda} f(u), \end{aligned}$$

for all $u, v \in E$ and all $\lambda \in \mathbb{C}$.

Remark: Instead of defining semilinear maps, we could have defined the vector space $\overline{E}$ as the vector space with the same carrier set E whose addition is the same as that of E, but whose multiplication by a complex number is given by

$$(\lambda, u) \mapsto \overline{\lambda} u.$$

Then it is easy to check that a function $f\colon E \to \mathbb{C}$ is semilinear iff $f\colon \overline{E} \to \mathbb{C}$ is linear.

We can now define sesquilinear forms and Hermitian forms.

Definition 14.2. Given a complex vector space E, a function $\varphi\colon E \times E \to \mathbb{C}$ is a *sesquilinear form* if it is linear in its first argument and semilinear in its second argument, which means that

$$\begin{aligned} \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v), \\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2), \\ \varphi(\lambda u, v) &= \lambda \varphi(u, v), \\ \varphi(u, \mu v) &= \overline{\mu} \varphi(u, v), \end{aligned}$$

for all $u, v,\ u_1, u_2,\ v_1, v_2 \in E$, and all $\lambda, \mu \in \mathbb{C}$. A function $\varphi\colon E \times E \to \mathbb{C}$ is a *Hermitian form* if it is sesquilinear and if

$$\varphi(v, u) = \overline{\varphi(u, v)}$$

for all all $u, v \in E$.

Obviously, $\varphi(0, v) = \varphi(u, 0) = 0$. Also note that if $\varphi\colon E \times E \to \mathbb{C}$ is sesquilinear, we have

$$\varphi(\lambda u + \mu v, \lambda u + \mu v) = |\lambda|^2 \varphi(u, u) + \lambda \overline{\mu} \varphi(u, v) + \overline{\lambda} \mu \varphi(v, u) + |\mu|^2 \varphi(v, v),$$

and if $\varphi\colon E \times E \to \mathbb{C}$ is Hermitian, we have

$$\varphi(\lambda u + \mu v, \lambda u + \mu v) = |\lambda|^2 \varphi(u, u) + 2\Re(\lambda \overline{\mu} \varphi(u, v)) + |\mu|^2 \varphi(v, v).$$

Note that restricted to real coefficients, a sesquilinear form is bilinear (we sometimes say $\mathbb{R}$-bilinear).

Definition 14.3. Given a sesquilinear form $\varphi\colon E \times E \to \mathbb{C}$, the function $\Phi\colon E \to \mathbb{C}$ defined such that $\Phi(u) = \varphi(u, u)$ for all $u \in E$ is called the *quadratic form* associated with φ.

The standard example of a Hermitian form on $\mathbb{C}^n$ is the map φ defined such that

$$\varphi((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = x_1\overline{y_1} + x_2\overline{y_2} + \cdots + x_n\overline{y_n}.$$

This map is also positive definite, but before dealing with these issues, we show the following useful proposition.

Proposition 14.1. *Given a complex vector space E, the following properties hold:*

(1) A sesquilinear form $\varphi\colon E \times E \to \mathbb{C}$ is a Hermitian form iff $\varphi(u, u) \in \mathbb{R}$ for all $u \in E$.

(2) If $\varphi\colon E \times E \to \mathbb{C}$ is a sesquilinear form, then

$$\begin{aligned} 4\varphi(u, v) &= \varphi(u + v, u + v) - \varphi(u - v, u - v) \\ &\quad + i\varphi(u + iv, u + iv) - i\varphi(u - iv, u - iv), \end{aligned}$$

and

$$2\varphi(u, v) = (1 + i)(\varphi(u, u) + \varphi(v, v)) - \varphi(u - v, u - v) - i\varphi(u - iv, u - iv).$$

These are called **polarization identities**.

Proof. (1) If φ is a Hermitian form, then

$$\varphi(v, u) = \overline{\varphi(u, v)}$$

implies that

$$\varphi(u, u) = \overline{\varphi(u, u)},$$

and thus $\varphi(u, u) \in \mathbb{R}$. If φ is sesquilinear and $\varphi(u, u) \in \mathbb{R}$ for all $u \in E$, then

$$\varphi(u + v, u + v) = \varphi(u, u) + \varphi(u, v) + \varphi(v, u) + \varphi(v, v),$$

which proves that

$$\varphi(u, v) + \varphi(v, u) = \alpha,$$

where α is real, and changing u to iu, we have

$$i(\varphi(u, v) - \varphi(v, u)) = \beta,$$

where β is real, and thus

$$\varphi(u, v) = \frac{\alpha - i\beta}{2} \quad \text{and} \quad \varphi(v, u) = \frac{\alpha + i\beta}{2},$$

proving that φ is Hermitian.

(2) These identities are verified by expanding the right-hand side, and we leave them as an exercise. $\square$

Proposition 14.1 shows that a sesquilinear form is completely determined by the quadratic form $\Phi(u) = \varphi(u, u)$, even if φ is not Hermitian. This is false for a real bilinear form, unless it is symmetric. For example, the bilinear form $\varphi\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined such that

$$\varphi((x_1, y_1), (x_2, y_2)) = x_1 y_2 - x_2 y_1$$

is not identically zero, and yet it is null on the diagonal. However, a real symmetric bilinear form is indeed determined by its values on the diagonal, as we saw in Chapter 12.

As in the Euclidean case, Hermitian forms for which $\varphi(u, u) \geq 0$ play an important role.

Definition 14.4. Given a complex vector space E, a Hermitian form $\varphi\colon E \times E \to \mathbb{C}$ is *positive* if $\varphi(u, u) \geq 0$ for all $u \in E$, and *positive definite* if $\varphi(u, u) > 0$ for all $u \neq 0$. A pair $\langle E, \varphi \rangle$ where E is a complex vector space and φ is a Hermitian form on E is called a *pre-Hilbert space* if φ is positive, and a *Hermitian (or unitary) space* if φ is positive definite.

We warn our readers that some authors, such as Lang [111], define a pre-Hilbert space as what we define as a Hermitian space. We prefer following the terminology used in Schwartz [150] and Bourbaki [27]. The quantity $\varphi(u, v)$ is usually called the *Hermitian product* of u and v. We will occasionally call it the inner product of u and v.

Given a pre-Hilbert space $\langle E, \varphi \rangle$, as in the case of a Euclidean space, we also denote $\varphi(u, v)$ by

$$u \cdot v \quad \text{or} \quad \langle u, v \rangle \quad \text{or} \quad (u|v),$$

and $\sqrt{\Phi(u)}$ by $\|u\|$.

Example 14.1. The complex vector space $\mathbb{C}^n$ under the Hermitian form

$$\varphi((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = x_1 \overline{y_1} + x_2 \overline{y_2} + \cdots + x_n \overline{y_n}$$

is a Hermitian space.

Example 14.2. Let ℓ^2 denote the set of all countably infinite sequences $x = (x_i)_{i \in \mathbb{N}}$ of complex numbers such that $\sum_{i=0}^{\infty} |x_i|^2$ is defined (i.e., the sequence $\sum_{i=0}^{n} |x_i|^2$ converges as $n \to \infty$). It can be shown that the map $\varphi\colon \ell^2 \times \ell^2 \to \mathbb{C}$ defined such that

$$\varphi\left((x_i)_{i \in \mathbb{N}}, (y_i)_{i \in \mathbb{N}}\right) = \sum_{i=0}^{\infty} x_i \overline{y_i}$$

is well defined, and ℓ^2 is a Hermitian space under φ. Actually, ℓ^2 is even a Hilbert space.

Example 14.3. Let $\mathcal{C}_{\mathrm{piece}}[a, b]$ be the set of bounded piecewise continuous functions $f\colon [a, b] \to \mathbb{C}$ under the Hermitian form

$$\langle f, g \rangle = \int_a^b f(x) \overline{g(x)} dx.$$

It is easy to check that this Hermitian form is positive, but it is not definite. Thus, under this Hermitian form, $\mathcal{C}_{\mathrm{piece}}[a, b]$ is only a pre-Hilbert space.

Example 14.4. Let $\mathcal{C}[a,b]$ be the set of complex-valued continuous functions $f\colon [a,b] \to \mathbb{C}$ under the Hermitian form

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx.$$

It is easy to check that this Hermitian form is positive definite. Thus, $\mathcal{C}[a,b]$ is a Hermitian space.

Example 14.5. Let $E = \mathrm{M}_n(\mathbb{C})$ be the vector space of complex $n \times n$ matrices. If we view a matrix $A \in \mathrm{M}_n(\mathbb{C})$ as a "long" column vector obtained by concatenating together its columns, we can define the Hermitian product of two matrices $A, B \in \mathrm{M}_n(\mathbb{C})$ as

$$\langle A, B\rangle = \sum_{i,j=1}^{n} a_{ij}\bar{b}_{ij},$$

which can be conveniently written as

$$\langle A, B\rangle = \mathrm{tr}(A^{\top}\overline{B}) = \mathrm{tr}(B^* A).$$

Since this can be viewed as the standard Hermitian product on $\mathbb{C}^{n^2}$, it is a Hermitian product on $\mathrm{M}_n(\mathbb{C})$. The corresponding norm

$$\|A\|_F = \sqrt{\mathrm{tr}(A^*A)}$$

is the Frobenius norm (see Section 9.2).

If E is finite-dimensional and if $\varphi\colon E \times E \to \mathbb{R}$ is a sequilinear form on E, given any basis $(e_1, \ldots, e_n)$ of E, we can write $x = \sum_{i=1}^n x_i e_i$ and $y = \sum_{j=1}^n y_j e_j$, and we have

$$\varphi(x, y) = \varphi\left(\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{n} y_j e_j\right) = \sum_{i,j=1}^{n} x_i \overline{y}_j \varphi(e_i, e_j).$$

If we let $G = (g_{ij})$ be the matrix given by $g_{ij} = \varphi(e_j, e_i)$, and if x and y are the column vectors associated with $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, then we can write

$$\varphi(x, y) = x^{\top} G^{\top}\, \overline{y} = y^* G x,$$

where $\overline{y}$ corresponds to $(\overline{y}_1, \ldots, \overline{y}_n)$. As in Section 12.1, we are committing the slight abuse of notation of letting x denote both the vector $x = \sum_{i=1}^n x_i e_i$ and the column vector associated with $(x_1, \ldots, x_n)$ (and similarly for y). The "correct" expression for $\varphi(x, y)$ is

$$\varphi(x, y) = \mathbf{y}^* G \mathbf{x} = \mathbf{x}^{\top} G^{\top} \overline{\mathbf{y}}.$$

Observe that in $\varphi(x, y) = y^* G x$, the matrix involved is the transpose of the matrix $(\varphi(e_i, e_j))$. The reason for this is that we want G to be positive definite when φ is positive definite, not $G^{\top}$.

Furthermore, observe that φ is Hermitian iff $G = G^*$, and φ is positive definite iff the matrix G is positive definite, that is,

$$(Gx)^\top \overline{x} = x^* G x > 0 \quad \text{for all } x \in \mathbb{C}^n,\ x \neq 0.$$

Definition 14.5. The matrix G associated with a Hermitian product is called the *Gram matrix* of the Hermitian product with respect to the basis $(e_1, \ldots, e_n)$.

Conversely, if A is a Hermitian positive definite $n \times n$ matrix, it is easy to check that the Hermitian form

$$\langle x, y\rangle = y^* A x$$

is positive definite. If we make a change of basis from the basis $(e_1, \ldots, e_n)$ to the basis $(f_1, \ldots, f_n)$, and if the change of basis matrix is P (where the jth column of P consists of the coordinates of f_j over the basis $(e_1, \ldots, e_n)$), then with respect to coordinates x' and y' over the basis $(f_1, \ldots, f_n)$, we have

$$y^* G x = (y')^* P^* G P x',$$

so the matrix of our inner product over the basis $(f_1, \ldots, f_n)$ is $P^* G P$. We summarize these facts in the following proposition.

Proposition 14.2. *Let E be a finite-dimensional vector space, and let $(e_1, \ldots, e_n)$ be a basis of E.*

1. *For any Hermitian inner product $\langle -, -\rangle$ on E, if $G = (g_{ij})$ with $g_{ij} = \langle e_j, e_i\rangle$ is the Gram matrix of the Hermitian product $\langle -, -\rangle$ w.r.t. the basis $(e_1, \ldots, e_n)$, then G is Hermitian positive definite.*

2. *For any change of basis matrix P, the Gram matrix of $\langle -, -\rangle$ with respect to the new basis is $P^* G P$.*

3. *If A is any $n \times n$ Hermitian positive definite matrix, then*

$$\langle x, y\rangle = y^* A x$$

is a Hermitian product on E.

We will see later that a Hermitian matrix is positive definite iff its eigenvalues are all positive.

The following result reminiscent of the first polarization identity of Proposition 14.1 can be used to prove that two linear maps are identical.

Proposition 14.3. *Given any Hermitian space E with Hermitian product $\langle -, -\rangle$, for any linear map $f \colon E \to E$, if $\langle f(x), x\rangle = 0$ for all $x \in E$, then $f = 0$.*

Proof. Compute $\langle f(x+y), x+y\rangle$ and $\langle f(x-y), x-y\rangle$:

$$\begin{aligned}\langle f(x+y), x+y\rangle &= \langle f(x), x\rangle + \langle f(x), y\rangle + \langle f(y), x\rangle + \langle y, y\rangle\\ \langle f(x-y), x-y\rangle &= \langle f(x), x\rangle - \langle f(x), y\rangle - \langle f(y), x\rangle + \langle y, y\rangle;\end{aligned}$$

then subtract the second equation from the first to obtain

$$\langle f(x+y), x+y\rangle - \langle f(x-y), x-y\rangle = 2(\langle f(x), y\rangle + \langle f(y), x\rangle).$$

If $\langle f(u), u\rangle = 0$ for all $u \in E$, we get

$$\langle f(x), y\rangle + \langle f(y), x\rangle = 0 \quad \text{for all } x, y \in E.$$

Then the above equation also holds if we replace x by ix, and we obtain

$$i\langle f(x), y\rangle - i\langle f(y), x\rangle = 0, \quad \text{for all } x, y \in E,$$

so we have

$$\begin{aligned}\langle f(x), y\rangle + \langle f(y), x\rangle &= 0\\ \langle f(x), y\rangle - \langle f(y), x\rangle &= 0,\end{aligned}$$

which implies that $\langle f(x), y\rangle = 0$ for all $x, y \in E$. Since $\langle -, -\rangle$ is positive definite, we have $f(x) = 0$ for all $x \in E$; that is, $f = 0$. ☐

One should be careful not to apply Proposition 14.3 to a linear map on a real Euclidean space because it is false! The reader should find a counterexample.

The Cauchy–Schwarz inequality and the Minkowski inequalities extend to pre-Hilbert spaces and to Hermitian spaces.

Proposition 14.4. *Let $\langle E, \varphi\rangle$ be a pre-Hilbert space with associated quadratic form Φ. For all $u, v \in E$, we have the* Cauchy–Schwarz inequality

$$|\varphi(u, v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Furthermore, if $\langle E, \varphi\rangle$ is a Hermitian space, the equality holds iff u and v are linearly dependent.

We also have the Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}.$$

Furthermore, if $\langle E, \varphi\rangle$ is a Hermitian space, the equality holds iff u and v are linearly dependent, where in addition, if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some real λ such that $\lambda > 0$.

Proof. For all $u, v \in E$ and all $\mu \in \mathbb{C}$, we have observed that

$$\varphi(u + \mu v, u + \mu v) = \varphi(u, u) + 2\Re(\overline{\mu}\varphi(u, v)) + |\mu|^2\varphi(v, v).$$

Let $\varphi(u, v) = \rho e^{i\theta}$, where $|\varphi(u, v)| = \rho$ ($\rho \geq 0$). Let $F \colon \mathbb{R} \to \mathbb{R}$ be the function defined such that

$$F(t) = \Phi(u + te^{i\theta}v),$$

for all $t \in \mathbb{R}$. The above shows that

$$F(t) = \varphi(u, u) + 2t|\varphi(u, v)| + t^2\varphi(v, v) = \Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v).$$

Since φ is assumed to be positive, we have $F(t) \geq 0$ for all $t \in \mathbb{R}$. If $\Phi(v) = 0$, we must have $\varphi(u, v) = 0$, since otherwise, $F(t)$ could be made negative by choosing t negative and small enough. If $\Phi(v) > 0$, in order for $F(t)$ to be nonnegative, the equation

$$\Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v) = 0$$

must not have distinct real roots, which is equivalent to

$$|\varphi(u, v)|^2 \leq \Phi(u)\Phi(v).$$

Taking the square root on both sides yields the Cauchy–Schwarz inequality.

For the second part of the claim, if φ is positive definite, we argue as follows. If u and v are linearly dependent, it is immediately verified that we get an equality. Conversely, if

$$|\varphi(u, v)|^2 = \Phi(u)\Phi(v),$$

then there are two cases. If $\Phi(v) = 0$, since φ is positive definite, we must have $v = 0$, so u and v are linearly dependent. Otherwise, the equation

$$\Phi(u) + 2t|\varphi(u, v)| + t^2\Phi(v) = 0$$

has a double root t_0, and thus

$$\Phi(u + t_0e^{i\theta}v) = 0.$$

Since φ is positive definite, we must have

$$u + t_0e^{i\theta}v = 0,$$

which shows that u and v are linearly dependent.

If we square the Minkowski inequality, we get

$$\Phi(u + v) \leq \Phi(u) + \Phi(v) + 2\sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

However, we observed earlier that

$$\Phi(u + v) = \Phi(u) + \Phi(v) + 2\Re(\varphi(u, v)).$$

Thus, it is enough to prove that

$$\Re(\varphi(u,v)) \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

but this follows from the Cauchy–Schwarz inequality

$$|\varphi(u,v)| \leq \sqrt{\Phi(u)}\sqrt{\Phi(v)}$$

and the fact that $\Re z \leq |z|$.

If φ is positive definite and u and v are linearly dependent, it is immediately verified that we get an equality. Conversely, if equality holds in the Minkowski inequality, we must have

$$\Re(\varphi(u,v)) = \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

which implies that

$$|\varphi(u,v)| = \sqrt{\Phi(u)}\sqrt{\Phi(v)},$$

since otherwise, by the Cauchy–Schwarz inequality, we would have

$$\Re(\varphi(u,v)) \leq |\varphi(u,v)| < \sqrt{\Phi(u)}\sqrt{\Phi(v)}.$$

Thus, equality holds in the Cauchy–Schwarz inequality, and

$$\Re(\varphi(u,v)) = |\varphi(u,v)|.$$

But then we proved in the Cauchy–Schwarz case that u and v are linearly dependent. Since we also just proved that $\varphi(u,v)$ is real and nonnegative, the coefficient of proportionality between u and v is indeed nonnegative. $\square$

As in the Euclidean case, if $\langle E, \varphi\rangle$ is a Hermitian space, the Minkowski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

shows that the map $u \mapsto \sqrt{\Phi(u)}$ is a *norm* on E. The norm induced by φ is called the *Hermitian norm induced by* φ. We usually denote $\sqrt{\Phi(u)}$ by $\|u\|$, and the Cauchy–Schwarz inequality is written as

$$|u \cdot v| \leq \|u\|\|v\|.$$

Since a Hermitian space is a normed vector space, it is a topological space under the topology induced by the norm (a basis for this topology is given by the open balls $B_0(u,\rho)$ of center u and radius $\rho > 0$, where

$$B_0(u,\rho) = \{v \in E \mid \|v - u\| < \rho\}.$$

If E has finite dimension, every linear map is continuous; see Chapter 9 (or Lang [111, 112], Dixmier [51], or Schwartz [150, 151]). The Cauchy–Schwarz inequality

$$|u \cdot v| \leq \|u\|\|v\|$$

shows that $\varphi\colon E \times E \to \mathbb{C}$ is continuous, and thus, that $\| \; \|$ is continuous.

If $\langle E, \varphi \rangle$ is only pre-Hilbertian, $\|u\|$ is called a *seminorm*. In this case, the condition

$$\|u\| = 0 \quad \text{implies} \quad u = 0$$

is not necessarily true. However, the Cauchy–Schwarz inequality shows that if $\|u\| = 0$, then $u \cdot v = 0$ for all $v \in E$.

Remark: As in the case of real vector spaces, a norm on a complex vector space is induced by some positive definite Hermitian product $\langle -, - \rangle$ iff it satisfies the *parallelogram law*:

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

This time the Hermitian product is recovered using the polarization identity from Proposition 14.1:

$$4\langle u, v \rangle = \|u + v\|^2 - \|u - v\|^2 + i \, \|u + iv\|^2 - i \, \|u - iv\|^2 .$$

It is easy to check that $\langle u, u \rangle = \|u\|^2$, and

$$\begin{aligned} \langle v, u \rangle &= \overline{\langle u, v \rangle} \\ \langle iu, v \rangle &= i \langle u, v \rangle, \end{aligned}$$

so it is enough to check linearity in the variable u, and only for real scalars. This is easily done by applying the proof from Section 12.1 to the real and imaginary part of $\langle u, v \rangle$; the details are left as an exercise.

We will now basically mirror the presentation of Euclidean geometry given in Chapter 12 rather quickly, leaving out most proofs, except when they need to be seriously amended.

14.2 Orthogonality, Duality, Adjoint of a Linear Map

In this section we assume that we are dealing with Hermitian spaces. We denote the Hermitian inner product by $u \cdot v$ or $\langle u, v \rangle$. The concepts of orthogonality, orthogonal family of vectors, orthonormal family of vectors, and orthogonal complement of a set of vectors are unchanged from the Euclidean case (Definition 12.2).

For example, the set $\mathcal{C}[-\pi, \pi]$ of continuous functions $f\colon [-\pi, \pi] \to \mathbb{C}$ is a Hermitian space under the product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx,$$

and the family $(e^{ikx})_{k \in \mathbb{Z}}$ is orthogonal.

Propositions 12.4 and 12.5 hold without any changes. It is easy to show that

$$\left\| \sum_{i=1}^{n} u_i \right\|^2 = \sum_{i=1}^{n} \|u_i\|^2 + \sum_{1 \leq i < j \leq n} 2\Re(u_i \cdot u_j).$$

Analogously to the case of Euclidean spaces of finite dimension, the Hermitian product induces a canonical bijection (i.e., independent of the choice of bases) between the vector space E and the space E^*. This is one of the places where conjugation shows up, but in this case, troubles are minor.

Given a Hermitian space E, for any vector $u \in E$, let $\varphi^l_u \colon E \to \mathbb{C}$ be the map defined such that

$$\varphi^l_u(v) = \overline{u \cdot v}, \quad \text{for all } v \in E.$$

Similarly, for any vector $v \in E$, let $\varphi^r_v \colon E \to \mathbb{C}$ be the map defined such that

$$\varphi^r_v(u) = u \cdot v, \quad \text{for all } u \in E.$$

Since the Hermitian product is linear in its first argument u, the map φ^r_v is a linear form in E^*, and since it is semilinear in its second argument v, the map φ^l_u is also a linear form in E^*. Thus, we have two maps $\flat^l \colon E \to E^*$ and $\flat^r \colon E \to E^*$, defined such that

$$\flat^l(u) = \varphi^l_u, \quad \text{and} \quad \flat^r(v) = \varphi^r_v.$$

Proposition 14.5. *The equations $\varphi^l_u = \varphi^r_u$ and $\flat^l = \flat^r$ hold.*

Proof. Indeed, for all $u, v \in E$, we have

$$\begin{aligned} \flat^l(u)(v) &= \varphi^l_u(v) \\ &= \overline{u \cdot v} \\ &= v \cdot u \\ &= \varphi^r_u(v) \\ &= \flat^r(u)(v). \end{aligned}$$

□

Therefore, we use the notation φ_u for both φ^l_u and φ^r_u, and $\flat$ for both $\flat^l$ and $\flat^r$.

Theorem 14.6. *Let E be a Hermitian space E. The map $\flat \colon E \to E^*$ defined such that*

$$\flat(u) = \varphi^l_u = \varphi^r_u \quad \textit{for all } u \in E$$

is semilinear and injective. When E is also of finite dimension, the map $\flat \colon \overline{E} \to E^*$ is a canonical isomorphism.

Proof. That $\flat \colon E \to E^*$ is a semilinear map follows immediately from the fact that $\flat = \flat^r$, and that the Hermitian product is semilinear in its second argument. If $\varphi_u = \varphi_v$, then $\varphi_u(w) = \varphi_v(w)$ for all $w \in E$, which by definition of φ_u and φ_v means that

$$w \cdot u = w \cdot v$$

for all $w \in E$, which by semilinearity on the right is equivalent to

$$w \cdot (v - u) = 0 \quad \text{for all } w \in E,$$

which implies that $u = v$, since the Hermitian product is positive definite. Thus, $\flat \colon E \to E^*$ is injective. Finally, when E is of finite dimension n, E^* is also of dimension n, and then $\flat \colon E \to E^*$ is bijective. Since $\flat$ is semilinar, the map $\flat \colon \overline{E} \to E^*$ is an isomorphism. □

The inverse of the isomorphism $\flat\colon \overline{E} \to E^*$ is denoted by $\sharp\colon E^* \to \overline{E}$.

As a corollary of the isomorphism $\flat\colon \overline{E} \to E^*$ we have the following result.

Proposition 14.7. *If E is a Hermitian space of finite dimension, then every linear form $f \in E^*$ corresponds to a unique $v \in E$, such that*

$$f(u) = u \cdot v, \quad \textit{for every } u \in E.$$

In particular, if f is not the zero form, the kernel of f, which is a hyperplane H, is precisely the set of vectors that are orthogonal to v.

Remarks:

1. The "musical map" $\flat\colon \overline{E} \to E^*$ is not surjective when E has infinite dimension. This result can be salvaged by restricting our attention to continuous linear maps and by assuming that the vector space E is a *Hilbert space*.

2. *Dirac's "bra-ket" notation.* Dirac invented a notation widely used in quantum mechanics for denoting the linear form $\varphi_u = \flat(u)$ associated to the vector $u \in E$ *via* the duality induced by a Hermitian inner product. Dirac's proposal is to denote the vectors u in E by $|u\rangle$, and call them *kets*; the notation $|u\rangle$ is pronounced "ket u." Given two kets (vectors) $|u\rangle$ and $|v\rangle$, their inner product is denoted by

$$\langle u|v\rangle$$

(instead of $|u\rangle \cdot |v\rangle$). The notation $\langle u|v\rangle$ for the inner product of $|u\rangle$ and $|v\rangle$ anticipates duality. Indeed, we define the dual (usually called adjoint) *bra* u of ket u, denoted by $\langle u|$, as the linear form whose value on any ket v is given by the inner product, so

$$\langle u|(|v\rangle) = \langle u|v\rangle.$$

Thus, bra $u = \langle u|$ is Dirac's notation for our $\flat(u)$. Since the map $\flat$ is semi-linear, we have

$$\langle \lambda u| = \overline{\lambda}\langle u|.$$

Using the bra-ket notation, given an orthonormal basis $(|u_1\rangle, \ldots, |u_n\rangle)$, ket v (a vector) is written as

$$|v\rangle = \sum_{i=1}^{n} \langle v|u_i\rangle |u_i\rangle,$$

and the corresponding linear form bra v is written as

$$\langle v| = \sum_{i=1}^{n} \overline{\langle v|u_i\rangle} \langle u_i| = \sum_{i=1}^{n} \langle u_i|v\rangle \langle u_i|$$

over the dual basis $(\langle u_1|, \ldots, \langle u_n|)$. As cute as it looks, we do not recommend using the Dirac notation.

The existence of the isomorphism $\flat\colon \overline{E} \to E^*$ is crucial to the existence of adjoint maps. Indeed, Theorem 14.6 allows us to define the adjoint of a linear map on a Hermitian space. Let E be a Hermitian space of finite dimension n, and let $f\colon E \to E$ be a linear map. For every $u \in E$, the map

$$v \mapsto \overline{u \cdot f(v)}$$

is clearly a linear form in E^*, and by Theorem 14.6, there is a unique vector in E denoted by $f^*(u)$, such that

$$\overline{f^*(u) \cdot v} = \overline{u \cdot f(v)},$$

that is,

$$f^*(u) \cdot v = u \cdot f(v), \quad \text{for every } v \in E.$$

The following proposition shows that the map f^* is linear.

Proposition 14.8. *Given a Hermitian space E of finite dimension, for every linear map $f\colon E \to E$ there is a unique linear map $f^*\colon E \to E$ such that*

$$f^*(u) \cdot v = u \cdot f(v), \quad \textit{for all } u, v \in E.$$

Proof. Careful inspection of the proof of Proposition 12.8 reveals that it applies unchanged. The only potential problem is in proving that $f^*(\lambda u) = \lambda f^*(u)$, but everything takes place in the first argument of the Hermitian product, and there, we have linearity. $\square$

Definition 14.6. Given a Hermitian space E of finite dimension, for every linear map $f\colon E \to E$, the unique linear map $f^*\colon E \to E$ such that

$$f^*(u) \cdot v = u \cdot f(v), \quad \text{for all } u, v \in E$$

given by Proposition 14.8 is called the *adjoint of f (w.r.t. to the Hermitian product)*.

The fact that

$$v \cdot u = \overline{u \cdot v}$$

implies that the adjoint f^* of f is also characterized by

$$f(u) \cdot v = u \cdot f^*(v),$$

for all $u, v \in E$.

Given two Hermitian spaces E and F, where the Hermitian product on E is denoted by $\langle -, - \rangle_1$ and the Hermitian product on F is denoted by $\langle -, - \rangle_2$, given any linear map $f\colon E \to F$, it is immediately verified that the proof of Proposition 14.8 can be adapted to show that there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the *adjoint* of f.

As in the Euclidean case, the following properties immediately follow from the definition of the adjoint map.

Proposition 14.9. *(1) For any linear map $f\colon E \to F$, we have*

$$f^{**} = f.$$

(2) For any two linear maps $f, g\colon E \to F$ and any scalar $\lambda \in \mathbb{R}$:

$$\begin{aligned}(f+g)^* &= f^* + g^* \\ (\lambda f)^* &= \overline{\lambda} f^*.\end{aligned}$$

(3) If E, F, G are Hermitian spaces with respective inner products $\langle -,-\rangle_1, \langle -,-\rangle_2$, and $\langle -,-\rangle_3$, and if $f\colon E \to F$ and $g\colon F \to G$ are two linear maps, then

$$(g \circ f)^* = f^* \circ g^*.$$

As in the Euclidean case, a linear map $f\colon E \to E$ (where E is a finite-dimensional Hermitian space) is *self-adjoint* if $f = f^*$. The map f is *positive semidefinite* iff

$$\langle f(x), x\rangle \geq 0 \quad \text{all } x \in E;$$

positive definite iff

$$\langle f(x), x\rangle > 0 \quad \text{all } x \in E,\ x \neq 0.$$

An interesting corollary of Proposition 14.3 is that a positive semidefinite linear map must be self-adjoint. In fact, we can prove a slightly more general result.

Proposition 14.10. *Given any finite-dimensional Hermitian space E with Hermitian product $\langle -,-\rangle$, for any linear map $f\colon E \to E$, if $\langle f(x), x\rangle \in \mathbb{R}$ for all $x \in E$, then f is self-adjoint. In particular, any positive semidefinite linear map $f\colon E \to E$ is self-adjoint.*

Proof. Since $\langle f(x), x\rangle \in \mathbb{R}$ for all $x \in E$, we have

$$\begin{aligned}\langle f(x), x\rangle &= \overline{\langle f(x), x\rangle} \\ &= \langle x, f(x)\rangle \\ &= \langle f^*(x), x\rangle,\end{aligned}$$

so we have

$$\langle (f - f^*)(x), x\rangle = 0 \quad \text{all } x \in E,$$

and Proposition 14.3 implies that $f - f^* = 0$. ☐

Beware that Proposition 14.10 is false if E is a real Euclidean space.

As in the Euclidean case, Theorem 14.6 can be used to show that any Hermitian space of finite dimension has an orthonormal basis. The proof is unchanged.

Proposition 14.11. *Given any nontrivial Hermitian space E of finite dimension $n \geq 1$, there is an orthonormal basis $(u_1, \ldots, u_n)$ for E.*

The *Gram–Schmidt orthonormalization procedure* also applies to Hermitian spaces of finite dimension, without any changes from the Euclidean case!

Proposition 14.12. *Given a nontrivial Hermitian space E of finite dimension $n \geq 1$, from any basis $(e_1, \ldots, e_n)$ for E we can construct an orthonormal basis $(u_1, \ldots, u_n)$ for E with the property that for every k, $1 \leq k \leq n$, the families $(e_1, \ldots, e_k)$ and $(u_1, \ldots, u_k)$ generate the same subspace.*

Remark: The remarks made after Proposition 12.10 also apply here, except that in the QR-decomposition, Q is a unitary matrix.

As a consequence of Proposition 12.9 (or Proposition 14.12), given any Hermitian space of finite dimension n, if $(e_1, \ldots, e_n)$ is an orthonormal basis for E, then for any two vectors $u = u_1e_1 + \cdots + u_ne_n$ and $v = v_1e_1 + \cdots + v_ne_n$, the Hermitian product $u \cdot v$ is expressed as

$$u \cdot v = (u_1e_1 + \cdots + u_ne_n) \cdot (v_1e_1 + \cdots + v_ne_n) = \sum_{i=1}^{n} u_i\overline{v_i},$$

and the norm $\|u\|$ as

$$\|u\| = \|u_1e_1 + \cdots + u_ne_n\| = \left(\sum_{i=1}^{n} |u_i|^2\right)^{1/2}.$$

The fact that a Hermitian space always has an orthonormal basis implies that any Gram matrix G can be written as

$$G = Q^*Q,$$

for some invertible matrix Q. Indeed, we know that in a change of basis matrix, a Gram matrix G becomes $G' = P^*GP$. If the basis corresponding to G' is orthonormal, then $G' = I$, so $G = (P^{-1})^*P^{-1}$.

Proposition 12.11 also holds unchanged.

Proposition 14.13. *Given any nontrivial Hermitian space E of finite dimension $n \geq 1$, for any subspace F of dimension k, the orthogonal complement $F^{\perp}$ of F has dimension $n - k$, and $E = F \oplus F^{\perp}$. Furthermore, we have $F^{\perp\perp} = F$.*

14.3 Linear Isometries (Also Called Unitary Transformations)

In this section we consider linear maps between Hermitian spaces that preserve the Hermitian norm. All definitions given for Euclidean spaces in Section 12.5 extend to Hermitian spaces, except that orthogonal transformations are called unitary transformation, but Proposition 12.12 extends only with a modified Condition (2). Indeed, the old proof that (2) implies (3) does not work, and the implication is in fact false! It can be repaired by strengthening Condition (2). For the sake of completeness, we state the Hermitian version of Definition 12.5.

Definition 14.7. Given any two nontrivial Hermitian spaces E and F of the same finite dimension n, a function $f\colon E \to F$ is *a unitary transformation, or a linear isometry*, if it is linear and

$$\|f(u)\| = \|u\|, \quad \text{for all } u \in E.$$

Proposition 12.12 can be salvaged by strengthening Condition (2).

Proposition 14.14. *Given any two nontrivial Hermitian spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, the following properties are equivalent:*

(1) f is a linear map and $\|f(u)\| = \|u\|$, for all $u \in E$;

(2) $\|f(v) - f(u)\| = \|v - u\|$ and $f(iu) = if(u)$, for all $u, v \in E$.

(3) $f(u) \cdot f(v) = u \cdot v$, for all $u, v \in E$.

Furthermore, such a map is bijective.

Proof. The proof that (2) implies (3) given in Proposition 12.12 needs to be revised as follows. We use the polarization identity

$$2\varphi(u, v) = (1 + i)(\|u\|^2 + \|v\|^2) - \|u - v\|^2 - i\|u - iv\|^2.$$

Since $f(iv) = if(v)$, we get $f(0) = 0$ by setting $v = 0$, so the function f preserves distance and norm, and we get

$$\begin{aligned}
2\varphi(f(u), f(v)) &= (1 + i)(\|f(u)\|^2 + \|f(v)\|^2) - \|f(u) - f(v)\|^2 \\
&\quad - i\|f(u) - if(v)\|^2 \\
&= (1 + i)(\|f(u)\|^2 + \|f(v)\|^2) - \|f(u) - f(v)\|^2 \\
&\quad - i\|f(u) - f(iv)\|^2 \\
&= (1 + i)(\|u\|^2 + \|v\|^2) - \|u - v\|^2 - i\|u - iv\|^2 \\
&= 2\varphi(u, v),
\end{aligned}$$

which shows that f preserves the Hermitian inner product as desired. The rest of the proof is unchanged. $\square$

Remarks:

(i) In the Euclidean case, we proved that the assumption

$$\|f(v) - f(u)\| = \|v - u\| \quad \text{for all } u, v \in E \text{ and } f(0) = 0 \tag{2'}$$

implies (3). For this we used the polarization identity

$$2u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2.$$

In the Hermitian case the polarization identity involves the complex number i. In fact, the implication (2′) implies (3) is false in the Hermitian case! Conjugation $z \mapsto \overline{z}$ satisfies (2′) since

$$|\overline{z_2} - \overline{z_1}| = |\overline{z_2 - z_1}| = |z_2 - z_1|,$$

and yet, it is not linear!

(ii) If we modify (2) by changing the second condition by now requiring that there be some $\tau \in E$ such that

$$f(\tau + iu) = f(\tau) + i(f(\tau + u) - f(\tau))$$

for all $u \in E$, then the function $g \colon E \to E$ defined such that

$$g(u) = f(\tau + u) - f(\tau)$$

satisfies the old conditions of (2), and the implications (2) $\to$ (3) and (3) $\to$ (1) prove that g is linear, and thus that f is affine. In view of the first remark, some condition involving i is needed on f, in addition to the fact that f is distance-preserving.

14.4 The Unitary Group, Unitary Matrices

In this section, as a mirror image of our treatment of the isometries of a Euclidean space, we explore some of the fundamental properties of the unitary group and of unitary matrices. As an immediate corollary of the Gram–Schmidt orthonormalization procedure, we obtain the QR-decomposition for invertible matrices. In the Hermitian framework, the matrix of the adjoint of a linear map is not given by the transpose of the original matrix, but by the conjugate of the original matrix. For the reader's convenience we recall the following definitions from Section 9.2.

Definition 14.8. Given a complex $m \times n$ matrix A, the *transpose* $A^\top$ *of* A is the $n \times m$ matrix $A^\top = \left(a^\top_{i\,j}\right)$ defined such that

$$a^\top_{i\,j} = a_{j\,i},$$

and the *conjugate* $\overline{A}$ *of* A is the $m \times n$ matrix $\overline{A} = (b_{i\,j})$ defined such that

$$b_{i\,j} = \overline{a_{i\,j}}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. The *adjoint A^* of A* is the matrix defined such that

$$A^* = \overline{(A^\top)} = \left(\overline{A}\right)^\top.$$

Proposition 14.15. *Let E be any Hermitian space of finite dimension n, and let $f\colon E \to E$ be any linear map. The following properties hold:*

(1) The linear map $f\colon E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, if the matrix of f is A, then the matrix of f^* is the adjoint A^* of A, and f is an isometry iff A satisfies the identities

$$A\,A^* = A^*A = I_n,$$

where I_n denotes the identity matrix of order n, iff the columns of A form an orthonormal basis of $\mathbb{C}^n$, iff the rows of A form an orthonormal basis of $\mathbb{C}^n$.

Proof. (1) The proof is identical to that of Proposition 12.14 (1).

(2) If $(e_1, \ldots, e_n)$ is an orthonormal basis for E, let $A = (a_{i\,j})$ be the matrix of f, and let $B = (b_{i\,j})$ be the matrix of f^*. Since f^* is characterized by

$$f^*(u) \cdot v = u \cdot f(v)$$

for all $u, v \in E$, using the fact that if $w = w_1 e_1 + \cdots + w_n e_n$, we have $w_k = w \cdot e_k$, for all k, $1 \leq k \leq n$; letting $u = e_i$ and $v = e_j$, we get

$$b_{j\,i} = f^*(e_i) \cdot e_j = e_i \cdot f(e_j) = \overline{f(e_j) \cdot e_i} = \overline{a_{i\,j}},$$

for all i, j, $1 \leq i, j \leq n$. Thus, $B = A^*$. Now if X and Y are arbitrary matrices over the basis $(e_1, \ldots, e_n)$, denoting as usual the jth column of X by X^j, and similarly for Y, a simple calculation shows that

$$Y^*X = (X^j \cdot Y^i)_{1 \leq i,j \leq n}.$$

Then it is immediately verified that if $X = Y = A$, then $A^*A = A\,A^* = I_n$ iff the column vectors $(A^1, \ldots, A^n)$ form an orthonormal basis. Thus, from (1), we see that (2) is clear. ☐

Proposition 12.14 shows that the inverse of an isometry f is its adjoint f^*. Proposition 12.14 also motivates the following definition.

Definition 14.9. A complex $n \times n$ matrix is a *unitary matrix* if

$$A\,A^* = A^*A = I_n.$$

Remarks:

(1) The conditions $A\,A^* = I_n$, $A^*A = I_n$, and $A^{-1} = A^*$ are equivalent. Given any two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, if P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$, it is easy to show that the matrix P is unitary. The proof of Proposition 14.14 (3) also shows that if f is an isometry, then the image of an orthonormal basis $(u_1, \ldots, u_n)$ is an orthonormal basis.

(2) Using the explicit formula for the determinant, we see immediately that

$$\det(\overline{A}) = \overline{\det(A)}.$$

If f is a unitary transformation and A is its matrix with respect to any orthonormal basis, from $AA^* = I$, we get

$$\det(AA^*) = \det(A)\det(A^*) = \det(A)\overline{\det(A^\top)} = \det(A)\overline{\det(A)} = |\det(A)|^2,$$

and so $|\det(A)| = 1$. It is clear that the isometries of a Hermitian space of dimension n form a group, and that the isometries of determinant $+1$ form a subgroup.

This leads to the following definition.

Definition 14.10. Given a Hermitian space E of dimension n, the set of isometries $f\colon E \to E$ forms a subgroup of $\mathbf{GL}(E, \mathbb{C})$ denoted by $\mathbf{U}(E)$, or $\mathbf{U}(n)$ when $E = \mathbb{C}^n$, called the *unitary group (of E)*. For every isometry f we have $|\det(f)| = 1$, where $\det(f)$ denotes the determinant of f. The isometries such that $\det(f) = 1$ are called *rotations, or proper isometries, or proper unitary transformations*, and they form a subgroup of the special linear group $\mathbf{SL}(E, \mathbb{C})$ (and of $\mathbf{U}(E)$), denoted by $\mathbf{SU}(E)$, or $\mathbf{SU}(n)$ when $E = \mathbb{C}^n$, called the *special unitary group (of E)*. The isometries such that $\det(f) \neq 1$ are called *improper isometries, or improper unitary transformations, or flip transformations*.

A very important example of unitary matrices is provided by Fourier matrices (up to a factor of $\sqrt{n}$), matrices that arise in the various versions of the discrete Fourier transform. For more on this topic, see the problems, and Strang [169, 172].

The group $\mathbf{SU}(2)$ turns out to be the group of *unit quaternions*, invented by Hamilton. This group plays an important role in the representation of rotations in $\mathbf{SO}(3)$ used in computer graphics and robotics; see Chapter 16.

Now that we have the definition of a unitary matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decomposition for matrices.

Definition 14.11. Given any complex $n \times n$ matrix A, a *QR-decomposition* of A is any pair of $n \times n$ matrices (U, R), where U is a unitary matrix and R is an upper triangular matrix such that $A = UR$.

Proposition 14.16. *Given any $n \times n$ complex matrix A, if A is invertible, then there is a unitary matrix U and an upper triangular matrix R with positive diagonal entries such that $A = UR$.*

The proof is absolutely the same as in the real case!

Remark: If A is invertible and if $A = U_1R_1 = U_2R_2$ are two QR-decompositions for A, then

$$R_1R_2^{-1} = U_1^*U_2.$$

Then it is easy to show that there is a diagonal matrix D with diagonal entries such that $|d_{ii}| = 1$ for $i = 1, \ldots, n$, and $U_2 = U_1D$, $R_2 = D^*R_1$.

We have the following version of the Hadamard inequality for complex matrices. The proof is essentially the same as in the Euclidean case but it uses Proposition 14.16 instead of Proposition 12.16.

Proposition 14.17. *(Hadamard) For any complex $n \times n$ matrix $A = (a_{ij})$, we have*

$$|\det(A)| \leq \prod_{i=1}^{n}\biggl(\sum_{j=1}^{n}|a_{ij}|^2\biggr)^{1/2} \quad \textit{and} \quad |\det(A)| \leq \prod_{j=1}^{n}\biggl(\sum_{i=1}^{n}|a_{ij}|^2\biggr)^{1/2}.$$

Moreover, equality holds iff either A has orthogonal rows in the left inequality or orthogonal columns in the right inequality.

We also have the following version of Proposition 12.18 for Hermitian matrices. The proof of Proposition 12.18 goes through because the Cholesky decomposition for a Hermitian positive definite A matrix holds in the form $A = B^*B$, where B is upper triangular with positive diagonal entries. The details are left to the reader.

Proposition 14.18. *(Hadamard) For any complex $n \times n$ matrix $A = (a_{ij})$, if A is Hermitian positive semidefinite, then we have*

$$\det(A) \leq \prod_{i=1}^{n} a_{ii}.$$

Moreover, if A is positive definite, then equality holds iff A is a diagonal matrix.

14.5 Hermitian Reflections and QR-Decomposition

If A is an $n \times n$ complex singular matrix, there is some (not necessarily unique) QR-decomposition $A = QR$ with Q a unitary matrix which is a product of Householder reflections and R an upper triangular matrix, but the proof is more involved. One way to proceed is to generalize the notion of hyperplane reflection. This is not really surprising since in the Hermitian case there are improper isometries whose determinant can be any unit complex number. Hyperplane reflections are generalized as follows.

Definition 14.12. Let E be a Hermitian space of finite dimension. For any hyperplane H, for any nonnull vector w orthogonal to H, so that $E = H \oplus G$, where $G = \mathbb{C}w$, a *Hermitian reflection about H of angle θ* is a linear map of the form $\rho_{H,\theta}\colon E \to E$, defined such that

$$\rho_{H,\theta}(u) = p_H(u) + e^{i\theta} p_G(u),$$

for any unit complex number $e^{i\theta} \neq 1$ (i.e. $\theta \neq k2\pi$). For any nonzero vector $w \in E$, we denote by $\rho_{w,\theta}$ the Hermitian reflection given by $\rho_{H,\theta}$, where H is the hyperplane orthogonal to w.

Since $u = p_H(u) + p_G(u)$, the Hermitian reflection $\rho_{w,\theta}$ is also expressed as

$$\rho_{w,\theta}(u) = u + (e^{i\theta} - 1) p_G(u),$$

or as

$$\rho_{w,\theta}(u) = u + (e^{i\theta} - 1) \frac{(u \cdot w)}{\|w\|^2}\, w.$$

Note that the case of a standard hyperplane reflection is obtained when $e^{i\theta} = -1$, i.e., $\theta = \pi$. In this case,

$$\rho_{w,\pi}(u) = u - 2\, \frac{(u \cdot w)}{\|w\|^2}\, w,$$

and the matrix of such a reflection is a Householder matrix, as in Section 13.1, except that w may be a complex vector.

We leave as an easy exercise to check that $\rho_{w,\theta}$ is indeed an isometry, and that the inverse of $\rho_{w,\theta}$ is $\rho_{w,-\theta}$. If we pick an orthonormal basis $(e_1, \ldots, e_n)$ such that $(e_1, \ldots, e_{n-1})$ is an orthonormal basis of H, the matrix of $\rho_{w,\theta}$ is

$$\begin{pmatrix} I_{n-1} & 0 \\ 0 & e^{i\theta} \end{pmatrix}$$

We now come to the main surprise. Given any two distinct vectors u and v such that $\|u\| = \|v\|$, there isn't always a hyperplane reflection mapping u to v, but this can be done using two Hermitian reflections!

Proposition 14.19. *Let E be any nontrivial Hermitian space.*

(1) *For any two vectors $u, v \in E$ such that $u \neq v$ and $\|u\| = \|v\|$, if $u \cdot v = e^{i\theta}|u \cdot v|$, then the (usual) reflection s about the hyperplane orthogonal to the vector $v - e^{-i\theta}u$ is such that $s(u) = e^{i\theta}v$.*

(2) *For any nonnull vector $v \in E$, for any unit complex number $e^{i\theta} \neq 1$, there is a Hermitian reflection $\rho_{v,\theta}$ such that*

$$\rho_{v,\theta}(v) = e^{i\theta} v.$$

As a consequence, for u and v as in (1), we have $\rho_{v,-\theta} \circ s(u) = v$.

Proof. (1) Consider the (usual) reflection about the hyperplane orthogonal to $w = v - e^{-i\theta}u$. We have

$$s(u) = u - 2\,\frac{(u\cdot(v - e^{-i\theta}u))}{\|v - e^{-i\theta}u\|^2}\,(v - e^{-i\theta}u).$$

We need to compute

$$-2u\cdot(v - e^{-i\theta}u) \quad\text{and}\quad (v - e^{-i\theta}u)\cdot(v - e^{-i\theta}u).$$

Since $u\cdot v = e^{i\theta}|u\cdot v|$, we have

$$e^{-i\theta}u\cdot v = |u\cdot v| \quad\text{and}\quad e^{i\theta}v\cdot u = |u\cdot v|.$$

Using the above and the fact that $\|u\| = \|v\|$, we get

$$\begin{aligned} -2u\cdot(v - e^{-i\theta}u) &= 2e^{i\theta}\,\|u\|^2 - 2u\cdot v,\\ &= 2e^{i\theta}(\|u\|^2 - |u\cdot v|), \end{aligned}$$

and

$$\begin{aligned} (v - e^{-i\theta}u)\cdot(v - e^{-i\theta}u) &= \|v\|^2 + \|u\|^2 - e^{-i\theta}u\cdot v - e^{i\theta}v\cdot u,\\ &= 2(\|u\|^2 - |u\cdot v|), \end{aligned}$$

and thus,

$$-2\,\frac{(u\cdot(v - e^{-i\theta}u))}{\|(v - e^{-i\theta}u)\|^2}\,(v - e^{-i\theta}u) = e^{i\theta}(v - e^{-i\theta}u).$$

But then,

$$s(u) = u + e^{i\theta}(v - e^{-i\theta}u) = u + e^{i\theta}v - u = e^{i\theta}v,$$

and $s(u) = e^{i\theta}v$, as claimed.

(2) This part is easier. Consider the Hermitian reflection

$$\rho_{v,\theta}(u) = u + (e^{i\theta} - 1)\,\frac{(u\cdot v)}{\|v\|^2}\,v.$$

We have

$$\begin{aligned} \rho_{v,\theta}(v) &= v + (e^{i\theta} - 1)\,\frac{(v\cdot v)}{\|v\|^2}\,v,\\ &= v + (e^{i\theta} - 1)v,\\ &= e^{i\theta}v. \end{aligned}$$

Thus, $\rho_{v,\theta}(v) = e^{i\theta}v$. Since $\rho_{v,\theta}$ is linear, changing the argument v to $e^{i\theta}v$, we get

$$\rho_{v,-\theta}(e^{i\theta}v) = v,$$

and thus, $\rho_{v,-\theta}\circ s(u) = v$. ∎

Remarks:

(1) If we use the vector $v + e^{-i\theta}u$ instead of $v - e^{-i\theta}u$, we get $s(u) = -e^{i\theta}v$.

(2) Certain authors, such as Kincaid and Cheney [102] and Ciarlet [41], use the vector $u + e^{i\theta}v$ instead of the vector $v + e^{-i\theta}u$. The effect of this choice is that they also get $s(u) = -e^{i\theta}v$.

(3) If $v = \|u\|\, e_1$, where e_1 is a basis vector, $u \cdot e_1 = a_1$, where a_1 is just the coefficient of u over the basis vector e_1. Then, since $u \cdot e_1 = e^{i\theta}|a_1|$, the choice of the plus sign in the vector $\|u\|\, e_1 + e^{-i\theta}u$ has the effect that the coefficient of this vector over e_1 is $\|u\| + |a_1|$, and no cancellations takes place, which is preferable for numerical stability (we need to divide by the square norm of this vector).

We now show that the QR-decomposition in terms of (complex) Householder matrices holds for complex matrices. We need the version of Proposition 14.19 and a trick at the end of the argument, but the proof is basically unchanged.

Proposition 14.20. *Let E be a nontrivial Hermitian space of dimension n. Given any orthonormal basis $(e_1, \ldots, e_n)$, for any n-tuple of vectors $(v_1, \ldots, v_n)$, there is a sequence of $n - 1$ isometries $h_1, \ldots, h_{n-1}$, such that h_i is a (standard) hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by*

$$r_j = h_{n-1} \circ \cdots \circ h_2 \circ h_1(v_j), \quad 1 \leq j \leq n,$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $(1 \leq j \leq n)$. Equivalently, the matrix R whose columns are the components of the r_j over the basis $(e_1, \ldots, e_n)$ is an upper triangular matrix. Furthermore, if we allow one more isometry h_n of the form

$$h_n = \rho_{e_n, \varphi_n} \circ \cdots \circ \rho_{e_1, \varphi_1}$$

after $h_1, \ldots, h_{n-1}$, we can ensure that the diagonal entries of R are nonnegative.

Proof. The proof is very similar to the proof of Proposition 13.3, but it needs to be modified a little bit since Proposition 14.19 is weaker than Proposition 13.2. We explain how to modify the induction step, leaving the base case and the rest of the proof as an exercise.

As in the proof of Proposition 13.3, the vectors $(e_1, \ldots, e_k)$ form a basis for the subspace denoted as U'_k, the vectors $(e_{k+1}, \ldots, e_n)$ form a basis for the subspace denoted as U''_k, the subspaces U'_k and U''_k are orthogonal, and $E = U'_k \oplus U''_k$. Let

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}).$$

We can write

$$u_{k+1} = u'_{k+1} + u''_{k+1},$$

where $u'_{k+1} \in U'_k$ and $u''_{k+1} \in U''_k$. Let

$$r_{k+1,k+1} = \left\|u''_{k+1}\right\|, \quad \text{and} \quad e^{i\theta_{k+1}}|u''_{k+1} \cdot e_{k+1}| = u''_{k+1} \cdot e_{k+1}.$$

If $u''_{k+1} = e^{i\theta_{k+1}} r_{k+1,k+1}\, e_{k+1}$, we let $h_{k+1} = \mathrm{id}$. Otherwise, by Proposition 14.19(1) (with $u = u''_{k+1}$ and $v = r_{k+1,k+1}\, e_{k+1}$), there is a unique hyperplane reflection h_{k+1} such that

$$h_{k+1}(u''_{k+1}) = e^{i\theta_{k+1}} r_{k+1,k+1}\, e_{k+1},$$

where h_{k+1} is the reflection about the hyperplane H_{k+1} orthogonal to the vector

$$w_{k+1} = r_{k+1,k+1}\, e_{k+1} - e^{-i\theta_{k+1}} u''_{k+1}.$$

At the end of the induction, we have a triangular matrix R, but the diagonal entries $e^{i\theta_j} r_{j,j}$ of R may be complex. Letting

$$h_n = \rho_{e_n, -\theta_n} \circ \cdots \circ \rho_{e_1, -\theta_1},$$

we observe that the diagonal entries of the matrix of vectors

$$r'_j = h_n \circ h_{n-1} \circ \cdots \circ h_2 \circ h_1(v_j)$$

is triangular with nonnegative entries. $\square$

Remark: For numerical stability, it is preferable to use $w_{k+1} = r_{k+1,k+1}\, e_{k+1} + e^{-i\theta_{k+1}} u''_{k+1}$ instead of $w_{k+1} = r_{k+1,k+1}\, e_{k+1} - e^{-i\theta_{k+1}} u''_{k+1}$. The effect of that choice is that the diagonal entries in R will be of the form $-e^{i\theta_j} r_{j,j} = e^{i(\theta_j + \pi)} r_{j,j}$. Of course, we can make these entries nonegative by applying

$$h_n = \rho_{e_n, \pi - \theta_n} \circ \cdots \circ \rho_{e_1, \pi - \theta_1}$$

after h_{n-1}.

As in the Euclidean case, Proposition 14.20 immediately implies the QR-decomposition for arbitrary complex $n \times n$-matrices, where Q is now unitary (see Kincaid and Cheney [102] and Ciarlet [41]).

Proposition 14.21. *For every complex $n \times n$-matrix A, there is a sequence $H_1, \ldots, H_{n-1}$ of matrices, where each H_i is either a Householder matrix or the identity, and an upper triangular matrix R, such that*

$$R = H_{n-1} \cdots H_2 H_1 A.$$

As a corollary, there is a pair of matrices Q, R, where Q is unitary and R is upper triangular, such that $A = QR$ (a QR-decomposition of A). Furthermore, R can be chosen so that its diagonal entries are nonnegative. This can be achieved by a diagonal matrix D with entries such that $|d_{ii}| = 1$ for $i = 1, \ldots, n$, and we have $A = \widetilde{Q}\widetilde{R}$ with

$$\widetilde{Q} = H_1 \cdots H_{n-1} D, \quad \widetilde{R} = D^* R,$$

where $\widetilde{R}$ is upper triangular and has nonnegative diagonal entries.

Proof. It is essentially identical to the proof of Proposition 13.4, and we leave the details as an exercise. For the last statement, observe that $h_n \circ \cdots \circ h_1$ is also an isometry. $\square$

14.6 Orthogonal Projections and Involutions

In this section we begin by assuming that the field K is not a field of characteristic 2. Recall that a linear map $f\colon E \to E$ is an *involution* iff $f^2 = \mathrm{id}$, and is *idempotent* iff $f^2 = f$. We know from Proposition 6.9 that if f is idempotent, then

$$E = \mathrm{Im}(f) \oplus \mathrm{Ker}\,(f),$$

and that the restriction of f to its image is the identity. For this reason, a linear idempotent map is called a *projection*. The connection between involutions and projections is given by the following simple proposition.

Proposition 14.22. *For any linear map $f\colon E \to E$, we have $f^2 = \mathrm{id}$ iff $\frac{1}{2}(\mathrm{id} - f)$ is a projection iff $\frac{1}{2}(\mathrm{id} + f)$ is a projection; in this case, f is equal to the difference of the two projections $\frac{1}{2}(\mathrm{id} + f)$ and $\frac{1}{2}(\mathrm{id} - f)$.*

Proof. We have

$$\left(\frac{1}{2}(\mathrm{id} - f)\right)^2 = \frac{1}{4}(\mathrm{id} - 2f + f^2)$$

so

$$\left(\frac{1}{2}(\mathrm{id} - f)\right)^2 = \frac{1}{2}(\mathrm{id} - f) \quad \text{iff} \quad f^2 = \mathrm{id}.$$

We also have

$$\left(\frac{1}{2}(\mathrm{id} + f)\right)^2 = \frac{1}{4}(\mathrm{id} + 2f + f^2),$$

so

$$\left(\frac{1}{2}(\mathrm{id} + f)\right)^2 = \frac{1}{2}(\mathrm{id} + f) \quad \text{iff} \quad f^2 = \mathrm{id}.$$

Obviously, $f = \frac{1}{2}(\mathrm{id} + f) - \frac{1}{2}(\mathrm{id} - f)$. ∎

Proposition 14.23. *For any linear map $f\colon E \to E$, let $U^+ = \mathrm{Ker}\,(\frac{1}{2}(\mathrm{id} - f))$ and let $U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$. If $f^2 = \mathrm{id}$, then*

$$U^+ = \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right) = \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right),$$

and so, $f(u) = u$ on U^+ and $f(u) = -u$ on U^-.

Proof. If $f^2 = \mathrm{id}$, then

$$(\mathrm{id} - f) \circ (\mathrm{id} + f) = \mathrm{id} - f^2 = \mathrm{id} - \mathrm{id} = 0,$$

which implies that

$$\mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right) \subseteq \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right).$$

Conversely, if $u \in \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right)$, then $f(u) = u$, so

$$\frac{1}{2}(\mathrm{id} + f)(u) = \frac{1}{2}(u + u) = u,$$

and thus

$$\mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right) \subseteq \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right).$$

Therefore,

$$U^+ = \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - f)\right) = \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} + f)\right),$$

and so, $f(u) = u$ on U^+ and $f(u) = -u$ on U^-. ☐

We now assume that $K = \mathbb{C}$. The involutions of E that are unitary transformations are characterized as follows.

Proposition 14.24. *Let $f \in \mathbf{GL}(E)$ be an involution. The following properties are equivalent:*

(a) The map f is unitary; that is, $f \in \mathbf{U}(E)$.

(b) The subspaces $U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$ and $U^+ = \mathrm{Im}(\frac{1}{2}(\mathrm{id} + f))$ are orthogonal.

Furthermore, if E is finite-dimensional, then (a) and (b) are equivalent to (c) below:

(c) The map is self-adjoint; that is, $f = f^*$.

Proof. If f is unitary, then from $\langle f(u), f(v)\rangle = \langle u, v\rangle$ for all $u, v \in E$, we see that if $u \in U^+$ and $v \in U^-$, we get

$$\langle u, v\rangle = \langle f(u), f(v)\rangle = \langle u, -v\rangle = -\langle u, v\rangle,$$

so $2\langle u, v\rangle = 0$, which implies $\langle u, v\rangle = 0$, that is, U^+ and U^- are orthogonal. Thus, (a) implies (b).

Conversely, if (b) holds, since $f(u) = u$ on U^+ and $f(u) = -u$ on U^-, we see that $\langle f(u), f(v)\rangle = \langle u, v\rangle$ if $u, v \in U^+$ or if $u, v \in U^-$. Since $E = U^+ \oplus U^-$ and since U^+ and U^- are orthogonal, we also have $\langle f(u), f(v)\rangle = \langle u, v\rangle$ for all $u, v \in E$, and (b) implies (a).

If E is finite-dimensional, the adjoint f^* of f exists, and we know that $f^{-1} = f^*$. Since f is an involution, $f^2 = \mathrm{id}$, which implies that $f^* = f^{-1} = f$. ☐

A unitary involution is the identity on $U^+ = \mathrm{Im}(\frac{1}{2}(\mathrm{id} + f))$, and $f(v) = -v$ for all $v \in U^- = \mathrm{Im}(\frac{1}{2}(\mathrm{id} - f))$. Furthermore, E is an orthogonal direct sum $E = U^+ \oplus U^-$. We say that f is an *orthogonal reflection* about U^+. In the special case where U^+ is a hyperplane, we say that f is a *hyperplane reflection*. We already studied hyperplane reflections in the Euclidean case; see Chapter 13.

If $f\colon E \to E$ is a projection ($f^2 = f$), then

$$(\mathrm{id} - 2f)^2 = \mathrm{id} - 4f + 4f^2 = \mathrm{id} - 4f + 4f = \mathrm{id},$$

so $\mathrm{id} - 2f$ is an involution. As a consequence, we get the following result.

Proposition 14.25. *If $f\colon E \to E$ is a projection ($f^2 = f$), then $\mathrm{Ker}\,(f)$ and $\mathrm{Im}(f)$ are orthogonal iff $f^* = f$.*

Proof. Apply Proposition 14.24 to $g = \mathrm{id} - 2f$. Since $\mathrm{id} - g = 2f$ we have

$$U^+ = \mathrm{Ker}\left(\frac{1}{2}(\mathrm{id} - g)\right) = \mathrm{Ker}\,(f)$$

and

$$U^- = \mathrm{Im}\left(\frac{1}{2}(\mathrm{id} - g)\right) = \mathrm{Im}(f),$$

which proves the proposition. $\square$

A projection such that $f = f^*$ is called an *orthogonal projection*.

If $(a_1 \ldots, a_k)$ are k linearly independent vectors in $\mathbb{R}^n$, let us determine the matrix P of the orthogonal projection onto the subspace of $\mathbb{R}^n$ spanned by $(a_1, \ldots, a_k)$. Let A be the $n \times k$ matrix whose jth column consists of the coordinates of the vector a_j over the canonical basis $(e_1, \ldots, e_n)$.

Any vector in the subspace $(a_1, \ldots, a_k)$ is a linear combination of the form Ax, for some $x \in \mathbb{R}^k$. Given any $y \in \mathbb{R}^n$, the orthogonal projection $Py = Ax$ of y onto the subspace spanned by $(a_1, \ldots, a_k)$ is the vector Ax such that $y - Ax$ is orthogonal to the subspace spanned by $(a_1, \ldots, a_k)$ (prove it). This means that $y - Ax$ is orthogonal to every a_j, which is expressed by

$$A^\top (y - Ax) = 0;$$

that is,

$$A^\top A x = A^\top y.$$

The matrix $A^\top A$ is invertible because A has full rank k, thus we get

$$x = (A^\top A)^{-1} A^\top y,$$

and so

$$Py = Ax = A(A^\top A)^{-1} A^\top y.$$

Therefore, the matrix P of the projection onto the subspace spanned by $(a_1 \ldots, a_k)$ is given by

$$P = A(A^\top A)^{-1} A^\top.$$

The reader should check that $P^2 = P$ and $P^\top = P$.

14.7 Dual Norms

In the remark following the proof of Proposition 9.10, we explained that if $(E, \| \|)$ and $(F, \| \|)$ are two normed vector spaces and if we let $\mathcal{L}(E; F)$ denote the set of all continuous (equivalently, bounded) linear maps from E to F, then, we can define the *operator norm* (or *subordinate norm*) $\| \|$ on $\mathcal{L}(E; F)$ as follows: for every $f \in \mathcal{L}(E; F)$,

$$\|f\| = \sup_{\substack{x \in E \\ x \neq 0}} \frac{\|f(x)\|}{\|x\|} = \sup_{\substack{x \in E \\ \|x\|=1}} \|f(x)\| .$$

In particular, if $F = \mathbb{C}$, then $\mathcal{L}(E; F) = E'$ is the *dual space* of E, and we get the operator norm denoted by $\| \|_*$ given by

$$\|f\|_* = \sup_{\substack{x \in E \\ \|x\|=1}} |f(x)|.$$

The norm $\| \|_*$ is called the *dual norm* of $\| \|$ on E'.

Let us now assume that E is a finite-dimensional Hermitian space, in which case $E' = E^*$. Theorem 14.6 implies that for every linear form $f \in E^*$, there is a unique vector $y \in E$ so that

$$f(x) = \langle x, y \rangle,$$

for all $x \in E$, and so we can write

$$\|f\|_* = \sup_{\substack{x \in E \\ \|x\|=1}} |\langle x, y \rangle|.$$

The above suggests defining a norm $\| \|^D$ on E.

Definition 14.13. If E is a finite-dimensional Hermitian space and $\| \|$ is any norm on E, for any $y \in E$ we let

$$\|y\|^D = \sup_{\substack{x \in E \\ \|x\|=1}} |\langle x, y \rangle|,$$

be the *dual norm* of $\| \|$ (on E). If E is a real Euclidean space, then the dual norm is defined by

$$\|y\|^D = \sup_{\substack{x \in E \\ \|x\|=1}} \langle x, y \rangle$$

for all $y \in E$.

Beware that $\| \|$ is generally *not* the Hermitian norm associated with the Hermitian inner product. The dual norm shows up in convex programming; see Boyd and Vandenberghe [29], Chapters 2, 3, 6, 9.

The fact that $\| \|^D$ is a norm follows from the fact that $\| \|_*$ is a norm and can also be checked directly. It is worth noting that the triangle inequality for $\| \|^D$ comes "for free," in the sense that it holds for any function $p\colon E \to \mathbb{R}$.

Proposition 14.26. *For any function $p\colon E \to \mathbb{R}$, if we define p^D by*

$$p^D(x) = \sup_{p(z)=1} |\langle z, x\rangle|,$$

then we have

$$p^D(x+y) \leq p^D(x) + p^D(y).$$

Proof. We have

$$\begin{aligned} p^D(x+y) &= \sup_{p(z)=1} |\langle z, x+y\rangle| \\ &= \sup_{p(z)=1} (|\langle z, x\rangle + \langle z, y\rangle|) \\ &\leq \sup_{p(z)=1} (|\langle z, x\rangle| + |\langle z, y\rangle|) \\ &\leq \sup_{p(z)=1} |\langle z, x\rangle| + \sup_{p(z)=1} |\langle z, y\rangle| \\ &= p^D(x) + p^D(y). \end{aligned}$$

∎

Definition 14.14. If $p\colon E \to \mathbb{R}$ is a function such that

(1) $p(x) \geq 0$ for all $x \in E$, and $p(x) = 0$ iff $x = 0$;

(2) $p(\lambda x) = |\lambda| p(x)$, for all $x \in E$ and all $\lambda \in \mathbb{C}$;

(3) p is continuous, in the sense that for some basis $(e_1, \ldots, e_n)$ of E, the function

$$(x_1, \ldots, x_n) \mapsto p(x_1 e_1 + \cdots + x_n e_n)$$

from $\mathbb{C}^n$ to $\mathbb{R}$ is continuous,

then we say that p is a *pre-norm*.

Obviously, every norm is a pre-norm, but a pre-norm may not satisfy the triangle inequality.

Corollary 14.27. *The dual norm of any pre-norm is actually a norm.*

Proposition 14.28. *For all $y \in E$, we have*

$$\|y\|^D = \sup_{\substack{x \in E \\ \|x\|=1}} |\langle x, y\rangle| = \sup_{\substack{x \in E \\ \|x\|=1}} \Re\langle x, y\rangle.$$

Proof. Since E is finite dimensional, the unit sphere $S^{n-1} = \{x \in E \mid \|x\| = 1\}$ is compact, so there is some $x_0 \in S^{n-1}$ such that

$$\|y\|^D = |\langle x_0, y\rangle|.$$

If $\langle x_0, y\rangle = \rho e^{i\theta}$, with $\rho \geq 0$, then

$$|\langle e^{-i\theta} x_0, y\rangle| = |e^{-i\theta}\langle x_0, y\rangle| = |e^{-i\theta}\rho e^{i\theta}| = \rho,$$

so

$$\|y\|^D = \rho = \langle e^{-i\theta} x_0, y\rangle, \tag{$*$}$$

with $\left\|e^{-i\theta} x_0\right\| = \|x_0\| = 1$. On the other hand,

$$\Re\langle x, y\rangle \leq |\langle x, y\rangle|,$$

so by $(*)$ we get

$$\|y\|^D = \sup_{\substack{x\in E\\ \|x\|=1}} |\langle x, y\rangle| = \sup_{\substack{x\in E\\ \|x\|=1}} \Re\langle x, y\rangle,$$

as claimed. □

Proposition 14.29. *For all $x, y \in E$, we have*

$$\begin{aligned} |\langle x, y\rangle| &\leq \|x\| \, \|y\|^D \\ |\langle x, y\rangle| &\leq \|x\|^D \, \|y\| . \end{aligned}$$

Proof. If $x = 0$, then $\langle x, y\rangle = 0$ and these inequalities are trivial. If $x \neq 0$, since $\|x/\|x\|\| = 1$, by definition of $\|y\|^D$, we have

$$|\langle x/\|x\|, y\rangle| \leq \sup_{\|z\|=1} |\langle z, y\rangle| = \|y\|^D,$$

which yields

$$|\langle x, y\rangle| \leq \|x\| \, \|y\|^D .$$

The second inequality holds because $|\langle x, y\rangle| = |\langle y, x\rangle|$. □

It is not hard to show that for all $y \in \mathbb{C}^n$,

$$\begin{aligned} \|y\|_1^D &= \|y\|_\infty \\ \|y\|_\infty^D &= \|y\|_1 \\ \|y\|_2^D &= \|y\|_2 . \end{aligned}$$

Thus, the Euclidean norm is autodual. More generally, the following proposition holds.

Proposition 14.30. *If $p, q \geq 1$ and $1/p + 1/q = 1$, or $p = 1$ and $q = \infty$, or $p = \infty$ and $q = 1$, then for all $y \in \mathbb{C}^n$, we have*

$$\|y\|_p^D = \|y\|_q .$$

Proof. By Hölder's inequality (Corollary 9.2), for all $x, y \in \mathbb{C}^n$, we have

$$|\langle x, y\rangle| \leq \|x\|_p \, \|y\|_q \,,$$

so

$$\|y\|_p^D = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_p = 1}} |\langle x, y\rangle| \leq \|y\|_q \,.$$

For the converse, we consider the cases $p = 1$, $1 < p < +\infty$, and $p = +\infty$. First assume $p = 1$. The result is obvious for $y = 0$, so assume $y \neq 0$. Given y, if we pick $x_j = 1$ for some index j such that $\|y\|_\infty = \max_{1 \leq i \leq n} |y_i| = |y_j|$, and $x_k = 0$ for $k \neq j$, then $|\langle x, y\rangle| = |y_j| = \|y\|_\infty$, so $\|y\|_1^D = \|y\|_\infty$.

Now we turn to the case $1 < p < +\infty$. Then we also have $1 < q < +\infty$, and the equation $1/p + 1/q = 1$ is equivalent to $pq = p + q$, that is, $p(q-1) = q$. Pick $z_j = y_j |y_j|^{q-2}$ for $j = 1, \ldots, n$, so that

$$\|z\|_p = \left(\sum_{j=1}^n |z_j|^p\right)^{1/p} = \left(\sum_{j=1}^n |y_j|^{(q-1)p}\right)^{1/p} = \left(\sum_{j=1}^n |y_j|^q\right)^{1/p}.$$

Then if $x = z / \|z\|_p$, we have

$$|\langle x, y\rangle| = \frac{\left|\sum_{j=1}^n z_j \overline{y_j}\right|}{\|z\|_p} = \frac{\left|\sum_{j=1}^n y_j \overline{y_j} |y_j|^{q-2}\right|}{\|z\|_p} = \frac{\sum_{j=1}^n |y_j|^q}{\left(\sum_{j=1}^n |y_j|^q\right)^{1/p}} = \left(\sum_{j=1}^n |y_j|^q\right)^{1/q} = \|y\|_q \,.$$

Thus $\|y\|_p^D = \|y\|_q$.

Finally, if $p = \infty$, then pick $x_j = y_j / |y_j|$ if $y_j \neq 0$, and $x_j = 0$ if $y_j = 0$. Then

$$|\langle x, y\rangle| = \left|\sum_{y_j \neq 0}^n y_j \overline{y_j} / |y_j|\right| = \sum_{y_j \neq 0} |y_j| = \|y\|_1 \,.$$

Thus $\|y\|_\infty^D = \|y\|_1$. $\square$

We can show that the dual of the spectral norm is the *trace norm* (or *nuclear norm*) also discussed in Section 22.5. Recall from Proposition 9.10 that the spectral norm $\|A\|_2$ of a matrix A is the square root of the largest eigenvalue of $A^* A$, that is, the largest singular value of A.

Proposition 14.31. *The dual of the spectral norm is given by*

$$\|A\|_2^D = \sigma_1 + \cdots + \sigma_r,$$

where $\sigma_1 > \cdots > \sigma_r > 0$ are the singular values of $A \in \mathrm{M}_n(\mathbb{C})$ (which has rank r).

Proof. In this case the inner product on $\mathrm{M}_n(\mathbb{C})$ is the Frobenius inner product $\langle A, B\rangle = \mathrm{tr}(B^*A)$, and the dual norm of the spectral norm is given by

$$\|A\|_2^D = \sup\{|\mathrm{tr}(A^*B)| \mid \|B\|_2 = 1\}.$$

If we factor A using an SVD as $A = V\Sigma U^*$, where U and V are unitary and Σ is a diagonal matrix whose r nonzero entries are the singular values $\sigma_1 > \cdots > \sigma_r > 0$, where r is the rank of A, then

$$|\mathrm{tr}(A^*B)| = |\mathrm{tr}(U\Sigma V^*B)| = |\mathrm{tr}(\Sigma V^*BU)|,$$

so if we pick $B = VU^*$, a unitary matrix such that $\|B\|_2 = 1$, we get

$$|\mathrm{tr}(A^*B)| = \mathrm{tr}(\Sigma) = \sigma_1 + \cdots + \sigma_r,$$

and thus

$$\|A\|_2^D \geq \sigma_1 + \cdots + \sigma_r.$$

Since $\|B\|_2 = 1$ and U and V are unitary, by Proposition 9.10 we have $\|V^*BU\|_2 = \|B\|_2 = 1$. If $Z = V^*BU$, by definition of the operator norm

$$1 = \|Z\|_2 = \sup\{\|Zx\|_2 \mid \|x\|_2 = 1\},$$

so by picking x to be the canonical vector e_j, we see that $\|Z^j\|_2 \leq 1$ where Z^j is the jth column of Z, so $|z_{jj}| \leq 1$, and since

$$|\mathrm{tr}(\Sigma V^*BU)| = |\mathrm{tr}(\Sigma Z)| = \left|\sum_{j=1}^r \sigma_j z_{jj}\right| \leq \sum_{j=1}^r \sigma_j |z_{jj}| \leq \sum_{j=1}^r \sigma_j,$$

and we conclude that

$$|\mathrm{tr}(\Sigma V^*BU)| \leq \sum_{j=1}^r \sigma_j.$$

The above implies that

$$\|A\|_2^D \leq \sigma_1 + \cdots + \sigma_r,$$

and since we also have $\|A\|_2^D \geq \sigma_1 + \cdots + \sigma_r$, we conclude that

$$\|A\|_2^D = \sigma_1 + \cdots + \sigma_r,$$

proving our proposition. ☐

Definition 14.15. Given any complex matrix $n \times n$ matrix A of rank r, its *nuclear norm* (or *trace norm*) is given by

$$\|A\|_N = \sigma_1 + \cdots + \sigma_r.$$

The nuclear norm can be generalized to $m \times n$ matrices (see Section 22.5). The nuclear norm $\sigma_1 + \cdots + \sigma_r$ of an $m \times n$ matrix A (where r is the rank of A) is denoted by $\|A\|_N$. The nuclear norm plays an important role in *matrix completion*. The problem is this. Given a matrix A_0 with missing entries (missing data), one would like to fill in the missing entries in A_0 to obtain a matrix A of minimal rank. For example, consider the matrices

$$A_0 = \begin{pmatrix} 1 & 2 \\ * & * \end{pmatrix}, \qquad B_0 = \begin{pmatrix} 1 & * \\ * & 4 \end{pmatrix}, \qquad C_0 = \begin{pmatrix} 1 & 2 \\ 3 & * \end{pmatrix}.$$

All can be completed with rank 1. For A_0, use any multiple of $(1, 2)$ for the second row. For B_0, use any numbers b and c such that $bc = 4$. For C_0, the only possibility is $d = 6$.

A famous example of this problem is the *Netflix competition*. The ratings of m films by n viewers goes into A_0. But the customers didn't see all the movies. Many ratings were missing. Those had to be predicted by a recommender system. The nuclear norm gave a good solution that needed to be adjusted for human psychology.

Since the rank of a matrix is not a norm, in order to solve the matrix completion problem we can use the following "convex relaxation." Let A_0 be an incomplete $m \times n$ matrix:

Minimize $\|A\|_N$ subject to $A = A_0$ in the known entries.

The above problem has been extensively studied, in particular by Candès and Recht. Roughly, they showed that if A is an $n \times n$ matrix of rank r and K entries are known in A, then if K is large enough ($K > Cn^{5/4} r \log n$), with high probability, the recovery of A is perfect. See Strang [171] for details (Section III.5).

We close this section by stating the following duality theorem.

Theorem 14.32. *If E is a finite-dimensional Hermitian space, then for any norm $\| \;\|$ on E, we have*

$$\|y\|^{DD} = \|y\|$$

for all $y \in E$.

Proof. By Proposition 14.29, we have

$$|\langle x, y \rangle| \leq \|x\|^D \|y\|,$$

so we get

$$\|y\|^{DD} = \sup_{\|x\|^D = 1} |\langle x, y \rangle| \leq \|y\|, \quad \text{for all } y \in E.$$

It remains to prove that

$$\|y\| \leq \|y\|^{DD}, \quad \text{for all } y \in E.$$

Proofs of this fact can be found in Horn and Johnson [95] (Section 5.5), and in Serre [156] (Chapter 7). The proof makes use of the fact that a nonempty, closed, convex set has a supporting hyperplane through each of its boundary points, a result known as *Minkowski's* *lemma*. For a geometric interpretation of supporting hyperplane see Figure 14.1. This result is a consequence of the *Hahn–Banach theorem*; see Gallier [72]. We give the proof in the case where E is a real Euclidean space. Some minor modifications have to be made when dealing with complex vector spaces and are left as an exercise.

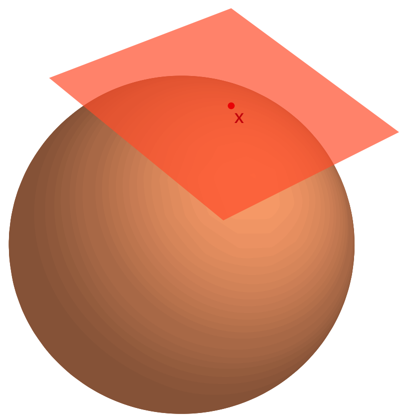


Figure 14.1: The orange tangent plane is a supporting hyperplane to the unit ball in $\mathbb{R}^3$ since this ball is entirely contained in "one side" of the tangent plane.

Since the unit ball $B = \{z \in E \mid \|z\| \leq 1\}$ is closed and convex, the Minkowski lemma says for every x such that $\|x\| = 1$, there is an affine map g of the form

$$g(z) = \langle z, w\rangle - \langle x, w\rangle$$

with $\|w\| = 1$, such that $g(x) = 0$ and $g(z) \leq 0$ for all z such that $\|z\| \leq 1$. Then it is clear that

$$\sup_{\|z\|=1} \langle z, w\rangle = \langle x, w\rangle,$$

and so

$$\|w\|^D = \langle x, w\rangle.$$

It follows that

$$\|x\|^{DD} \geq \langle w/\left\|w\right\|^D, x\rangle = \frac{\langle x, w\rangle}{\|w\|^D} = 1 = \|x\|$$

for all x such that $\|x\| = 1$. By homogeneity, this is true for all $y \in E$, which completes the proof in the real case. When E is a complex vector space, we have to view the unit ball B as a closed convex set in $\mathbb{R}^{2n}$ and we use the fact that there is real affine map of the form

$$g(z) = \Re\langle z, w\rangle - \Re\langle x, w\rangle$$

such that $g(x) = 0$ and $g(z) \leq 0$ for all z with $\|z\| = 1$, so that $\|w\|^D = \Re\langle x, w\rangle$. $\square$

More details on dual norms and unitarily invariant norms can be found in Horn and Johnson [95] (Chapters 5 and 7).

14.8 Summary

The main concepts and results of this chapter are listed below:

- *Semilinear maps*.
- *Sesquilinear forms*; *Hermitian forms*.
- *Quadratic form* associated with a sesquilinear form.
- *Polarization identities*.
- *Positive* and *positive definite* Hermitian forms; *pre-Hilbert spaces*, *Hermitian spaces*.
- *Gram matrix* associated with a Hermitian product.
- The *Cauchy–Schwarz inequality* and the *Minkowski inequality*.
- *Hermitian inner product*, *Hermitian norm*.
- The *parallelogram law*.
- The musical isomorphisms $\flat\colon \overline{E} \to E^*$ and $\sharp\colon E^* \to \overline{E}$; Theorem 14.6 ($E$ is finite-dimensional).
- The *adjoint* of a linear map (with respect to a Hermitian inner product).
- Existence of orthonormal bases in a Hermitian space (Proposition 14.11).
- *Gram–Schmidt orthonormalization procedure*.
- *Linear isometries* (*unitary transformations*).
- The *unitary group*, *unitary matrices*.
- The *unitary group* $\mathbf{U}(n)$.
- The *special unitary group* $\mathbf{SU}(n)$.
- QR-Decomposition for arbitrary complex matrices.
- The *Hadamard inequality* for complex matrices.
- The *Hadamard inequality* for Hermitian positive semidefinite matrices.
- Orthogonal projections and involutions; orthogonal reflections.
- Dual norms.
- Nuclear norm (also called trace norm).
- Matrix completion.

14.9 Problems

Problem 14.1. Let $(E, \langle -, - \rangle)$ be a Hermitian space of finite dimension. Prove that if $f\colon E \to E$ is a self-adjoint linear map (that is, $f^* = f$), then $\langle f(x), x \rangle \in \mathbb{R}$ for all $x \in E$.

Problem 14.2. Prove the polarization identities of Proposition 14.1.

Problem 14.3. Let E be a real Euclidean space. Give an example of a nonzero linear map $f\colon E \to E$ such that $\langle f(u), u \rangle = 0$ for all $u \in E$.

Problem 14.4. Prove Proposition 14.9.

Problem 14.5. (1) Prove that every matrix in $\mathbf{SU}(2)$ is of the form

$$A = \begin{pmatrix} a + ib & c + id \\ -c + id & a - ib \end{pmatrix}, \quad a^2 + b^2 + c^2 + d^2 = 1,\ a, b, c, d \in \mathbb{R},$$

(2) Prove that the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

all belong to $\mathbf{SU}(2)$ and are linearly independent over $\mathbb{C}$.

(3) Prove that the linear span of $\mathbf{SU}(2)$ over $\mathbb{C}$ is the complex vector space $\mathrm{M}_2(\mathbb{C})$ of all complex 2×2 matrices.

Problem 14.6. The purpose of this problem is to prove that the linear span of $\mathbf{SU}(n)$ over $\mathbb{C}$ is $\mathrm{M}_n(\mathbb{C})$ for all $n \geq 3$. One way to prove this result is to adapt the method of Problem 12.12, so please review this problem.

Every complex matrix $A \in \mathrm{M}_n(\mathbb{C})$ can be written as

$$A = \frac{A + A^*}{2} + \frac{A - A^*}{2}$$

where the first matrix is Hermitian and the second matrix is skew-Hermitian. Observe that if $A = (z_{ij})$ is a Hermitian matrix, that is $A^* = A$, then $z_{ji} = \overline{z}_{ij}$, so if $z_{ij} = a_{ij} + ib_{ij}$ with $a_{ij}, b_{ij} \in \mathbb{R}$, then $a_{ij} = a_{ji}$ and $b_{ij} = -b_{ji}$. On the other hand, if $A = (z_{ij})$ is a skew-Hermitian matrix, that is $A^* = -A$, then $z_{ji} = -\overline{z}_{ij}$, so $a_{ij} = -a_{ji}$ and $b_{ij} = b_{ji}$.

The Hermitian and the skew-Hermitian matrices do not form complex vector spaces because they are not closed under multiplication by a complex number, but we can get around this problem by treating the real part and the complex part of these matrices separately and using multiplication by reals.

(1) Consider the matrices of the form

$$R_c^{i,j} = \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 0 & 0 & \cdots & 0 & i & & & \\ & & & 0 & 1 & \cdots & 0 & 0 & & & \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots & & & \\ & & & 0 & 0 & \cdots & 1 & 0 & & & \\ & & & i & 0 & \cdots & 0 & 0 & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix}.$$

Prove that $(R_c^{i,j})^* R_c^{i,j} = I$ and $\det(R_c^{i,j}) = +1$. Use the matrices $R^{i,j}, R_c^{i,j} \in \mathbf{SU}(n)$ and the matrices $(R^{i,j} - (R^{i,j})^*)/2$ (from Problem 12.12) to form the real part of a skew-Hermitian matrix and the matrices $(R_c^{i,j} - (R_c^{i,j})^*)/2$ to form the imaginary part of a skew-Hermitian matrix. Deduce that the matrices in $\mathbf{SU}(n)$ span all skew-Hermitian matrices.

(2) Consider matrices of the form

Type 1

$$S_c^{1,2} = \begin{pmatrix} 0 & -i & 0 & 0 & \dots & 0 \\ i & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Type 2

$$S_c^{i,i+1} = \begin{pmatrix} -1 & & & & & & & & \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & 1 & & & & & \\ & & & & 0 & -i & & & \\ & & & & i & 0 & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}.$$

Type 3

$$S_c^{i,j} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & 0 & \cdots & 0 & -i & & \\ & & & 0 & -1 & \cdots & 0 & 0 & & \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots & & \\ & & & 0 & 0 & \cdots & 1 & 0 & & \\ & & & i & 0 & \cdots & 0 & 0 & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix}.$$

Prove that $S^{i,j}, S_c^{i,j} \in \mathbf{SU}(n)$, and using diagonal matrices as in Problem 12.12, prove that the matrices $S^{i,j}$ can be used to form the real part of a Hermitian matrix and the matrices $S_c^{i,j}$ can be used to form the imaginary part of a Hermitian matrix.

(3) Use (1) and (2) to prove that the matrices in $\mathbf{SU}(n)$ span all Hermitian matrices. It follows that $\mathbf{SU}(n)$ spans $\mathrm{M}_n(\mathbb{C})$ for $n \geq 3$.

Problem 14.7. Consider the complex matrix

$$A = \begin{pmatrix} i & 1 \\ 1 & -i \end{pmatrix}.$$

Check that this matrix is symmetric but not Hermitian. Prove that

$$\det(\lambda I - A) = \lambda^2,$$

and so the eigenvalues of A are $0, 0$.

Problem 14.8. Let $(E, \langle -, - \rangle)$ be a Hermitian space of finite dimension and let $f \colon E \to E$ be a linear map. Prove that the following conditions are equivalent.

(1) $f \circ f^* = f^* \circ f$ (f is normal).

(2) $\langle f(x), f(y) \rangle = \langle f^*(x), f^*(y) \rangle$ for all $x, y \in E$.

(3) $\|f(x)\| = \|f^*(x)\|$ for all $x \in E$.

(4) The map f can be diagonalized with respect to an orthonormal basis of eigenvectors.

(5) There exist some linear maps $g, h \colon E \to E$ such that, $g = g^*$, $\langle x, g(x) \rangle \geq 0$ for all $x \in E$, $h^{-1} = h^*$, and $f = g \circ h = h \circ g$.

(6) There exist some linear map $h \colon E \to E$ such that $h^{-1} = h^*$ and $f^* = h \circ f$.

(7) There is a polynomial P (with complex coefficients) such that $f^* = P(f)$.

Problem 14.9. Recall from Problem 13.7 that a complex $n \times n$ matrix H is *upper Hessenberg* if $h_{jk} = 0$ for all (j, k) such that $j - k \geq 0$. Adapt the proof of Problem 13.7 to prove that given any complex $n \times n$-matrix A, there are $n - 2 \geq 1$ complex matrices $H_1, \ldots, H_{n-2}$, Householder matrices or the identity, such that

$$B = H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2}$$

is upper Hessenberg.

Problem 14.10. Prove that all $y \in \mathbb{C}^n$,

$$\begin{aligned}
\|y\|_1^D &= \|y\|_\infty \\
\|y\|_\infty^D &= \|y\|_1 \\
\|y\|_2^D &= \|y\|_2 .
\end{aligned}$$

Problem 14.11. The purpose of this problem is to complete each of the matrices A_0, B_0, C_0 of Section 14.7 to a matrix A in such way that the nuclear norm $\|A\|_N$ is minimized.

(1) Prove that the squares σ_1^2 and σ_2^2 of the singular values of

$$A = \begin{pmatrix} 1 & 2 \\ c & d \end{pmatrix}$$

are the zeros of the equation

$$\lambda^2 - (5 + c^2 + d^2)\lambda + (2c - d)^2 = 0.$$

(2) Using the fact that

$$\|A\|_N = \sigma_1 + \sigma_2 = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2},$$

prove that

$$\|A\|_N^2 = 5 + c^2 + d^2 + 2|2c - d|.$$

Consider the cases where $2c - d \geq 0$ and $2c - d \leq 0$, and show that in both cases we must have $c = -2d$, and that the minimum of $f(c, d) = 5 + c^2 + d^2 + 2|2c - d|$ is achieved by $c = d = 0$. Conclude that the matrix A completing A_0 that minimizes $\|A\|_N$ is

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}.$$

(3) Prove that the squares σ_1^2 and σ_2^2 of the singular values of

$$A = \begin{pmatrix} 1 & b \\ c & 4 \end{pmatrix}$$

are the zeros of the equation

$$\lambda^2 - (17 + b^2 + c^2)\lambda + (4 - bc)^2 = 0.$$

(4) Prove that

$$\|A\|_N^2 = 17 + b^2 + c^2 + 2|4 - bc|.$$

Consider the cases where $4 - bc \geq 0$ and $4 - bc \leq 0$, and show that in both cases we must have $b^2 = c^2$. Then show that the minimum of $f(c, d) = 17 + b^2 + c^2 + 2|4 - bc|$ is achieved by $b = c$ with $-2 \leq b \leq 2$. Conclude that the matrices A completing B_0 that minimize $\|A\|_N$ are given by

$$A = \begin{pmatrix} 1 & b \\ b & 4 \end{pmatrix}, \qquad -2 \leq b \leq 2.$$

(5) Prove that the squares σ_1^2 and σ_2^2 of the singular values of

$$A = \begin{pmatrix} 1 & 2 \\ 3 & d \end{pmatrix}$$

are the zeros of the equation

$$\lambda^2 - (14 + d^2)\lambda + (6 - d)^2 = 0$$

(6) Prove that

$$\|A\|_N^2 = 14 + d^2 + 2|6 - d|.$$

Consider the cases where $6 - d \geq 0$ and $6 - d \leq 0$, and show that the minimum of $f(c, d) = 14 + d^2 + 2|6 - d|$ is achieved by $d = 1$. Conclude that the the matrix A completing C_0 that minimizes $\|A\|_N$ is given by

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}.$$

Problem 14.12. Prove Theorem 14.32 when E is a finite-dimensional Hermitian space.

Chapter 15

Eigenvectors and Eigenvalues

In this chapter all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$ or $K = \mathbb{C}$.

15.1 Eigenvectors and Eigenvalues of a Linear Map

Given a finite-dimensional vector space E, let $f\colon E \to E$ be any linear map. If by luck there is a basis $(e_1, \ldots, e_n)$ of E with respect to which f is represented by a *diagonal matrix*

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix},$$

then the action of f on E is very simple; in every "direction" e_i, we have

$$f(e_i) = \lambda_i e_i.$$

We can think of f as a transformation that stretches or shrinks space along the direction $e_1, \ldots, e_n$ (at least if E is a real vector space). In terms of matrices, the above property translates into the fact that there is an invertible matrix P and a diagonal matrix D such that a matrix A can be factored as

$$A = PDP^{-1}.$$

When this happens, we say that f (or A) is *diagonalizable*, the λ_i's are called the *eigenvalues* of f, and the e_i's are *eigenvectors* of f. For example, we will see that every symmetric matrix can be diagonalized. Unfortunately, not every matrix can be diagonalized. For example, the matrix

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

can't be diagonalized. Sometimes a matrix fails to be diagonalizable because its eigenvalues do not belong to the field of coefficients, such as

$$A_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

whose eigenvalues are $\pm i$. This is not a serious problem because A_2 can be diagonalized over the complex numbers. However, A_1 is a "fatal" case! Indeed, its eigenvalues are both 1 and the problem is that A_1 does not have enough eigenvectors to span E.

The next best thing is that there is a basis with respect to which f is represented by an *upper triangular* matrix. In this case we say that f can be *triangularized*, or that f is *triangularizable*. As we will see in Section 15.2, if all the eigenvalues of f belong to the field of coefficients K, then f can be triangularized. In particular, this is the case if $K = \mathbb{C}$.

Now an alternative to triangularization is to consider the representation of f with respect to *two* bases $(e_1, \ldots, e_n)$ and $(f_1, \ldots, f_n)$, rather than a single basis. In this case, if $K = \mathbb{R}$ or $K = \mathbb{C}$, it turns out that we can even pick these bases to be *orthonormal*, and we get a diagonal matrix Σ with *nonnegative entries*, such that

$$f(e_i) = \sigma_i f_i, \quad 1 \leq i \leq n.$$

The nonzero σ_i's are the *singular values* of f, and the corresponding representation is the *singular value decomposition*, or *SVD*. The SVD plays a very important role in applications, and will be considered in detail in Chapter 22.

In this section we focus on the possibility of diagonalizing a linear map, and we introduce the relevant concepts to do so. Given a vector space E over a field K, let id denote the identity map on E.

The notion of eigenvalue of a linear map $f\colon E \to E$ defined on an infinite-dimensional space E is quite subtle because it cannot be defined in terms of eigenvectors as in the finite-dimensional case. The problem is that the map $\lambda\,\mathrm{id} - f$ (with $\lambda \in \mathbb{C}$) could be noninvertible (because it is not surjective) and yet injective. In finite dimension this cannot happen, so until further notice we *assume that E is of finite dimension n*.

Definition 15.1. Given any vector space E of finite dimension n and any linear map $f\colon E \to E$, a scalar $\lambda \in K$ is called an *eigenvalue, or proper value, or characteristic value of f* if there is some *nonzero* vector $u \in E$ such that

$$f(u) = \lambda u.$$

Equivalently, λ is an eigenvalue of f if $\mathrm{Ker}\,(\lambda\,\mathrm{id} - f)$ is nontrivial (i.e., $\mathrm{Ker}\,(\lambda\,\mathrm{id} - f) \neq \{0\}$) iff $\lambda\,\mathrm{id} - f$ is *not* invertible (this is where the fact that E is finite-dimensional is used; a linear map from E to itself is injective iff it is invertible). A vector $u \in E$ is called an *eigenvector, or proper vector, or characteristic vector of f* if $u \neq 0$ and if there is some $\lambda \in K$ such that

$$f(u) = \lambda u;$$

the scalar λ is then an eigenvalue, and we say that u is an *eigenvector associated with* λ. Given any eigenvalue $\lambda \in K$, the nontrivial subspace $\mathrm{Ker}\,(\lambda\,\mathrm{id} - f)$ consists of all the eigenvectors associated with λ together with the zero vector; this subspace is denoted by $E_\lambda(f)$, or $E(\lambda, f)$, or even by E_λ, and is called the *eigenspace associated with* λ, *or proper subspace associated with* λ.

Note that distinct eigenvectors may correspond to the same eigenvalue, but distinct eigenvalues correspond to disjoint sets of eigenvectors.

Remark: As we emphasized in the remark following Definition 9.4, we *require an eigenvector to be nonzero*. This requirement seems to have more benefits than inconveniences, even though it may considered somewhat inelegant because the set of all eigenvectors associated with an eigenvalue is not a subspace since the zero vector is excluded.

The next proposition shows that the eigenvalues of a linear map $f\colon E \to E$ are the roots of a polynomial associated with f.

Proposition 15.1. *Let E be any vector space of finite dimension n and let f be any linear map $f\colon E \to E$. The eigenvalues of f are the roots (in K) of the polynomial*

$$\det(\lambda\,\mathrm{id} - f).$$

Proof. A scalar $\lambda \in K$ is an eigenvalue of f iff there is some vector $u \neq 0$ in E such that

$$f(u) = \lambda u$$

iff

$$(\lambda\,\mathrm{id} - f)(u) = 0$$

iff $(\lambda\,\mathrm{id} - f)$ is not invertible iff, by Proposition 7.13,

$$\det(\lambda\,\mathrm{id} - f) = 0. \qquad \square$$

In view of the importance of the polynomial $\det(\lambda\,\mathrm{id} - f)$, we have the following definition.

Definition 15.2. Given any vector space E of dimension n, for any linear map $f\colon E \to E$, the polynomial $P_f(X) = \chi_f(X) = \det(X\,\mathrm{id} - f)$ is called the *characteristic polynomial of* f. For any square matrix A, the polynomial $P_A(X) = \chi_A(X) = \det(XI - A)$ is called the *characteristic polynomial of* A.

Note that we already encountered the characteristic polynomial in Section 7.7; see Definition 7.9.

Given any basis $(e_1, \ldots, e_n)$, if $A = M(f)$ is the matrix of f w.r.t. $(e_1, \ldots, e_n)$, we can compute the characteristic polynomial $\chi_f(X) = \det(X\,\mathrm{id} - f)$ of f by expanding the following determinant:

$$\det(XI - A) = \begin{vmatrix} X - a_{1\,1} & -a_{1\,2} & \ldots & -a_{1\,n} \\ -a_{2\,1} & X - a_{2\,2} & \ldots & -a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n\,1} & -a_{n\,2} & \ldots & X - a_{n\,n} \end{vmatrix}.$$

If we expand this determinant, we find that

$$\chi_A(X) = \det(XI - A) = X^n - (a_{1\,1} + \cdots + a_{n\,n})X^{n-1} + \cdots + (-1)^n \det(A).$$

The sum $\mathrm{tr}(A) = a_{1\,1} + \cdots + a_{n\,n}$ of the diagonal elements of A is called the *trace of* A. Since we proved in Section 7.7 that the characteristic polynomial only depends on the linear map f, the above shows that $\mathrm{tr}(A)$ has the same value for all matrices A representing f. Thus, the *trace of a linear map* is well-defined; we have $\mathrm{tr}(f) = \mathrm{tr}(A)$ for any matrix A representing f.

Remark: The characteristic polynomial of a linear map is sometimes defined as $\det(f - X\,\mathrm{id})$. Since

$$\det(f - X\,\mathrm{id}) = (-1)^n \det(X\,\mathrm{id} - f),$$

this makes essentially no difference but the version $\det(X\,\mathrm{id} - f)$ has the small advantage that the coefficient of X^n is $+1$.

If we write

$$\chi_A(X) = \det(XI - A) = X^n - \tau_1(A)X^{n-1} + \cdots + (-1)^k \tau_k(A)X^{n-k} + \cdots + (-1)^n \tau_n(A),$$

then we just proved that

$$\tau_1(A) = \mathrm{tr}(A) \quad \text{and} \quad \tau_n(A) = \det(A).$$

It is also possible to express $\tau_k(A)$ in terms of determinants of certain submatrices of A. For any nonempty ordered subset, $I \subseteq \{1, \ldots, n\}$, say $I = \{i_1 < \cdots < i_k\}$, let $A_{I,I}$ be the $k \times k$ submatrix of A whose jth column consists of the elements $a_{i_h\, i_j}$, where $h = 1, \ldots, k$. Equivalently, $A_{I,I}$ is the matrix obtained from A by first selecting the columns whose indices belong to I, and then the rows whose indices also belong to I. Then it can be shown that

$$\tau_k(A) = \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I = \{i_1,\ldots,i_k\} \\ i_1 < \cdots < i_k}} \det(A_{I,I});$$

see Jacobson [98], Section 3.10, just after Formula (33).

If all the roots, $\lambda_1, \ldots, \lambda_n$, of the polynomial $\det(XI - A)$ belong to the field K, then we can write

$$\chi_A(X) = \det(XI - A) = (X - \lambda_1) \cdots (X - \lambda_n),$$

where some of the λ_i's may appear more than once. Consequently,

$$\chi_A(X) = \det(XI - A) = X^n - \sigma_1(\lambda)X^{n-1} + \cdots + (-1)^k \sigma_k(\lambda)X^{n-k} + \cdots + (-1)^n \sigma_n(\lambda),$$

where

$$\sigma_k(\lambda) = \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \prod_{i \in I} \lambda_i,$$

the *kth elementary symmetric polynomial (or function)* of the λ_i's, where $\lambda = (\lambda_1, \ldots, \lambda_n)$. The elementary symmetric polynomial $\sigma_k(\lambda)$ is often denoted $E_k(\lambda)$, but this notation may be confusing in the context of linear algebra. For $n = 5$, the elementary symmetric polynomials are listed below:

$$\begin{aligned}
\sigma_0(\lambda) &= 1 \\
\sigma_1(\lambda) &= \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 \\
\sigma_2(\lambda) &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_1\lambda_5 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_2\lambda_5 \\
&\quad + \lambda_3\lambda_4 + \lambda_3\lambda_5 + \lambda_4\lambda_5 \\
\sigma_3(\lambda) &= \lambda_3\lambda_4\lambda_5 + \lambda_2\lambda_4\lambda_5 + \lambda_2\lambda_3\lambda_5 + \lambda_2\lambda_3\lambda_4 + \lambda_1\lambda_4\lambda_5 \\
&\quad + \lambda_1\lambda_3\lambda_5 + \lambda_1\lambda_3\lambda_4 + \lambda_1\lambda_2\lambda_5 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_2\lambda_3 \\
\sigma_4(\lambda) &= \lambda_1\lambda_2\lambda_3\lambda_4 + \lambda_1\lambda_2\lambda_3\lambda_5 + \lambda_1\lambda_2\lambda_4\lambda_5 + \lambda_1\lambda_3\lambda_4\lambda_5 + \lambda_2\lambda_3\lambda_4\lambda_5 \\
\sigma_5(\lambda) &= \lambda_1\lambda_2\lambda_3\lambda_4\lambda_5.
\end{aligned}$$

Since

$$\begin{aligned}
\chi_A(X) &= X^n - \tau_1(A)X^{n-1} + \cdots + (-1)^k \tau_k(A)X^{n-k} + \cdots + (-1)^n \tau_n(A) \\
&= X^n - \sigma_1(\lambda)X^{n-1} + \cdots + (-1)^k \sigma_k(\lambda)X^{n-k} + \cdots + (-1)^n \sigma_n(\lambda),
\end{aligned}$$

we have

$$\sigma_k(\lambda) = \tau_k(A), \quad k = 1, \ldots, n,$$

and in particular, the product of the eigenvalues of f is equal to $\det(A) = \det(f)$, and the sum of the eigenvalues of f is equal to the trace $\mathrm{tr}(A) = \mathrm{tr}(f)$, of f; for the record,

$$\begin{aligned}
\mathrm{tr}(f) &= \lambda_1 + \cdots + \lambda_n \\
\det(f) &= \lambda_1 \cdots \lambda_n,
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of f (and A), where some of the λ_i's may appear more than once. In particular, f is not invertible iff it admits 0 has an eigenvalue (since f is singular iff $\lambda_1 \cdots \lambda_n = \det(f) = 0$).

Remark: Depending on the field K, the characteristic polynomial $\chi_A(X) = \det(XI - A)$ may or may not have roots in K. This motivates considering *algebraically closed fields*, which are fields K such that every polynomial with coefficients in K has all its root in K. For example, over $K = \mathbb{R}$, not every polynomial has real roots. If we consider the matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

then the characteristic polynomial $\det(XI - A)$ has no real roots unless $\theta = k\pi$. However, over the field $\mathbb{C}$ of complex numbers, every polynomial has roots. For example, the matrix above has the roots $\cos\theta \pm i\sin\theta = e^{\pm i\theta}$.

Remark: It is possible to show that every linear map f over a complex vector space E must have some (complex) eigenvalue without having recourse to determinants (and the characteristic polynomial). Let $n = \dim(E)$, pick any nonzero vector $u \in E$, and consider the sequence

$$u, f(u), f^2(u), \ldots, f^n(u).$$

Since the above sequence has $n+1$ vectors and E has dimension n, these vectors must be linearly dependent, so there are some complex numbers $c_0, \ldots, c_m$, not all zero, such that

$$c_0 f^m(u) + c_1 f^{m-1}(u) + \cdots + c_m u = 0,$$

where $m \leq n$ is the largest integer such that the coefficient of $f^m(u)$ is nonzero (m must exits since we have a nontrivial linear dependency). Now because the field $\mathbb{C}$ is algebraically closed, the polynomial

$$c_0 X^m + c_1 X^{m-1} + \cdots + c_m$$

can be written as a product of linear factors as

$$c_0 X^m + c_1 X^{m-1} + \cdots + c_m = c_0(X - \lambda_1)\cdots(X - \lambda_m)$$

for some complex numbers $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$, not necessarily distinct. But then since $c_0 \neq 0$,

$$c_0 f^m(u) + c_1 f^{m-1}(u) + \cdots + c_m u = 0$$

is equivalent to

$$(f - \lambda_1\,\mathrm{id}) \circ \cdots \circ (f - \lambda_m\,\mathrm{id})(u) = 0.$$

If all the linear maps $f - \lambda_i\,\mathrm{id}$ were injective, then $(f - \lambda_1\,\mathrm{id}) \circ \cdots \circ (f - \lambda_m\,\mathrm{id})$ would be injective, contradicting the fact that $u \neq 0$. Therefore, some linear map $f - \lambda_i\,\mathrm{id}$ must have a nontrivial kernel, which means that there is some $v \neq 0$ so that

$$f(v) = \lambda_i v;$$

that is, λ_i is some eigenvalue of f and v is some eigenvector of f.

As nice as the above argument is, it does not provide a method for *finding* the eigenvalues of f, and even if we prefer avoiding determinants as much as possible, we are forced to deal with the characteristic polynomial $\det(X\,\mathrm{id} - f)$.

Definition 15.3. Let A be an $n \times n$ matrix over a field K. Assume that all the roots of the characteristic polynomial $\chi_A(X) = \det(XI - A)$ of A belong to K, which means that we can write

$$\det(XI - A) = (X - \lambda_1)^{k_1} \cdots (X - \lambda_m)^{k_m},$$

where $\lambda_1, \ldots, \lambda_m \in K$ are the distinct roots of $\det(XI - A)$ and $k_1 + \cdots + k_m = n$. The integer k_i is called the *algebraic multiplicity* of the eigenvalue λ_i, and the dimension of the eigenspace $E_{\lambda_i} = \mathrm{Ker}(\lambda_i I - A)$ is called the *geometric multiplicity* of λ_i. We denote the algebraic multiplicity of λ_i by $\mathrm{alg}(\lambda_i)$, and its geometric multiplicity by $\mathrm{geo}(\lambda_i)$.

By definition, the sum of the algebraic multiplicities is equal to n, but the sum of the geometric multiplicities can be strictly smaller.

Proposition 15.2. *Let A be an $n \times n$ matrix over a field K and assume that all the roots of the characteristic polynomial $\chi_A(X) = \det(XI - A)$ of A belong to K. For every eigenvalue λ_i of A in K, the geometric multiplicity of λ_i is always less than or equal to its algebraic multiplicity, that is,*

$$\mathrm{geo}(\lambda_i) \leq \mathrm{alg}(\lambda_i).$$

Proof. To see this, if n_i is the dimension of the eigenspace E_{λ_i} associated with the eigenvalue λ_i, we can form a basis of K^n obtained by picking a basis of E_{λ_i} and completing this linearly independent family to a basis of K^n. With respect to this new basis, our matrix is of the form

$$A' = \begin{pmatrix} \lambda_i I_{n_i} & B \\ 0 & D \end{pmatrix},$$

and a simple determinant calculation shows that

$$\det(XI - A) = \det(XI - A') = (X - \lambda_i)^{n_i} \det(XI_{n-n_i} - D).$$

Therefore, $(X - \lambda_i)^{n_i}$ divides the characteristic polynomial of A', and thus, the characteristic polynomial of A. It follows that n_i is less than or equal to the algebraic multiplicity of λ_i. ☐

The following proposition shows an interesting property of eigenspaces.

Proposition 15.3. *Let E be any vector space of finite dimension n and let f be any linear map. If $u_1, \ldots, u_m$ are eigenvectors associated with pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ of f (in K), then the family $(u_1, \ldots, u_m)$ is linearly independent.*

Proof. Assume that $(u_1, \ldots, u_m)$ is linearly dependent. Then there exists $\mu_1, \ldots, \mu_k \in K$ such that

$$\mu_1 u_{i_1} + \cdots + \mu_k u_{i_k} = 0,$$

where $1 \leq k \leq m$, $\mu_i \neq 0$ for all i, $1 \leq i \leq k$, $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$, and no proper subfamily of $(u_{i_1}, \ldots, u_{i_k})$ is linearly dependent (in other words, we consider a dependency relation with k minimal). Applying f to this dependency relation, we get

$$\mu_1 \lambda_{i_1} u_{i_1} + \cdots + \mu_k \lambda_{i_k} u_{i_k} = 0,$$

and if we multiply the original dependency relation by λ_{i_1} and subtract it from the above, we get

$$\mu_2(\lambda_{i_2} - \lambda_{i_1})u_{i_2} + \cdots + \mu_k(\lambda_{i_k} - \lambda_{i_1})u_{i_k} = 0,$$

which is a nontrivial linear dependency among a proper subfamily of $(u_{i_1}, \ldots, u_{i_k})$ since the λ_j are all distinct and the μ_i are nonzero, a contradiction. $\square$

As a corollary of Proposition 15.3 we have the following result.

Corollary 15.4. *If $\lambda_1, \ldots, \lambda_m$ are all the pairwise distinct eigenvalues of f (in K), where $m \leq n$), we have a direct sum*

$$E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m}$$

of the eigenspaces E_{λ_i}.

Unfortunately, it is not always the case that

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m}.$$

When it does is captured by the following definition.

Definition 15.4. If all the distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ of f belong to K and

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_m},$$

we say that f is *diagonalizable* (and similarly for any matrix associated with f).

Indeed, picking a basis in each E_{λ_i}, we obtain a matrix which is a diagonal matrix consisting of the eigenvalues, each λ_i occurring a number of times equal to the dimension of E_{λ_i}. This happens if the algebraic multiplicity and the geometric multiplicity of every eigenvalue are equal. *In particular, when the characteristic polynomial has n distinct roots in K*, then the geometric multiplicity and the geometric multiplicity of every eigenvalue are equal to 1, so f is diagonalizable. It can also be shown that symmetric matrices have real eigenvalues and can be diagonalized.

For a negative example, we leave it as exercise to show that the matrix

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

cannot be diagonalized, even though 1 is an eigenvalue. The problem is that the eigenspace of 1 only has dimension 1. The matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

cannot be diagonalized either, because it has no real eigenvalues, unless $\theta = k\pi$. However, over the field of complex numbers, it can be diagonalized.

15.2 Reduction to Upper Triangular Form

Unfortunately, not every linear map on a complex vector space can be diagonalized. The next best thing is to "triangularize," which means to find a basis over which the matrix has zero entries below the main diagonal. Fortunately, such a basis always exist.

We say that a square matrix A is an *upper triangular matrix* if it has the following shape,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} & \dots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & a_{2\,3} & \dots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{3\,3} & \dots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \dots & 0 & a_{n\,n} \end{pmatrix},$$

i.e., $a_{i\,j} = 0$ whenever $j < i$, $1 \leq i, j \leq n$.

Theorem 15.5. *Given any finite dimensional vector space over a field K, for any linear map $f\colon E \to E$, there is a basis $(u_1, \dots, u_n)$ with respect to which f is represented by an upper triangular matrix (in $\mathrm{M}_n(K)$) iff all the eigenvalues of f belong to K. Equivalently, for every $n \times n$ matrix $A \in \mathrm{M}_n(K)$, there is an invertible matrix P and an upper triangular matrix T (both in $\mathrm{M}_n(K)$) such that*

$$A = PTP^{-1}$$

iff all the eigenvalues of A belong to K.

Proof. If there is a basis $(u_1, \dots, u_n)$ with respect to which f is represented by an upper triangular matrix T in $\mathrm{M}_n(K)$, then since the eigenvalues of f are the diagonal entries of T, all the eigenvalues of f belong to K.

For the converse, we proceed by induction on the dimension n of E. For $n = 1$ the result is obvious. If $n > 1$, since by assumption f has all its eigenvalues in K, pick some eigenvalue $\lambda_1 \in K$ of f, and let u_1 be some corresponding (nonzero) eigenvector. We can find $n - 1$ vectors $(v_2, \dots, v_n)$ such that $(u_1, v_2, \dots, v_n)$ is a basis of E, and let F be the subspace of dimension $n - 1$ spanned by $(v_2, \dots, v_n)$. In the basis $(u_1, v_2 \dots, v_n)$, the matrix of f is of the form

$$U = \begin{pmatrix} \lambda_1 & a_{1\,2} & \dots & a_{1\,n} \\ 0 & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n\,2} & \dots & a_{n\,n} \end{pmatrix},$$

since its first column contains the coordinates of $\lambda_1 u_1$ over the basis $(u_1, v_2, \dots, v_n)$. If we let $p\colon E \to F$ be the projection defined such that $p(u_1) = 0$ and $p(v_i) = v_i$ when $2 \leq i \leq n$, the linear map $g\colon F \to F$ defined as the restriction of $p \circ f$ to F is represented by the

$(n-1) \times (n-1)$ matrix $V = (a_{i\,j})_{2 \leq i,j \leq n}$ over the basis $(v_2, \ldots, v_n)$. We need to prove that all the eigenvalues of g belong to K. However, since the entries in the first column of U are all zero for $i = 2, \ldots, n$, we get

$$\chi_U(X) = \det(XI - U) = (X - \lambda_1)\det(XI - V) = (X - \lambda_1)\chi_V(X),$$

where $\chi_U(X)$ is the characteristic polynomial of U and $\chi_V(X)$ is the characteristic polynomial of V. It follows that $\chi_V(X)$ divides $\chi_U(X)$, and since all the roots of $\chi_U(X)$ are in K, all the roots of $\chi_V(X)$ are also in K. Consequently, we can apply the induction hypothesis, and there is a basis $(u_2, \ldots, u_n)$ of F such that g is represented by an upper triangular matrix $(b_{i\,j})_{1 \leq i,j \leq n-1}$. However,

$$E = Ku_1 \oplus F,$$

and thus $(u_1, \ldots, u_n)$ is a basis for E. Since p is the projection from $E = Ku_1 \oplus F$ onto F and $g\colon F \to F$ is the restriction of $p \circ f$ to F, we have

$$f(u_1) = \lambda_1 u_1$$

and

$$f(u_{i+1}) = a_{1\,i}u_1 + \sum_{j=1}^{i} b_{i\,j}u_{j+1}$$

for some $a_{1\,i} \in K$, when $1 \leq i \leq n-1$. But then the matrix of f with respect to $(u_1, \ldots, u_n)$ is upper triangular.

For the matrix version, we assume that A is the matrix of f with respect to some basis, Then we just proved that there is a change of basis matrix P such that $A = PTP^{-1}$ where T is upper triangular. $\square$

If $A = PTP^{-1}$ where T is upper triangular, note that the diagonal entries of T are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A. Indeed, A and T have the same characteristic polynomial. Also, if A is a real matrix whose eigenvalues are all real, then P can be chosen to real, and if A is a rational matrix whose eigenvalues are all rational, then P can be chosen rational. *Since any polynomial over $\mathbb{C}$ has all its roots in $\mathbb{C}$, Theorem 15.5 implies that every complex $n \times n$ matrix can be triangularized.*

If E is a Hermitian space (see Chapter 14), the proof of Theorem 15.5 can be easily adapted to prove that there is an *orthonormal* basis $(u_1, \ldots, u_n)$ with respect to which the matrix of f is upper triangular. This is usually known as *Schur's lemma*.

Theorem 15.6. *(Schur decomposition) Given any linear map $f\colon E \to E$ over a complex Hermitian space E, there is an orthonormal basis $(u_1, \ldots, u_n)$ with respect to which f is represented by an upper triangular matrix. Equivalently, for every $n \times n$ matrix $A \in \mathrm{M}_n(\mathbb{C})$, there is a unitary matrix U and an upper triangular matrix T such that*

$$A = UTU^*.$$

If A is real and if all its eigenvalues are real, then there is an orthogonal matrix Q and a real upper triangular matrix T such that

$$A = QTQ^\top.$$

Proof. During the induction, we choose F to be the orthogonal complement of $\mathbb{C}u_1$ and we pick orthonormal bases (use Propositions 14.13 and 14.12). If E is a real Euclidean space and if the eigenvalues of f are all real, the proof also goes through with real matrices (use Propositions 12.11 and 12.10). $\square$

If λ is an eigenvalue of the matrix A and if u is an eigenvector associated with λ, from

$$Au = \lambda u,$$

we obtain

$$A^2u = A(Au) = A(\lambda u) = \lambda Au = \lambda^2 u,$$

which shows that λ^2 is an eigenvalue of A^2 for the eigenvector u. An obvious induction shows that λ^k is an eigenvalue of A^k for the eigenvector u, for all $k \geq 1$. Now, if all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are in K, it follows that $\lambda_1^k, \ldots, \lambda_n^k$ are eigenvalues of A^k. However, it is not obvious that A^k does not have other eigenvalues. In fact, this can't happen, and this can be proven using Theorem 15.5.

Proposition 15.7. *Given any $n \times n$ matrix $A \in \mathrm{M}_n(K)$ with coefficients in a field K, if all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are in K, then for every polynomial $q(X) \in K[X]$, the eigenvalues of $q(A)$ are exactly $(q(\lambda_1), \ldots, q(\lambda_n))$.*

Proof. By Theorem 15.5, there is an upper triangular matrix T and an invertible matrix P (both in $\mathrm{M}_n(K)$) such that

$$A = PTP^{-1}.$$

Since A and T are similar, they have the same eigenvalues (with the same multiplicities), so the diagonal entries of T are the eigenvalues of A. Since

$$A^k = PT^kP^{-1}, \quad k \geq 1,$$

for any polynomial $q(X) = c_0X^m + \cdots + c_{m-1}X + c_m$, we have

$$\begin{aligned}
q(A) &= c_0A^m + \cdots + c_{m-1}A + c_mI \\
&= c_0PT^mP^{-1} + \cdots + c_{m-1}PTP^{-1} + c_mPIP^{-1} \\
&= P(c_0T^m + \cdots + c_{m-1}T + c_mI)P^{-1} \\
&= Pq(T)P^{-1}.
\end{aligned}$$

Furthermore, it is easy to check that $q(T)$ is upper triangular and that its diagonal entries are $q(\lambda_1), \ldots, q(\lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the diagonal entries of T, namely the eigenvalues of A. It follows that $q(\lambda_1), \ldots, q(\lambda_n)$ are the eigenvalues of $q(A)$. $\square$

Remark: There is another way to prove Proposition 15.7 that does not use Theorem 15.5, but instead uses the fact that given any field K, there is field extension $\overline{K}$ of K ($K \subseteq \overline{K}$) such that every polynomial $q(X) = c_0X^m + \cdots + c_{m-1}X + c_m$ (of degree $m \geq 1$) with coefficients $c_i \in K$ factors as

$$q(X) = c_0(X - \alpha_1) \cdots (X - \alpha_n), \quad \alpha_i \in \overline{K}, i = 1, \ldots, n.$$

The field $\overline{K}$ is called an *algebraically closed field* (and an algebraic closure of K).

Assume that all eigenvalues $\lambda_1, \ldots, \lambda_n$ of A belong to K. Let $q(X)$ be any polynomial (in $K[X]$) and let $\mu \in \overline{K}$ be any eigenvalue of $q(A)$ (this means that μ is a zero of the characteristic polynomial $\chi_{q(A)}(X) \in K[X]$ of $q(A)$. Since $\overline{K}$ is algebraically closed, $\chi_{q(A)}(X)$ has all its roots in $\overline{K}$). We claim that $\mu = q(\lambda_i)$ for some eigenvalue λ_i of A.

Proof. (After Lax [113], Chapter 6). Since $\overline{K}$ is algebraically closed, the polynomial $\mu - q(X)$ factors as

$$\mu - q(X) = c_0(X - \alpha_1) \cdots (X - \alpha_n),$$

for some $\alpha_i \in \overline{K}$. Now $\mu I - q(A)$ is a matrix in $\mathrm{M}_n(\overline{K})$, and since μ is an eigenvalue of $q(A)$, it must be singular. We have

$$\mu I - q(A) = c_0(A - \alpha_1 I) \cdots (A - \alpha_n I),$$

and since the left-hand side is singular, so is the right-hand side, which implies that some factor $A - \alpha_i I$ is singular. This means that α_i is an eigenvalue of A, say $\alpha_i = \lambda_i$. As $\alpha_i = \lambda_i$ is a zero of $\mu - q(X)$, we get

$$\mu = q(\lambda_i),$$

which proves that μ is indeed of the form $q(\lambda_i)$ for some eigenvalue λ_i of A. $\square$

Using Theorem 15.6, we can derive two very important results.

Proposition 15.8. *If A is a Hermitian matrix (i.e. $A^* = A$), then its eigenvalues are real and A can be diagonalized with respect to an orthonormal basis of eigenvectors. In matrix terms, there is a unitary matrix U and a real diagonal matrix D such that $A = UDU^*$. If A is a real symmetric matrix (i.e. $A^\top = A$), then its eigenvalues are real and A can be diagonalized with respect to an orthonormal basis of eigenvectors. In matrix terms, there is an orthogonal matrix Q and a real diagonal matrix D such that $A = QDQ^\top$.*

Proof. By Theorem 15.6, we can write $A = UTU^*$ where $T = (t_{ij})$ is upper triangular and U is a unitary matrix. If $A^* = A$, we get

$$UTU^* = UT^*U^*,$$

and this implies that $T = T^*$. Since T is an upper triangular matrix, T^* is a lower triangular matrix, which implies that T is a diagonal matrix. Furthermore, since $T = T^*$, we have

$t_{ii} = \overline{t_{ii}}$ for $i = 1, \ldots, n$, which means that the t_{ii} are real, so T is indeed a real diagonal matrix, say D.

If we apply this result to a (real) symmetric matrix A, we obtain the fact that all the eigenvalues of a symmetric matrix are real, and by applying Theorem 15.6 again, we conclude that $A = QDQ^\top$, where Q is orthogonal and D is a real diagonal matrix. $\square$

More general versions of Proposition 15.8 are proven in Chapter 17.

When a real matrix A has complex eigenvalues, there is a version of Theorem 15.6 involving only real matrices provided that we allow T to be block upper-triangular (the diagonal entries may be 2×2 matrices or real entries).

Theorem 15.6 is not a very practical result but it is a useful theoretical result to cope with matrices that cannot be diagonalized. For example, it can be used to prove that *every* complex matrix is the limit of a sequence of diagonalizable matrices that have distinct eigenvalues!

15.3 Location of Eigenvalues

If A is an $n \times n$ complex (or real) matrix A, it would be useful to know, even roughly, where the eigenvalues of A are located in the complex plane $\mathbb{C}$. The Gershgorin discs provide some precise information about this.

Definition 15.5. For any complex $n \times n$ matrix A, for $i = 1, \ldots, n$, let

$$R'_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i\,j}|$$

and let

$$G(A) = \bigcup_{i=1}^{n} \{z \in \mathbb{C} \mid |z - a_{i\,i}| \leq R'_i(A)\}.$$

Each disc $\{z \in \mathbb{C} \mid |z - a_{i\,i}| \leq R'_i(A)\}$ is called a *Gershgorin disc* and their union $G(A)$ is called the *Gershgorin domain*. An example of Gershgorin domain for $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & i & 6 \\ 7 & 8 & 1+i \end{pmatrix}$ is illustrated in Figure 15.1.

Although easy to prove, the following theorem is very useful:

Theorem 15.9. *(Gershgorin's disc theorem) For any complex $n \times n$ matrix A, all the eigenvalues of A belong to the Gershgorin domain $G(A)$. Furthermore the following properties hold:*

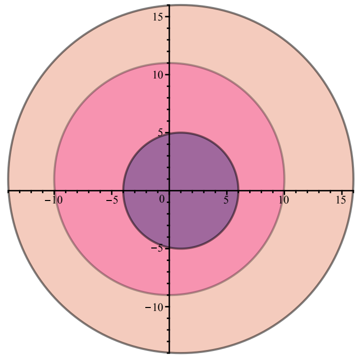


Figure 15.1: Let A be the 3×3 matrix specified at the end of Definition 15.5. For this particular A, we find that $R'_1(A) = 5$, $R'_2(A) = 10$, and $R'_3(A) = 15$. The blue/purple disk is $|z - 1| \leq 5$, the pink disk is $|z - i| \leq 10$, the peach disk is $|z - 1 - i| \leq 15$, and $G(A)$ is the union of these three disks.

(1) If A is strictly row diagonally dominant, that is

$$|a_{i\,i}| > \sum_{j=1,\, j\neq i}^{n} |a_{i\,j}|, \quad \textit{for } i = 1, \ldots, n,$$

then A is invertible.

(2) If A is strictly row diagonally dominant, and if $a_{i\,i} > 0$ for $i = 1, \ldots, n$, then every eigenvalue of A has a strictly positive real part.

Proof. Let λ be any eigenvalue of A and let u be a corresponding eigenvector (recall that we must have $u \neq 0$). Let k be an index such that

$$|u_k| = \max_{1 \leq i \leq n} |u_i|.$$

Since $Au = \lambda u$, we have

$$(\lambda - a_{k\,k})u_k = \sum_{\substack{j=1\\ j\neq k}}^{n} a_{k\,j} u_j,$$

which implies that

$$|\lambda - a_{k\,k}||u_k| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}||u_j| \leq |u_k| \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}|.$$

Since $u \neq 0$ and $|u_k| = \max_{1 \leq i \leq n} |u_i|$, we must have $|u_k| \neq 0$, and it follows that

$$|\lambda - a_{k\,k}| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{k\,j}| = R'_k(A),$$

and thus

$$\lambda \in \{z \in \mathbb{C} \mid |z - a_{k\,k}| \leq R'_k(A)\} \subseteq G(A),$$

as claimed.

(1) Strict row diagonal dominance implies that 0 does not belong to any of the Gershgorin discs, so all eigenvalues of A are nonzero, and A is invertible.

(2) If A is strictly row diagonally dominant and $a_{i\,i} > 0$ for $i = 1, \ldots, n$, then each of the Gershgorin discs lies strictly in the right half-plane, so every eigenvalue of A has a strictly positive real part. $\square$

In particular, Theorem 15.9 implies that if a symmetric matrix is strictly row diagonally dominant and has strictly positive diagonal entries, then it is positive definite. Theorem 15.9 is sometimes called the *Gershgorin–Hadamard theorem*.

Since A and $A^\top$ have the same eigenvalues (even for complex matrices) we also have a version of Theorem 15.9 for the discs of radius

$$C'_j(A) = \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{i\,j}|,$$

whose domain $G(A^\top)$ is given by

$$G(A^\top) = \bigcup_{i=1}^{n} \{z \in \mathbb{C} \mid |z - a_{i\,i}| \leq C'_i(A)\}.$$

Figure 15.2 shows $G(A^\top)$ for $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & i & 6 \\ 7 & 8 & 1+i \end{pmatrix}$.

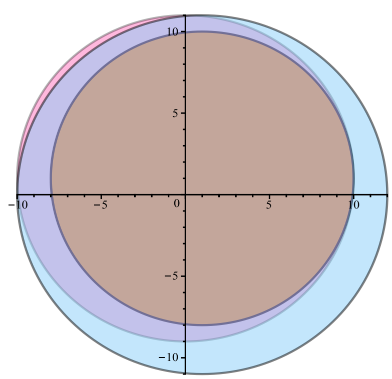


Figure 15.2: Let A be the 3×3 matrix specified at the end of Definition 15.5. For this particular A, we find that $C'_1(A) = 11$, $C'_2(A) = 10$, and $C'_3(A) = 9$. The pale blue disk is $|z - 1| \leq 11$, the pink disk is $|z - i| \leq 10$, the ocher disk is $|z - 1 - i| \leq 9$, and $G(A^\top)$ is the union of these three disks.

Thus we get the following:

Theorem 15.10. *For any complex $n \times n$ matrix A, all the eigenvalues of A belong to the intersection of the Gershgorin domains $G(A) \cap G(A^\top)$. See Figure 15.3. Furthermore the following properties hold:*

(1) If A is strictly column diagonally dominant, that is

$$|a_{i\,i}| > \sum_{i=1,\, i\neq j}^{n} |a_{i\,j}|, \quad \textit{for } j = 1, \ldots, n,$$

then A is invertible.

(2) If A is strictly column diagonally dominant, and if $a_{i\,i} > 0$ for $i = 1, \ldots, n$, then every eigenvalue of A has a strictly positive real part.

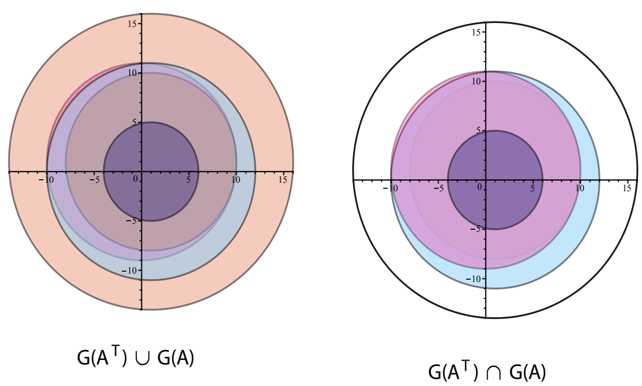


Figure 15.3: Let A be the 3×3 matrix specified at the end of Definition 15.5. The colored region in the second figure is $G(A) \cap G(A^\top)$.

There are refinements of Gershgorin's theorem and eigenvalue location results involving other domains besides discs; for more on this subject, see Horn and Johnson [95], Sections 6.1 and 6.2.

Remark: Neither strict row diagonal dominance nor strict column diagonal dominance are necessary for invertibility. Also, if we relax all strict inequalities to inequalities, then row diagonal dominance (or column diagonal dominance) is not a sufficient condition for invertibility.

15.4 Conditioning of Eigenvalue Problems

The following $n \times n$ matrix

$$A = \begin{pmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & 0 & \\ & & & & 1 & 0 \end{pmatrix}$$

has the eigenvalue 0 with multiplicity n. However, if we perturb the top rightmost entry of A by ϵ, it is easy to see that the characteristic polynomial of the matrix

$$A(\epsilon) = \begin{pmatrix} 0 & & & & & \epsilon \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & 0 & \\ & & & & 1 & 0 \end{pmatrix}$$

is $X^n - \epsilon$. It follows that if $n = 40$ and $\epsilon = 10^{-40}$, $A(10^{-40})$ has the eigenvalues $10^{-1}e^{k2\pi i/40}$ with $k = 1, \ldots, 40$. Thus, we see that a very small change ($\epsilon = 10^{-40}$) to the matrix A causes a significant change to the eigenvalues of A (from 0 to $10^{-1}e^{k2\pi i/40}$). Indeed, the relative error is 10^{-39}. Worse, due to machine precision, since very small numbers are treated as 0, the error on the computation of eigenvalues (for example, of the matrix $A(10^{-40})$) can be very large.

This phenomenon is similar to the phenomenon discussed in Section 9.5 where we studied the effect of a small perturbation of the coefficients of a linear system $Ax = b$ on its solution. In Section 9.5, we saw that the behavior of a linear system under small perturbations is governed by the condition number $\mathrm{cond}(A)$ of the matrix A. In the case of the eigenvalue problem (finding the eigenvalues of a matrix), we will see that the conditioning of the problem depends on the condition number of the change of basis matrix P used in reducing the matrix A to its diagonal form $D = P^{-1}AP$, rather than on the condition number of A itself. The following proposition in which we assume that A is diagonalizable and that the matrix norm $\| \, \|$ satisfies a special condition (satisfied by the operator norms $\| \, \|_p$ for $p = 1, 2, \infty$), is due to Bauer and Fike (1960).

Proposition 15.11. *Let $A \in \mathrm{M}_n(\mathbb{C})$ be a diagonalizable matrix, P be an invertible matrix, and D be a diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ such that*

$$A = PDP^{-1},$$

and let $\| \ \|$ be a matrix norm such that

$$\|\mathrm{diag}(\alpha_1, \ldots, \alpha_n)\| = \max_{1 \leq i \leq n} |\alpha_i|,$$

for every diagonal matrix. Then for every perturbation matrix ΔA, if we write

$$B_i = \{z \in \mathbb{C} \mid |z - \lambda_i| \leq \mathrm{cond}(P)\, \|\Delta A\|\},$$

for every eigenvalue λ of $A + \Delta A$, we have

$$\lambda \in \bigcup_{k=1}^{n} B_k.$$

Proof. Let λ be any eigenvalue of the matrix $A + \Delta A$. If $\lambda = \lambda_j$ for some j, then the result is trivial. Thus assume that $\lambda \neq \lambda_j$ for $j = 1, \ldots, n$. In this case the matrix $D - \lambda I$ is invertible (since its eigenvalues are $\lambda - \lambda_j$ for $j = 1, \ldots, n$), and we have

$$\begin{aligned} P^{-1}(A + \Delta A - \lambda I)P &= D - \lambda I + P^{-1}(\Delta A)P \\ &= (D - \lambda I)(I + (D - \lambda I)^{-1} P^{-1}(\Delta A)P). \end{aligned}$$

Since λ is an eigenvalue of $A + \Delta A$, the matrix $A + \Delta A - \lambda I$ is singular, so the matrix

$$I + (D - \lambda I)^{-1} P^{-1}(\Delta A)P$$

must also be singular. By Proposition 9.11(2), we have

$$1 \leq \left\|(D - \lambda I)^{-1} P^{-1}(\Delta A)P\right\|,$$

and since $\| \ \|$ is a matrix norm,

$$\left\|(D - \lambda I)^{-1} P^{-1}(\Delta A)P\right\| \leq \left\|(D - \lambda I)^{-1}\right\| \left\|P^{-1}\right\| \|\Delta A\| \|P\|,$$

so we have

$$1 \leq \left\|(D - \lambda I)^{-1}\right\| \left\|P^{-1}\right\| \|\Delta A\| \|P\|.$$

Now $(D - \lambda I)^{-1}$ is a diagonal matrix with entries $1/(\lambda_i - \lambda)$, so by our assumption on the norm,

$$\left\|(D - \lambda I)^{-1}\right\| = \frac{1}{\min_i(|\lambda_i - \lambda|)}.$$

As a consequence, since there is some index k for which $\min_i(|\lambda_i - \lambda|) = |\lambda_k - \lambda|$, we have

$$\left\|(D - \lambda I)^{-1}\right\| = \frac{1}{|\lambda_k - \lambda|},$$

and we obtain

$$|\lambda - \lambda_k| \leq \left\|P^{-1}\right\| \|\Delta A\| \|P\| = \mathrm{cond}(P)\, \|\Delta A\|,$$

which proves our result. $\square$

Proposition 15.11 implies that for any diagonalizable matrix A, if we define $\Gamma(A)$ by

$$\Gamma(A) = \inf\{\mathrm{cond}(P) \mid P^{-1}AP = D\},$$

then for every eigenvalue λ of $A + \Delta A$, we have

$$\lambda \in \bigcup_{k=1}^{n} \{z \in \mathbb{C}^n \mid |z - \lambda_k| \leq \Gamma(A) \, \|\Delta A\|\}.$$

Definition 15.6. The number $\Gamma(A) = \inf\{\mathrm{cond}(P) \mid P^{-1}AP = D\}$ is called the *conditioning of A relative to the eigenvalue problem*.

If A is a normal matrix, since by Theorem 17.22, A can be diagonalized with respect to a unitary matrix U, and since for the spectral norm $\|U\|_2 = 1$, we see that $\Gamma(A) = 1$. Therefore, normal matrices are very well conditionned w.r.t. the eigenvalue problem. In fact, for every eigenvalue λ of $A + \Delta A$ (with A normal), we have

$$\lambda \in \bigcup_{k=1}^{n} \{z \in \mathbb{C}^n \mid |z - \lambda_k| \leq \|\Delta A\|_2\}.$$

If A and $A+\Delta A$ are both symmetric (or Hermitian), there are sharper results; see Proposition 17.28.

Note that the matrix $A(\epsilon)$ from the beginning of the section is not normal.

15.5 Eigenvalues of the Matrix Exponential

The Schur decomposition yields a characterization of the eigenvalues of the matrix exponential e^A in terms of the eigenvalues of the matrix A. First we have the following proposition.

Proposition 15.12. *Let A and U be (real or complex) matrices and assume that U is invertible. Then*

$$e^{UAU^{-1}} = Ue^AU^{-1}.$$

Proof. A trivial induction shows that

$$UA^pU^{-1} = (UAU^{-1})^p,$$

and thus

$$\begin{aligned} e^{UAU^{-1}} &= \sum_{p \geq 0} \frac{(UAU^{-1})^p}{p!} = \sum_{p \geq 0} \frac{UA^pU^{-1}}{p!} \\ &= U \left(\sum_{p \geq 0} \frac{A^p}{p!} \right) U^{-1} = Ue^AU^{-1}, \end{aligned}$$

as claimed. $\square$

Proposition 15.13. *Given any complex $n \times n$ matrix A, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A. Furthermore, if u is an eigenvector of A for λ_i, then u is an eigenvector of e^A for e^{λ_i}.*

Proof. By Theorem 15.5, there is an invertible matrix P and an upper triangular matrix T such that

$$A = PTP^{-1}.$$

By Proposition 15.12,

$$e^{PTP^{-1}} = Pe^T P^{-1}.$$

Note that $e^T = \sum_{p \geq 0} \frac{T^p}{p!}$ is upper triangular since T^p is upper triangular for all $p \geq 0$. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the diagonal entries of T, the properties of matrix multiplication, when combined with an induction on p, imply that the diagonal entries of T^p are $\lambda_1^p, \lambda_2^p, \ldots, \lambda_n^p$. This in turn implies that the diagonal entries of e^T are $\sum_{p \geq 0} \frac{\lambda_i^p}{p!} = e^{\lambda_i}$ for $1 \leq i \leq n$. Since A and T are similar matrices, we know that they have the same eigenvalues, namely the diagonal entries $\lambda_1, \ldots, \lambda_n$ of T. Since $e^A = e^{PTP^{-1}} = Pe^T P^{-1}$, and e^T is upper triangular, we use the same argument to conclude that both e^A and e^T have the same eigenvalues, which are the diagonal entries of e^T, where the diagonal entries of e^T are of the form $e^{\lambda_1}, \ldots, e^{\lambda_n}$. Now, if u is an eigenvector of A for the eigenvalue λ, a simple induction shows that u is an eigenvector of A^n for the eigenvalue λ^n, from which is follows that

$$\begin{aligned} e^A u &= \left[I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots \right] u = u + Au + \frac{A^2}{2!} u + \frac{A^3}{3!} u + \ldots \\ &= u + \lambda u + \frac{\lambda^2}{2!} u + \frac{\lambda^3}{3!} u + \cdots = \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \right] u = e^\lambda u, \end{aligned}$$

which shows that u is an eigenvector of e^A for e^λ. $\square$

As a consequence, we obtain the following result.

Proposition 15.14. *For every complex (or real) square matrix A, we have*

$$\det(e^A) = e^{\operatorname{tr}(A)},$$

where $\operatorname{tr}(A)$ *is the trace of A, i.e., the sum $a_{1\,1} + \cdots + a_{n\,n}$ of its diagonal entries.*

Proof. The trace of a matrix A is equal to the sum of the eigenvalues of A. The determinant of a matrix is equal to the product of its eigenvalues, and if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then by Proposition 15.13, $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A, and thus

$$\det\left(e^A\right) = e^{\lambda_1} \cdots e^{\lambda_n} = e^{\lambda_1 + \cdots + \lambda_n} = e^{\operatorname{tr}(A)},$$

as desired. $\square$

If B is a skew symmetric matrix, since $\mathrm{tr}(B) = 0$, we deduce that $\det(e^B) = e^0 = 1$. This allows us to obtain the following result. Recall that the (real) vector space of skew symmetric matrices is denoted by $\mathfrak{so}(n)$.

Proposition 15.15. *For every skew symmetric matrix $B \in \mathfrak{so}(n)$, we have $e^B \in \mathbf{SO}(n)$, that is, e^B is a rotation.*

Proof. By Proposition 9.23, e^B is an orthogonal matrix. Since $\mathrm{tr}(B) = 0$, we deduce that $\det(e^B) = e^0 = 1$. Therefore, $e^B \in \mathbf{SO}(n)$. $\square$

Proposition 15.15 shows that the map $B \mapsto e^B$ is a map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$. It is not injective, but it can be shown (using one of the spectral theorems) that it is surjective.

If B is a (real) symmetric matrix, then

$$(e^B)^\top = e^{B^\top} = e^B,$$

so e^B is also symmetric. Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ of B are real, by Proposition 15.13, since the eigenvalues of e^B are $e^{\lambda_1}, \ldots, e^{\lambda_n}$ and the λ_i are real, we have $e^{\lambda_i} > 0$ for $i = 1, \ldots, n$, which implies that e^B is symmetric positive definite. In fact, it can be shown that for every symmetric positive definite matrix A, there is a *unique* symmetric matrix B such that $A = e^B$; see Gallier [72].

15.6 Summary

The main concepts and results of this chapter are listed below:

- *Diagonal matrix*.
- *Eigenvalues, eigenvectors*; the *eigenspace associated* with an eigenvalue.
- *Characteristic polynomial*.
- *Trace*.
- *Algebraic and geometric multiplicity*.
- Eigenspaces associated with distinct eigenvalues form a direct sum (Proposition 15.3).
- Reduction of a matrix to an upper-triangular matrix.
- *Schur decomposition*.
- The *Gershgorin's discs* can be used to locate the eigenvalues of a complex matrix; see Theorems 15.9 and 15.10.
- The conditioning of eigenvalue problems.
- Eigenvalues of the matrix exponential. The formula $\det(e^A) = e^{\mathrm{tr}(A)}$.

15.7 Problems

Problem 15.1. Let A be the following 2×2 matrix

$$A = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

(1) Prove that A has the eigenvalue 0 with multiplicity 2 and that $A^2 = 0$.

(2) Let A be any real 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Prove that if $bc > 0$, then A has two distinct real eigenvalues. Prove that if $a, b, c, d > 0$, then there is a positive eigenvector u associated with the largest of the two eigenvalues of A, which means that if $u = (u_1, u_2)$, then $u_1 > 0$ and $u_2 > 0$.

(3) Suppose now that A is any complex 2×2 matrix as in (2). Prove that if A has the eigenvalue 0 with multiplicity 2, then $A^2 = 0$. Prove that if A is real symmetric, then $A = 0$.

Problem 15.2. Let A be any complex $n \times n$ matrix. Prove that if A has the eigenvalue 0 with multiplicity n, then $A^n = 0$. Give an example of a matrix A such that $A^n = 0$ but $A \neq 0$.

Problem 15.3. Let A be a complex 2×2 matrix, and let λ_1 and λ_2 be the eigenvalues of A. Prove that if $\lambda_1 \neq \lambda_2$, then

$$e^A = \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I + \frac{e^{\lambda_1} - e^{\lambda_2}}{\lambda_1 - \lambda_2} A.$$

Problem 15.4. Let A be the real symmetric 2×2 matrix

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

(1) Prove that the eigenvalues of A are real and given by

$$\lambda_1 = \frac{a + c + \sqrt{4b^2 + (a-c)^2}}{2}, \quad \lambda_2 = \frac{a + c - \sqrt{4b^2 + (a-c)^2}}{2}.$$

(2) Prove that A has a double eigenvalue ($\lambda_1 = \lambda_2 = a$) if and only if $b = 0$ and $a = c$; that is, A is a diagonal matrix.

(3) Prove that the eigenvalues of A are nonnegative iff $b^2 \leq ac$ and $a + c \geq 0$.

(4) Prove that the eigenvalues of A are positive iff $b^2 < ac$, $a > 0$ and $c > 0$.

Problem 15.5. Find the eigenvalues of the matrices

$$A = \begin{pmatrix} 3 & 0 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 3 \end{pmatrix}, \quad C = A + B = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}.$$

Check that the eigenvalues of $A + B$ are not equal to the sums of eigenvalues of A plus eigenvalues of B.

Problem 15.6. Let A be a real symmetric $n \times n$ matrix and B be a real symmetric positive definite $n \times n$ matrix. We would like to solve the *generalized eigenvalue problem*: find $\lambda \in \mathbb{R}$ and $u \neq 0$ such that

$$Au = \lambda Bu. \tag{$*$}$$

(1) Use the Cholesky decomposition $B = CC^\top$ to show that λ and u are solutions of the generalized eigenvalue problem $(*)$ iff λ and v are solutions of the (ordinary) eigenvalue problem

$$C^{-1}A(C^\top)^{-1}v = \lambda v, \quad \text{with } v = C^\top u.$$

Check that $C^{-1}A(C^\top)^{-1}$ is symmetric.

(2) Prove that if $Au_1 = \lambda_1 Bu_1$, $Au_2 = \lambda_2 Bu_2$, with $u_1 \neq 0$, $u_2 \neq 0$ and $\lambda_1 \neq \lambda_2$, then $u_1^\top Bu_2 = 0$.

(3) Prove that $B^{-1}A$ and $C^{-1}A(C^\top)^{-1}$ have the same eigenvalues.

Problem 15.7. The sequence of *Fibonacci numbers*, $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots,$ is given by the recurrence

$$F_{n+2} = F_{n+1} + F_n,$$

with $F_0 = 0$ and $F_1 = 1$. In matrix form, we can write

$$\begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix}, \quad n \geq 1, \quad \begin{pmatrix} F_1 \\ F_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

(1) Show that

$$\begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

(2) Prove that the eigenvalues of the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

are

$$\lambda = \frac{1 \pm \sqrt{5}}{2}.$$

The number

$$\varphi = \frac{1 + \sqrt{5}}{2}$$

is called the *golden ratio*. Show that the eigenvalues of A are φ and $-\varphi^{-1}$.

(3) Prove that A is diagonalized as

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} \varphi & -\varphi^{-1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \varphi & 0 \\ 0 & -\varphi^{-1} \end{pmatrix} \begin{pmatrix} 1 & \varphi^{-1} \\ -1 & \varphi \end{pmatrix}.$$

Prove that

$$\begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} \varphi & -\varphi^{-1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \varphi^n \\ -(-\varphi^{-1})^n \end{pmatrix},$$

and thus

$$F_n = \frac{1}{\sqrt{5}}(\varphi^n - (-\varphi^{-1})^n) = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right], \quad n \geq 0.$$

Problem 15.8. Let A be an $n \times n$ matrix. For any subset I of $\{1, \ldots, n\}$, let $A_{I,I}$ be the matrix obtained from A by first selecting the columns whose indices belong to I, and then the rows whose indices also belong to I. Prove that

$$\tau_k(A) = \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \det(A_{I,I}).$$

Problem 15.9. (1) Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & -a_3 \\ 1 & 0 & -a_2 \\ 0 & 1 & -a_1 \end{pmatrix}.$$

Prove that the characteristic polynomial $\chi_A(z) = \det(zI - A)$ of A is given by

$$\chi_A(z) = z^3 + a_1 z^2 + a_2 z + a_3.$$

(2) Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & -a_4 \\ 1 & 0 & 0 & -a_3 \\ 0 & 1 & 0 & -a_2 \\ 0 & 0 & 1 & -a_1 \end{pmatrix}.$$

Prove that the characteristic polynomial $\chi_A(z) = \det(zI - A)$ of A is given by

$$\chi_A(z) = z^4 + a_1 z^3 + a_2 z^2 + a_3 z + a_4.$$

(3) Consider the $n \times n$ matrix (called a *companion matrix*)

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & 0 & \cdots & 0 & -a_{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & -a_2 \\ 0 & 0 & 0 & \cdots & 1 & -a_1 \end{pmatrix}.$$

Prove that the characteristic polynomial $\chi_A(z) = \det(zI - A)$ of A is given by

$$\chi_A(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + a_n.$$

Hint. Use induction.

Explain why finding the roots of a polynomial (with real or complex coefficients) and finding the eigenvalues of a (real or complex) matrix are equivalent problems, in the sense that if we have a method for solving one of these problems, then we have a method to solve the other.

Problem 15.10. Let A be a complex $n \times n$ matrix. Prove that if A is invertible and if the eigenvalues of A are $(\lambda_1, \ldots, \lambda_n)$, then the eigenvalues of A^{-1} are $(\lambda_1^{-1}, \ldots, \lambda_n^{-1})$. Prove that if u is an eigenvector of A for λ_i, then u is an eigenvector of A^{-1} for λ_i^{-1}.

Problem 15.11. Prove that every complex matrix is the limit of a sequence of diagonalizable matrices that have distinct eigenvalues.

Problem 15.12. Consider the following tridiagonal $n \times n$ matrices

$$A = \begin{pmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 1 & 0 & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & 0 & 1 & 0 \end{pmatrix}.$$

Observe that $A = 2I - S$ and show that the eigenvalues of A are $\lambda_k = 2 - \mu_k$, where the μ_k are the eigenvalues of S.

(2) Using Problem 10.6, prove that the eigenvalues of the matrix A are given by

$$\lambda_k = 4\sin^2\left(\frac{k\pi}{2(n+1)}\right), \quad k = 1, \ldots, n.$$

Show that A is symmetric positive definite.

(3) Find the condition number of A with respect to the 2-norm.

(4) Show that an eigenvector $(y_1^{(k)}, \ldots, y_n^{(k)})$ associated wih the eigenvalue λ_k is given by

$$y_j^{(k)} = \sin\left(\frac{kj\pi}{n+1}\right), \quad j = 1, \ldots, n.$$

Problem 15.13. Consider the following real tridiagonal symmetric $n \times n$ matrix

$$A = \begin{pmatrix} c & 1 & 0 & & \\ 1 & c & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & c & 1 \\ & & 0 & 1 & c \end{pmatrix}.$$

(1) Using Problem 10.6, prove that the eigenvalues of the matrix A are given by

$$\lambda_k = c + 2\cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n.$$

(2) Find a condition on c so that A is positive definite. It is satisfied by $c = 4$?

Problem 15.14. Let A be an $m \times n$ matrix and B be an $n \times m$ matrix (over $\mathbb{C}$).

(1) Prove that

$$\det(I_m - AB) = \det(I_n - BA).$$

Hint. Consider the matrices

$$X = \begin{pmatrix} I_m & A \\ B & I_n \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} I_m & 0 \\ -B & I_n \end{pmatrix}.$$

(2) Prove that

$$\lambda^n \det(\lambda I_m - AB) = \lambda^m \det(\lambda I_n - BA).$$

Hint. Consider the matrices

$$X = \begin{pmatrix} \lambda I_m & A \\ B & I_n \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} I_m & 0 \\ -B & \lambda I_n \end{pmatrix}.$$

Deduce that AB and BA have the same nonzero eigenvalues with the same multiplicity.

Problem 15.15. The purpose of this problem is to prove that the characteristic polynomial of the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 2 & 3 & 4 & 5 & \cdots & n+1 \\ 3 & 4 & 5 & 6 & \cdots & n+2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ n & n+1 & n+2 & n+3 & \cdots & 2n-1 \end{pmatrix}$$

is

$$P_A(\lambda) = \lambda^{n-2}\left(\lambda^2 - n^2\lambda - \frac{1}{12}n^2(n^2-1)\right).$$

(1) Prove that the characteristic polynomial $P_A(\lambda)$ is given by

$$P_A(\lambda) = \lambda^{n-2} P(\lambda),$$

with

$$P(\lambda) = \begin{vmatrix} \lambda - 1 & -2 & -3 & -4 & \cdots & -n+3 & -n+2 & -n+1 & -n \\ -\lambda - 1 & \lambda - 1 & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{vmatrix}.$$

(2) Prove that the sum of the roots λ_1, λ_2 of the (degree two) polynomial $P(\lambda)$ is

$$\lambda_1 + \lambda_2 = n^2.$$

The problem is thus to compute the product $\lambda_1 \lambda_2$ of these roots. Prove that

$$\lambda_1 \lambda_2 = P(0).$$

(3) The problem is now to evaluate $d_n = P(0)$, where

$$d_n = \begin{vmatrix} -1 & -2 & -3 & -4 & \cdots & -n+3 & -n+2 & -n+1 & -n \\ -1 & -1 & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{vmatrix}$$

I suggest the following strategy: cancel out the first entry in row 1 and row 2 by adding a suitable multiple of row 3 to row 1 and row 2, and then subtract row 2 from row 1.

Do this twice.

You will notice that the first two entries on row 1 and the first two entries on row 2 change, but the rest of the matrix looks the same, except that the dimension is reduced.

This suggests setting up a recurrence involving the entries u_k, v_k, x_k, y_k in the determinant

$$D_k = \begin{vmatrix} u_k & x_k & -3 & -4 & \cdots & -n+k-3 & -n+k-2 & -n+k-1 & -n+k \\ v_k & y_k & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{vmatrix},$$

starting with $k = 0$, with

$$u_0 = -1, \quad v_0 = -1, \quad x_0 = -2, \quad y_0 = -1,$$

and ending with $k = n-2$, so that

$$d_n = D_{n-2} = \begin{vmatrix} u_{n-3} & x_{n-3} & -3 \\ v_{n-3} & y_{n-3} & -1 \\ 1 & -2 & 1 \end{vmatrix} = \begin{vmatrix} u_{n-2} & x_{n-2} \\ v_{n-2} & y_{n-2} \end{vmatrix}.$$

Prove that we have the recurrence relations

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \\ x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} 2 & -2 & 1 & -1 \\ 0 & 2 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_k \\ v_k \\ x_k \\ y_k \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -2 \\ -1 \end{pmatrix}.$$

These appear to be nasty affine recurrence relations, so we will use the trick to convert this affine map to a linear map.

(4) Consider the linear map given by

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \\ x_{k+1} \\ y_{k+1} \\ 1 \end{pmatrix} = \begin{pmatrix} 2 & -2 & 1 & -1 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_k \\ v_k \\ x_k \\ y_k \\ 1 \end{pmatrix},$$

and show that its action on u_k, v_k, x_k, y_k is the same as the affine action of Part (3).

Use `Matlab` to find the eigenvalues of the matrix

$$T = \begin{pmatrix} 2 & -2 & 1 & -1 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

You will be stunned!

Let N be the matrix given by

$$N = T - I.$$

Prove that

$$N^4 = 0.$$

Use this to prove that

$$T^k = I + kN + \frac{1}{2}k(k-1)N^2 + \frac{1}{6}k(k-1)(k-2)N^3,$$

for all $k \geq 0$.

(5) Prove that

$$\begin{pmatrix} u_k \\ v_k \\ x_k \\ y_k \\ 1 \end{pmatrix} = T^k \begin{pmatrix} -1 \\ -1 \\ -2 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 & -2 & 1 & -1 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}^k \begin{pmatrix} -1 \\ -1 \\ -2 \\ -1 \\ 1 \end{pmatrix},$$

for $k \geq 0$.

Prove that

$$T^k = \begin{pmatrix} k+1 & -k(k+1) & k & -k^2 & \frac{1}{6}(k-1)k(2k-7) \\ 0 & k+1 & 0 & k & -\frac{1}{2}(k-1)k \\ -k & k^2 & 1-k & (k-1)k & -\frac{1}{3}k((k-6)k+11) \\ 0 & -k & 0 & 1-k & \frac{1}{2}(k-3)k \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and thus that

$$\begin{pmatrix} u_k \\ v_k \\ x_k \\ y_k \end{pmatrix} = \begin{pmatrix} \frac{1}{6}(2k^3+3k^2-5k-6) \\ -\frac{1}{2}(k^2+3k+2) \\ \frac{1}{3}(-k^3+k-6) \\ \frac{1}{2}(k^2+k-2) \end{pmatrix},$$

and that

$$\begin{vmatrix} u_k & x_k \\ v_k & y_k \end{vmatrix} = -1 - \frac{7}{3}k - \frac{23}{12}k^2 - \frac{2}{3}k^3 - \frac{1}{12}k^4.$$

As a consequence, prove that amazingly

$$d_n = D_{n-2} = -\frac{1}{12}n^2(n^2 - 1).$$

(6) Prove that the characteristic polynomial of A is indeed

$$P_A(\lambda) = \lambda^{n-2}\left(\lambda^2 - n^2\lambda - \frac{1}{12}n^2(n^2 - 1)\right).$$

Use the above to show that the two nonzero eigenvalues of A are

$$\lambda = \frac{n}{2}\left(n \pm \frac{\sqrt{3}}{3}\sqrt{4n^2 - 1}\right).$$

The negative eigenvalue λ_1 can also be expressed as

$$\lambda_1 = n^2\frac{(3 - 2\sqrt{3})}{6}\sqrt{1 - \frac{1}{4n^2}}.$$

Use this expression to explain the following phenomenon: if we add any number greater than or equal to $(2/25)n^2$ to every diagonal entry of A we get an invertible matrix. What about $0.077351n^2$? Try it!

Problem 15.16. Let A be a symmetric tridiagonal $n \times n$-matrix

$$A = \begin{pmatrix} b_1 & c_1 & & & & \\ c_1 & b_2 & c_2 & & & \\ & c_2 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-2} & b_{n-1} & c_{n-1} \\ & & & & c_{n-1} & b_n \end{pmatrix},$$

where it is assumed that $c_i \neq 0$ for all i, $1 \leq i \leq n-1$, and let A_k be the $k \times k$-submatrix consisting of the first k rows and columns of A, $1 \leq k \leq n$. We define the polynomials $P_k(x)$ as follows: $(0 \leq k \leq n)$.

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= b_1 - x, \\ P_k(x) &= (b_k - x)P_{k-1}(x) - c_{k-1}^2 P_{k-2}(x), \end{aligned}$$

where $2 \leq k \leq n$.

(1) Prove the following properties:

(i) $P_k(x)$ is the characteristic polynomial of A_k, where $1 \leq k \leq n$.

(ii) $\lim_{x \to -\infty} P_k(x) = +\infty$, where $1 \leq k \leq n$.

(iii) If $P_k(x) = 0$, then $P_{k-1}(x) P_{k+1}(x) < 0$, where $1 \leq k \leq n - 1$.

(iv) $P_k(x)$ has k distinct real roots that separate the $k + 1$ roots of $P_{k+1}(x)$, where $1 \leq k \leq n - 1$.

(2) Given any real number $\mu > 0$, for every k, $1 \leq k \leq n$, define the function $sg_k(\mu)$ as follows:

$$sg_k(\mu) = \begin{cases} \text{sign of } P_k(\mu) & \text{if } P_k(\mu) \neq 0, \\ \text{sign of } P_{k-1}(\mu) & \text{if } P_k(\mu) = 0. \end{cases}$$

We encode the sign of a positive number as $+$, and the sign of a negative number as $-$. Then let $E(k, \mu)$ be the ordered list

$$E(k, \mu) = \langle +, \, sg_1(\mu), \, sg_2(\mu), \, \ldots, \, sg_k(\mu) \rangle,$$

and let $N(k, \mu)$ be the number changes of sign between consecutive signs in $E(k, \mu)$.

Prove that $sg_k(\mu)$ is well defined and that $N(k, \mu)$ is the number of roots λ of $P_k(x)$ such that $\lambda < \mu$.

Remark: The above can be used to compute the eigenvalues of a (tridiagonal) symmetric matrix (the method of Givens-Householder).

Chapter 16

Unit Quaternions and Rotations in $\mathbf{SO}(3)$

This chapter is devoted to the representation of rotations in $\mathbf{SO}(3)$ in terms of unit quaternions. Since we already defined the unitary groups $\mathbf{SU}(n)$, the quickest way to introduce the *unit quaternions* is to define them as the elements of the group $\mathbf{SU}(2)$.

The skew field $\mathbb{H}$ of quaternions and the group $\mathbf{SU}(2)$ of unit quaternions are discussed in Section 16.1. In Section 16.2, we define a homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ and prove that its kernel is $\{-I, I\}$. We compute the rotation matrix R_q associated with the rotation r_q induced by a unit quaternion q in Section 16.3. In Section 16.4, we prove that the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is surjective by providing an algorithm to construct a quaternion from a rotation matrix. In Section 16.5 we define the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ where $\mathfrak{su}(2)$ is the real vector space of skew-Hermitian 2×2 matrices with zero trace. We prove that exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is surjective and give an algorithm for finding a logarithm. We discuss quaternion interpolation and prove the famous *slerp interpolation formula* due to Ken Shoemake in Section 16.6. This formula is used in robotics and computer graphics to deal with interpolation problems. In Section 16.7, we prove that there is no "nice" section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$, in the sense that any section of r is neither a homomorphism nor continuous.

16.1 The Group $\mathbf{SU}(2)$ of Unit Quaternions and the Skew Field $\mathbb{H}$ of Quaternions

Definition 16.1. The *unit quaternions* are the elements of the group $\mathbf{SU}(2)$, namely the group of 2×2 complex matrices of the form

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \quad \alpha, \beta \in \mathbb{C},\ \alpha\overline{\alpha} + \beta\overline{\beta} = 1.$$

The *quaternions* are the elements of the real vector space $\mathbb{H} = \mathbb{R}\,\mathbf{SU}(2)$.

Let $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ be the matrices

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

then $\mathbb{H}$ is the set of all matrices of the form

$$X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \quad a, b, c, d \in \mathbb{R}.$$

Indeed, every matrix in $\mathbb{H}$ is of the form

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}.$$

It is easy (but a bit tedious) to verify that the quaternions $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the famous identities discovered by Hamilton:

$$\begin{aligned}
\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} &= -\mathbf{1}, \\
\mathbf{ij} = -\mathbf{ji} &= \mathbf{k}, \\
\mathbf{jk} = -\mathbf{kj} &= \mathbf{i}, \\
\mathbf{ki} = -\mathbf{ik} &= \mathbf{j}.
\end{aligned}$$

Thus, the quaternions are a generalization of the complex numbers, but there are three square roots of $-\mathbf{1}$ and multiplication is not commutative.

Given any two quaternions $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $Y = a'\mathbf{1} + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k}$, Hamilton's famous formula

$$\begin{aligned}
XY = {} & (aa' - bb' - cc' - dd')\mathbf{1} + (ab' + ba' + cd' - dc')\mathbf{i} \\
& + (ac' + ca' + db' - bd')\mathbf{j} + (ad' + da' + bc' - cb')\mathbf{k}
\end{aligned}$$

looks mysterious, but it is simply the result of multiplying the two matrices

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} a' + ib' & c' + id' \\ -(c' - id') & a' - ib' \end{pmatrix}.$$

It is worth noting that this formula was discovered independently by Olinde Rodrigues in 1840, a few years before Hamilton (Veblen and Young [184]). However, Rodrigues was working with a different formalism, homogeneous transformations, and he did not discover the quaternions.

If

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R},$$

it is immediately verified that

$$XX^* = X^*X = (a^2 + b^2 + c^2 + d^2)\mathbf{1}.$$

Also observe that

$$X^* = \begin{pmatrix} a - ib & -(c + id) \\ c - id & a + ib \end{pmatrix} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}.$$

This implies that if $X \neq 0$, then X is invertible and its inverse is given by

$$X^{-1} = (a^2 + b^2 + c^2 + d^2)^{-1} X^*.$$

As a consequence, it can be verified that $\mathbb{H}$ is a skew field (a noncommutative field). It is also a real vector space of dimension 4 with basis $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k})$; thus as a vector space, $\mathbb{H}$ is isomorphic to $\mathbb{R}^4$.

Definition 16.2. A concise notation for the quaternion X defined by $\alpha = a + ib$ and $\beta = c + id$ is

$$X = [a, (b, c, d)].$$

We call a the *scalar part* of X and (b, c, d) the *vector part* of X. With this notation, $X^* = [a, -(b, c, d)]$, which is often denoted by $\overline{X}$. The quaternion $\overline{X}$ is called the *conjugate* of X. If q is a unit quaternion, then $\overline{q}$ is the multiplicative inverse of q.

16.2 Representation of Rotations in SO(3) by Quaternions in SU(2)

The key to representation of rotations in $\mathbf{SO}(3)$ by unit quaternions is a certain group homomorphism called the *adjoint representation of* $\mathbf{SU}(2)$. To define this mapping, first we define the real vector space $\mathfrak{su}(2)$ of skew Hermitian matrices.

Definition 16.3. The (real) vector space $\mathfrak{su}(2)$ of 2×2 *skew Hermitian matrices with zero trace* is given by

$$\mathfrak{su}(2) = \left\{ \begin{pmatrix} ix & y + iz \\ -y + iz & -ix \end{pmatrix} \;\middle|\; (x, y, z) \in \mathbb{R}^3 \right\}.$$

Observe that for every matrix $A \in \mathfrak{su}(2)$, we have $A^* = -A$, that is, A is skew Hermitian, and that $\mathrm{tr}(A) = 0$.

Definition 16.4. The *adjoint representation* of the group $\mathbf{SU}(2)$ is the group homomorphism $\mathrm{Ad}\colon \mathbf{SU}(2) \to \mathbf{GL}(\mathfrak{su}(2))$ defined such that for every $q \in \mathbf{SU}(2)$, with

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \in \mathbf{SU}(2),$$

we have

$$\mathrm{Ad}_q(A) = qAq^*, \quad A \in \mathfrak{su}(2),$$

where q^* is the inverse of q (since $\mathbf{SU}(2)$ is a unitary group) and is given by

$$q^* = \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

One needs to verify that the map Ad_q is an invertible linear map from $\mathfrak{su}(2)$ to itself, and that Ad is a group homomorphism, which is easy to do.

In order to associate a rotation ρ_q (in $\mathbf{SO}(3)$) to q, we need to embed $\mathbb{R}^3$ into $\mathbb{H}$ as the pure quaternions, by

$$\psi(x,y,z) = \begin{pmatrix} ix & y+iz \\ -y+iz & -ix \end{pmatrix}, \quad (x,y,z) \in \mathbb{R}^3.$$

Then q defines the map ρ_q (on $\mathbb{R}^3$) given by

$$\rho_q(x,y,z) = \psi^{-1}(q\psi(x,y,z)q^*).$$

Therefore, modulo the isomorphism ψ, the linear map ρ_q is the linear isomorphism Ad_q. In fact, it turns out that ρ_q is a rotation (and so is Ad_q), which we will prove shortly. So, the representation of rotations in $\mathbf{SO}(3)$ by unit quaternions is just the adjoint representation of $\mathbf{SU}(2)$; its image is a subgroup of $\mathbf{GL}(\mathfrak{su}(2))$ isomorphic to $\mathbf{SO}(3)$.

Technically, it is a bit simpler to embed $\mathbb{R}^3$ in the (real) vector spaces of Hermitian matrices with zero trace,

$$\left\{ \begin{pmatrix} x & z-iy \\ z+iy & -x \end{pmatrix} \;\middle|\; x,y,z \in \mathbb{R} \right\}.$$

Since the matrix $\psi(x,y,z)$ is skew-Hermitian, the matrix $-i\psi(x,y,z)$ is Hermitian, and we have

$$-i\psi(x,y,z) = \begin{pmatrix} x & z-iy \\ z+iy & -x \end{pmatrix} = x\sigma_3 + y\sigma_2 + z\sigma_1,$$

where $\sigma_1, \sigma_2, \sigma_3$ are the *Pauli spin matrices*

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Matrices of the form $x\sigma_3 + y\sigma_2 + z\sigma_1$ are Hermitian matrices with zero trace.

It is easy to see that every 2×2 Hermitian matrix with zero trace must be of this form. (observe that $(i\sigma_1, i\sigma_2, i\sigma_3)$ forms a basis of $\mathfrak{su}(2)$. Also, $\mathbf{i} = i\sigma_3$, $\mathbf{j} = i\sigma_2$, $\mathbf{k} = i\sigma_1$.)

Now, if $A = x\sigma_3 + y\sigma_2 + z\sigma_1$ is a Hermitian 2×2 matrix with zero trace, we have

$$(qAq^*)^* = qA^*q^* = qAq^*,$$

so qAq^* is also Hermitian, and

$$\mathrm{tr}(qAq^*) = \mathrm{tr}(Aq^*q) = \mathrm{tr}(A),$$

and qAq^* also has zero trace. Therefore, the map $A \mapsto qAq^*$ preserves the Hermitian matrices with zero trace. We also have

$$\det(x\sigma_3 + y\sigma_2 + z\sigma_1) = \det \begin{pmatrix} x & z-iy \\ z+iy & -x \end{pmatrix} = -(x^2+y^2+z^2),$$

and

$$\det(qAq^*) = \det(q)\det(A)\det(q^*) = \det(A) = -(x^2 + y^2 + z^2).$$

We can embed $\mathbb{R}^3$ into the space of Hermitian matrices with zero trace by

$$\varphi(x, y, z) = x\sigma_3 + y\sigma_2 + z\sigma_1.$$

Note that

$$\varphi = -i\psi \quad \text{and} \quad \varphi^{-1} = i\psi^{-1}.$$

Definition 16.5. The unit quaternion $q \in \mathbf{SU}(2)$ induces a map r_q on $\mathbb{R}^3$ by

$$r_q(x, y, z) = \varphi^{-1}(q\varphi(x, y, z)q^*) = \varphi^{-1}(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*).$$

The map r_q is clearly linear since φ is linear.

Proposition 16.1. *For every unit quaternion $q \in \mathbf{SU}(2)$, the linear map r_q is orthogonal, that is, $r_q \in \mathbf{O}(3)$.*

Proof. Since

$$-\|(x, y, z)\|^2 = -(x^2 + y^2 + z^2) = \det(x\sigma^3 + y\sigma^2 + z\sigma_1) = \det(\varphi(x, y, z)),$$

we have

$$\begin{aligned}-\|r_q(x, y, z)\|^2 &= \det(\varphi(r_q(x, y, z))) = \det(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*)\\ &= \det(x\sigma_3 + y\sigma_2 + z\sigma_1) = -\left\|(x, y, z)^2\right\|,\end{aligned}$$

and we deduce that r_q is an isometry. Thus, $r_q \in \mathbf{O}(3)$. ☐

In fact, r_q is a rotation, and we can show this by finding the fixed points of r_q. Let q be a unit quaternion of the form

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + ib$, $\beta = c + id$, and $a^2 + b^2 + c^2 + d^2 = 1$ $(a, b, c, d \in \mathbb{R})$.

If $b = c = d = 0$, then $q = I$ and r_q is the identity so we may assume that $(b, c, d) \neq (0, 0, 0)$.

Proposition 16.2. *If $(b, c, d) \neq (0, 0, 0)$, then the fixed points of r_q are solutions (x, y, z) of the linear system*

$$\begin{aligned}-dy + cz &= 0\\ cx - by &= 0\\ dx - bz &= 0.\end{aligned}$$

This linear system has the nontrivial solution (b, c, d) and has rank 2. Therefore, r_q has the eigenvalue 1 with multiplicity 1, and r_q is a rotation whose axis is determined by (b, c, d).

Proof. We have $r_q(x, y, z) = (x, y, z)$ iff

$$\varphi^{-1}(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*) = (x, y, z)$$

iff

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \varphi(x, y, z),$$

and since

$$\varphi(x, y, z) = x\sigma_3 + y\sigma_2 + z\sigma_1 = A$$

with

$$A = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix},$$

we see that $r_q(x, y, z) = (x, y, z)$ iff

$$qAq^* = A \quad \text{iff} \quad qA = Aq.$$

We have

$$qA = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} = \begin{pmatrix} \alpha x + \beta z + i\beta y & \alpha z - i\alpha y - \beta x \\ -\overline{\beta}x + \overline{\alpha}z + i\overline{\alpha}y & -\overline{\beta}z + i\overline{\beta}y - \overline{\alpha}x \end{pmatrix}$$

and

$$Aq = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} = \begin{pmatrix} \alpha x - \overline{\beta}z + i\overline{\beta}y & \beta x + \overline{\alpha}z - i\overline{\alpha}y \\ \alpha z + i\alpha y + \overline{\beta}x & \beta z + i\beta y - \overline{\alpha}x \end{pmatrix}.$$

By equating qA and Aq, we get

$$\begin{aligned} i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0 \\ 2\beta x + i(\alpha - \overline{\alpha})y + (\overline{\alpha} - \alpha)z &= 0 \\ 2\overline{\beta}x + i(\alpha - \overline{\alpha})y + (\alpha - \overline{\alpha})z &= 0 \\ i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0. \end{aligned}$$

The first and the fourth equation are identical and the third equation is obtained by conjugating the second, so the above system reduces to

$$\begin{aligned} i(\beta - \overline{\beta})y + (\beta + \overline{\beta})z &= 0 \\ 2\beta x + i(\alpha - \overline{\alpha})y + (\overline{\alpha} - \alpha)z &= 0. \end{aligned}$$

Replacing α by $a + ib$ and β by $c + id$, we get

$$\begin{aligned} -dy + cz &= 0 \\ cx - by + i(dx - bz) &= 0, \end{aligned}$$

which yields the equations

$$\begin{aligned}-dy + cz &= 0\\ cx - by &= 0\\ dx - bz &= 0.\end{aligned}$$

This linear system has the nontrivial solution (b, c, d) and the matrix of this system is

$$\begin{pmatrix}0 & -d & c\\ c & -b & 0\\ d & 0 & -b\end{pmatrix}.$$

Since $(b, c, d) \neq (0, 0, 0)$, this matrix always has a 2×2 submatrix which is nonsingular, so it has rank 2, and consequently its kernel is the one-dimensional space spanned by (b, c, d). Therefore, r_q has the eigenvalue 1 with multiplicity 1. If we had $\det(r_q) = -1$, then the eigenvalues of r_q would be either $(-1, 1, 1)$ or $(-1, e^{i\theta}, e^{-i\theta})$ with $\theta \neq k2\pi$ (with $k \in \mathbb{Z}$), contradicting the fact that 1 is an eigenvalue with multiplicity 1. Therefore, r_q is a rotation; in fact, its axis is determined by (b, c, d). $\square$

In summary, $q \mapsto r_q$ is a map r from $\mathbf{SU}(2)$ to $\mathbf{SO}(3)$.

Theorem 16.3. *The map $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is a homomorphism whose kernel is $\{I, -I\}$.*

Proof. This map is a homomorphism, because if $q_1, q_2 \in \mathbf{SU}(2)$, then

$$\begin{aligned}r_{q_2}(r_{q_1}(x, y, z)) &= \varphi^{-1}(q_2\varphi(r_{q_1}(x, y, z))q_2^*)\\ &= \varphi^{-1}(q_2\varphi(\varphi^{-1}(q_1\varphi(x, y, z)q_1^*))q_2^*)\\ &= \varphi^{-1}((q_2q_1)\varphi(x, y, z)(q_2q_1)^*)\\ &= r_{q_2q_1}(x, y, z).\end{aligned}$$

The computation that showed that if $(b, c, d) \neq (0, 0, 0)$, then r_q has the eigenvalue 1 with multiplicity 1 implies the following: if $r_q = I_3$, namely r_q has the eigenvalue 1 with multiplicity 3, then $(b, c, d) = (0, 0, 0)$. But then $a = \pm 1$, and so $q = \pm I_2$. Therefore, the kernel of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$. $\square$

Remark: Perhaps the quickest way to show that r maps $\mathbf{SU}(2)$ into $\mathbf{SO}(3)$ is to observe that the map r is continuous. Then, since it is known that $\mathbf{SU}(2)$ is connected, its image by r lies in the connected component of I, namely $\mathbf{SO}(3)$.

Proposition 16.2 showed that if $u = (b, c, d) \neq (0, 0, 0)$, then r_q is a rotation whose axis is determined by $u = (b, c, d)$. The angle θ of this rotation can also be determined. The following result is proven in Gallier [72] (Chapter 9).

Theorem 16.4. *Let $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ be the homomorphism of Definition 16.5. For every unit quaternion*

$$q = \begin{pmatrix} a+ib & c+id \\ -(c-id) & a-ib \end{pmatrix},$$

we have $r_q = I_3$ iff $u = (b,c,d) = 0$ iff $|a| = 1$. If $u \neq 0$, then either $a = 0$ and r_q is a rotation by π around the axis of rotation determined by the vector $u = (b,c,d)$, or $0 < |a| < 1$ and r_q is the rotation around the axis of rotation determined by the vector $u = (b,c,d)$ and the angle of rotation $\theta \neq \pi$ with $0 < \theta < 2\pi$, is given by

$$\tan(\theta/2) = \frac{\|u\|}{a}.$$

Here we are assuming that a basis (w_1, w_2) has been chosen in the plane orthogonal to $u = (b,c,d)$ such that (w_1, w_2, u) is positively oriented, that is, $\det(w_1, w_2, u) > 0$ (where w_1, w_2, u are expressed over the canonical basis (e_1, e_2, e_3), which is chosen to define positive orientation).

Remark: Under the orientation defined above, we have

$$\cos(\theta/2) = a, \quad 0 < \theta < 2\pi.$$

Note that the condition $0 < \theta < 2\pi$ implies that θ is uniquely determined by the above equation. This is not the case if we choose π such that $-\pi < \theta < \pi$ since both θ and $-\theta$ satisfy the equation, and this shows why the condition $0 < \theta < 2\pi$ is preferable. If $0 < a < 1$, then $0 < \theta < \pi$, and if $-1 < a < 0$, then $\pi < \theta < 2\pi$. In the second case, r_q is also the rotation of axis $-u$ and of angle $-(2\pi - \theta) = \theta - 2\pi$ with $0 < 2\pi - \theta < \pi$, but this time the orientation of the plane orthogonal to $-u = (b,c,d)$ is the opposite orientation from before. This orientation is given by (w_2, w_1), so that $(w_2, w_1, -u)$ has positive orientation. Since the quaternions q and $-q$ define the same rotation, we may assume that $a > 0$, in which case $0 < \theta < \pi$, but we have to remember that if $a < 0$ and if we pick $-q$ instead of q, the vector defining the axis of rotation becomes $-u$, which amounts to flipping the orientation of the plane orthogonal to the axis of rotation.

The map r is surjective, but this is not obvious. We will return to this point after finding the matrix representing r_q explicitly.

16.3 Matrix Representation of the Rotation r_q

Given a unit quaternion q of the form

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + ib$, $\beta = c + id$, and $a^2 + b^2 + c^2 + d^2 = 1$ $(a, b, c, d \in \mathbb{R})$, to find the matrix representing the rotation r_q we need to compute

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

First, we have

$$\begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix} = \begin{pmatrix} x\overline{\alpha} + z\overline{\beta} - iy\overline{\beta} & -x\beta + z\alpha - iy\alpha \\ z\overline{\alpha} + iy\overline{\alpha} - x\overline{\beta} & -z\beta - iy\beta - x\alpha \end{pmatrix}.$$

Next, we have

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x\overline{\alpha} + z\overline{\beta} - iy\overline{\beta} & -x\beta + z\alpha - iy\alpha \\ z\overline{\alpha} + iy\overline{\alpha} - x\overline{\beta} & -z\beta - iy\beta - x\alpha \end{pmatrix} =$$
$$\begin{pmatrix} (\alpha\overline{\alpha} - \beta\overline{\beta})x + i(\overline{\alpha}\beta - \alpha\overline{\beta})y + (\alpha\overline{\beta} + \overline{\alpha}\beta)z & -2\alpha\beta x - i(\alpha^2 + \beta^2)y + (\alpha^2 - \beta^2)z \\ -2\overline{\alpha}\overline{\beta}x + i(\overline{\alpha}^2 + \overline{\beta}^2)y + (\overline{\alpha}^2 - \overline{\beta}^2)z & -(\alpha\overline{\alpha} - \beta\overline{\beta})x - i(\overline{\alpha}\beta - \alpha\overline{\beta})y - (\alpha\overline{\beta} + \overline{\alpha}\beta)z \end{pmatrix}$$

Since $\alpha = a + ib$ and $\beta = c + id$, with $a, b, c, d \in \mathbb{R}$, we have

$$\begin{aligned}
\alpha\overline{\alpha} - \beta\overline{\beta} &= a^2 + b^2 - c^2 - d^2 \\
i(\overline{\alpha}\beta - \alpha\overline{\beta}) &= 2(bc - ad) \\
\alpha\overline{\beta} + \overline{\alpha}\beta &= 2(ac + bd) \\
-\alpha\beta &= -ac + bd - i(ad + bc) \\
-i(\alpha^2 + \beta^2) &= 2(ab + cd) - i(a^2 - b^2 + c^2 - d^2) \\
\alpha^2 - \beta^2 &= a^2 - b^2 - c^2 + d^2 + i2(ab - cd).
\end{aligned}$$

Using the above, we get

$$(\alpha\overline{\alpha} - \beta\overline{\beta})x + i(\overline{\alpha}\beta - \alpha\overline{\beta})y + (\alpha\overline{\beta} + \overline{\alpha}\beta)z = (a^2 + b^2 - c^2 - d^2)x + 2(bc - ad)y + 2(ac + bd)z,$$

and

$$\begin{aligned}
-2\alpha\beta x - i(\alpha^2 + \beta^2)y + (\alpha^2 - \beta^2)z &= 2(-ac + bd)x + 2(ab + cd)y + (a^2 - b^2 - c^2 + d^2)z \\
&\quad - i[2(ad + bc)x + (a^2 - b^2 + c^2 - d^2)y + 2(-ab + cd)z].
\end{aligned}$$

If we write

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \begin{pmatrix} x' & z' - iy' \\ z' + iy' & -x' \end{pmatrix},$$

we obtain

$$\begin{aligned}
x' &= (a^2 + b^2 - c^2 - d^2)x + 2(bc - ad)y + 2(ac + bd)z \\
y' &= 2(ad + bc)x + (a^2 - b^2 + c^2 - d^2)y + 2(-ab + cd)z \\
z' &= 2(-ac + bd)x + 2(ab + cd)y + (a^2 - b^2 - c^2 + d^2)z.
\end{aligned}$$

In summary, we proved the following result.

Proposition 16.5. *The matrix representing r_q is*

$$R_q = \begin{pmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2ac+2bd \\ 2bc+2ad & a^2-b^2+c^2-d^2 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}.$$

Since $a^2+b^2+c^2+d^2=1$, this matrix can also be written as

$$R_q = \begin{pmatrix} 2a^2+2b^2-1 & 2bc-2ad & 2ac+2bd \\ 2bc+2ad & 2a^2+2c^2-1 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & 2a^2+2d^2-1 \end{pmatrix}.$$

The above is the rotation matrix in Euler form induced by the quaternion q, which is the matrix corresponding to ρ_q. This is because

$$\varphi = -i\psi, \quad \varphi^{-1} = i\psi^{-1},$$

so

$$r_q(x,y,z) = \varphi^{-1}(q\varphi(x,y,z)q^*) = i\psi^{-1}(q(-i\psi(x,y,z))q^*) = \psi^{-1}(q\psi(x,y,z)q^*) = \rho_q(x,y,z),$$

and so $r_q = \rho_q$.

We showed that every unit quaternion $q \in \mathbf{SU}(2)$ induces a rotation $r_q \in \mathbf{SO}(3)$, but it is not obvious that every rotation can be represented by a quaternion. This can shown in various ways.

One way to is use the fact that every rotation in $\mathbf{SO}(3)$ is the composition of two reflections, and that every reflection σ of $\mathbb{R}^3$ can be represented by a quaternion q, in the sense that

$$\sigma(x,y,z) = -\varphi^{-1}(q\varphi(x,y,z)q^*).$$

Note the presence of the negative sign. This is the method used in Gallier [72] (Chapter 9).

16.4 An Algorithm to Find a Quaternion Representing a Rotation

Theorem 16.6. *The homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is surjective.*

Here is an algorithmic method to find a unit quaternion q representing a rotation matrix R, which provides a proof of Theorem 16.6.

Let

$$q = \begin{pmatrix} a+ib & c+id \\ -(c-id) & a-ib \end{pmatrix}, \quad a^2+b^2+c^2+d^2 = 1,\ a,b,c,d \in \mathbb{R}.$$

First observe that the trace of R_q is given by

$$\mathrm{tr}(R_q) = 3a^2 - b^2 - c^2 - d^2,$$

but since $a^2 + b^2 + c^2 + d^2 = 1$, we get $\mathrm{tr}(R_q) = 4a^2 - 1$, so

$$a^2 = \frac{\mathrm{tr}(R_q) + 1}{4}.$$

If $R \in \mathbf{SO}(3)$ is any rotation matrix and if we write

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33}, \end{pmatrix}$$

we are looking for a unit quaternion $q \in \mathbf{SU}(2)$ such that $R_q = R$. Therefore, we must have

$$a^2 = \frac{\mathrm{tr}(R) + 1}{4}.$$

We also know that

$$\mathrm{tr}(R) = 1 + 2\cos\theta,$$

where $\theta \in [0, \pi]$ is the angle of the rotation R, so we get

$$a^2 = \frac{\cos\theta + 1}{2} = \cos^2\left(\frac{\theta}{2}\right),$$

which implies that

$$|a| = \cos\left(\frac{\theta}{2}\right) \quad (0 \leq \theta \leq \pi).$$

Note that we may assume that $\theta \in [0, \pi]$, because if $\pi \leq \theta \leq 2\pi$, then $\theta - 2\pi \in [-\pi, 0]$, and then the rotation of angle $\theta - 2\pi$ and axis determined by the vector (b, c, d) is the same as the rotation of angle $2\pi - \theta \in [0, \pi]$ and axis determined by the vector $-(b, c, d)$. There are two cases.

Case 1. $\mathrm{tr}(R) \neq -1$, or equivalently $\theta \neq \pi$. In this case $a \neq 0$. Pick

$$a = \frac{\sqrt{\mathrm{tr}(R) + 1}}{2}.$$

Then by equating $R - R^\top$ and $R_q - R_q^\top$, we get

$$\begin{aligned} 4ab &= r_{32} - r_{23} \\ 4ac &= r_{13} - r_{31} \\ 4ad &= r_{21} - r_{12}, \end{aligned}$$

which yields

$$b = \frac{r_{32} - r_{23}}{4a}, \quad c = \frac{r_{13} - r_{31}}{4a}, \quad d = \frac{r_{21} - r_{12}}{4a}.$$

Case 2. $\mathrm{tr}(R) = -1$, or equivalently $\theta = \pi$. In this case $a = 0$. By equating $R + R^\top$ and $R_q + R_q^\top$, we get

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

By equating the diagonal terms of R and R_q, we also get

$$\begin{aligned} b^2 &= \frac{1 + r_{11}}{2} \\ c^2 &= \frac{1 + r_{22}}{2} \\ d^2 &= \frac{1 + r_{33}}{2}. \end{aligned}$$

Since $q \neq 0$ and $a = 0$, at least one of b, c, d is nonzero.

If $b \neq 0$, let

$$b = \frac{\sqrt{1 + r_{11}}}{\sqrt{2}},$$

and determine c, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31}. \end{aligned}$$

If $c \neq 0$, let

$$c = \frac{\sqrt{1 + r_{22}}}{\sqrt{2}},$$

and determine b, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

If $d \neq 0$, let

$$d = \frac{\sqrt{1 + r_{33}}}{\sqrt{2}},$$

and determine b, c using

$$\begin{aligned} 4bd &= r_{13} + r_{31} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

It is easy to check that whenever we computed a square root, if we had chosen a negative sign instead of a positive sign, we would obtain the quaternion $-q$. However, both q and $-q$ determine the same rotation r_q.

The above discussion involving the cases $\mathrm{tr}(R) \neq -1$ and $\mathrm{tr}(R) = -1$ is reminiscent of the procedure for finding a logarithm of a rotation matrix using the Rodrigues formula (see Section 12.7). This is not surprising, because if

$$B = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}$$

and if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$ (with $0 \leq \theta \leq 2\pi$), then the Rodrigues formula says that

$$e^B = I + \frac{\sin\theta}{\theta} B + \frac{(1-\cos\theta)}{\theta^2} B^2, \quad \theta \neq 0,$$

with $e^0 = I$. It is easy to check that $\mathrm{tr}(e^B) = 1 + 2\cos\theta$. Then it is an easy exercise to check that the quaternion q corresponding to the rotation $R = e^B$ (with $B \neq 0$) is given by

$$q = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \frac{u_3}{\theta}\right)\right].$$

So the method for finding the logarithm of a rotation R is essentially the same as the method for finding a quaternion defining R.

Remark: Geometrically, the group $\mathbf{SU}(2)$ is homeomorphic to the 3-sphere S^3 in $\mathbb{R}^4$,

$$S^3 = \{(x, y, z, t) \in \mathbb{R}^4 \mid x^2 + y^2 + z^2 + t^2 = 1\}.$$

However, since the kernel of the surjective homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$, as a topological space, $\mathbf{SO}(3)$ is homeomorphic to the quotient of S^3 obtained by identifying antipodal points (x, y, z, t) and $-(x, y, z, t)$. This quotient space is the (real) projective space $\mathbb{RP}^3$, and it is more complicated than S^3. The space S^3 is simply-connected, but $\mathbb{RP}^3$ is not.

16.5 The Exponential Map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$

Given any matrix $A \in \mathfrak{su}(2)$, with

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

it is easy to check that

$$A^2 = -\theta^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

with $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$. Then we have the following formula whose proof is very similar to the proof of the formula given in Proposition 9.22.

Proposition 16.7. *For every matrix $A \in \mathfrak{su}(2)$, with*

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$, then

$$e^A = \cos\theta I + \frac{\sin\theta}{\theta} A, \quad \theta \neq 0,$$

and $e^0 = I$.

Therefore, by the discussion at the end of the previous section, e^A is a unit quaternion representing the rotation of angle 2θ and axis (u_1, u_2, u_3) (or I when $\theta = k\pi$, $k \in \mathbb{Z}$). The above formula shows that we may assume that $0 \leq \theta \leq \pi$. Proposition 16.7 shows that the exponential yields a map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$. It is an analog of the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$.

Remark: Because $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ are real vector spaces of dimension 3, they are isomorphic, and it is easy to construct an isomorphism. In fact, $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ are isomorphic as Lie algebras, which means that there is a linear isomorphism preserving the the Lie bracket $[A, B] = AB - BA$. However, as observed earlier, the groups $\mathbf{SU}(2)$ and $\mathbf{SO}(3)$ are *not isomorphic*.

An equivalent, but often more convenient, formula is obtained by assuming that $u = (u_1, u_2, u_3)$ is a unit vector, equivalently $\det(A) = 1$, in which case $A^2 = -I$, so we have

$$e^{\theta A} = \cos\theta I + \sin\theta A.$$

Using the quaternion notation, this is read as

$$e^{\theta A} = [\cos\theta, \sin\theta\, u].$$

Proposition 16.8. *The exponential map* $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ *is surjective*

Proof. We give an algorithm to find the logarithm $A \in \mathfrak{su}(2)$ of a unit quaternion

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + bi$ and $\beta = c + id$.

If $q = I$ (*i.e.* $a = 1$), then $A = 0$. If $q = -I$ (*i.e.* $a = -1$), then

$$A = \pm\pi \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Otherwise, $a \neq \pm 1$ and $(b, c, d) \neq (0, 0, 0)$, and we are seeking some $A = \theta B \in \mathfrak{su}(2)$ with $\det(B) = 1$ and $0 < \theta < \pi$, such that, by Proposition 16.7,

$$q = e^{\theta B} = \cos\theta I + \sin\theta B.$$

Let

$$B = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

with $u = (u_1, u_2, u_3)$ a unit vector. We must have

$$a = \cos\theta, \quad e^{\theta B} - (e^{\theta B})^* = q - q^*.$$

Since $0 < \theta < \pi$, we have $\sin\theta \neq 0$, and

$$2\sin\theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix} = \begin{pmatrix} \alpha - \overline{\alpha} & 2\beta \\ -2\overline{\beta} & \overline{\alpha} - \alpha \end{pmatrix}.$$

Thus, we get

$$u_1 = \frac{1}{\sin\theta} b, \quad u_2 + iu_3 = \frac{1}{\sin\theta}(c + id);$$

that is,

$$\begin{aligned} \cos\theta &= a \qquad (0 < \theta < \pi) \\ (u_1, u_2, u_3) &= \frac{1}{\sin\theta}(b, c, d). \end{aligned}$$

Since $a^2 + b^2 + c^2 + d^2 = 1$ and $a = \cos\theta$, the vector $(b, c, d)/\sin\theta$ is a unit vector. Furthermore if the quaternion q is of the form $q = [\cos\theta, \sin\theta u]$ where $u = (u_1, u_2, u_3)$ is a unit vector (with $0 < \theta < \pi$), then

$$A = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix} \tag{$*_{\log}$}$$

is a logarithm of q. $\square$

Observe that not only is the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ surjective, but the above proof shows that it is injective on the open ball

$$\{\theta B \in \mathfrak{su}(2) \mid \det(B) = 1, 0 \leq \theta < \pi\}.$$

Also, unlike the situation where in computing the logarithm of a rotation matrix $R \in \mathbf{SO}(3)$ we needed to treat the case where $\mathrm{tr}(R) = -1$ (the angle of the rotation is π) in a special way, computing the logarithm of a quaternion (other than $\pm I$) does not require any case analysis; no special case is needed when the angle of rotation is π.

16.6 Quaternion Interpolation ⊛

We are now going to derive a formula for interpolating between two quaternions. This formula is due to Ken Shoemake, once a Penn student and my TA! Since rotations in $\mathbf{SO}(3)$ can be defined by quaternions, this has applications to computer graphics, robotics, and computer vision.

First we observe that multiplication of quaternions can be expressed in terms of the inner product and the cross-product in $\mathbb{R}^3$. Indeed, if $q_1 = [a, u_1]$ and $q_2 = [a_2, u_2]$, it can be verified that

$$q_1 q_2 = [a_1, u_1][a_2, u_2] = [a_1 a_2 - u_1 \cdot u_2,\ a_1 u_2 + a_2 u_1 + u_1 \times u_2]. \qquad (*_{\text{mult}})$$

We will also need the identity

$$u \times (u \times v) = (u \cdot v)u - (u \cdot u)v.$$

Given a quaternion q expressed as $q = [\cos\theta, \sin\theta\, u]$, where u is a unit vector, we can interpolate between I and q by finding the logs of I and q, interpolating in $\mathfrak{su}(2)$, and then exponentiating. We have

$$A = \log(I) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \log(q) = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

and so $q = e^B$. Since $\mathbf{SU}(2)$ is a compact Lie group and since the inner product on $\mathfrak{su}(2)$ given by

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$$

is $\mathrm{Ad}(\mathbf{SU}(2))$-invariant, it induces a biinvariant Riemannian metric on $\mathbf{SU}(2)$, and the curve

$$\lambda \mapsto e^{\lambda B}, \quad \lambda \in [0, 1]$$

is a geodesic from I to q in $\mathbf{SU}(2)$. We write $q^\lambda = e^{\lambda B}$. Given two quaternions q_1 and q_2, because the metric is left invariant, the curve

$$\lambda \mapsto Z(\lambda) = q_1(q_1^{-1} q_2)^\lambda, \quad \lambda \in [0, 1]$$

is a geodesic from q_1 to q_2. Remarkably, there is a closed-form formula for the interpolant $Z(\lambda)$.

Say $q_1 = [\cos\theta, \sin\theta\, u]$ and $q_2 = [\cos\varphi, \sin\varphi\, v]$, and assume that $q_1 \neq q_2$ and $q_1 \neq -q_2$. First, we compute $q^{-1} q_2$. Since $q^{-1} = [\cos\theta, -\sin\theta\, u]$, we have

$$q^{-1} q_2 = [\cos\theta\cos\varphi + \sin\theta\sin\varphi(u \cdot v),\ -\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u \times v)].$$

Define Ω by

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi(u \cdot v). \qquad (*_{\Omega})$$

Since $q_1 \neq q_2$ and $q_1 \neq -q_2$, we have $0 < \Omega < \pi$, so we get

$$q_1^{-1}q_2 = \left[\cos\Omega,\ \sin\Omega\,\frac{(-\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u\times v)}{\sin\Omega}\right],$$

where the term multiplying $\sin\Omega$ is a unit vector because q_1 and q_2 are unit quaternions, so $q_1^{-1}q_2$ is also a unit quaternion. By $(*_{\log})$, we have

$$(q_1^{-1}q_2)^\lambda = \left[\cos\lambda\Omega,\ \sin\lambda\Omega\,\frac{(-\sin\theta\cos\varphi\, u + \cos\theta\sin\varphi\, v - \sin\theta\sin\varphi(u\times v)}{\sin\Omega}\right].$$

Next we need to compute $q_1(q_1^{-1}q_2)^\lambda$. The scalar part of this product is

$$\begin{aligned} s = \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi(u\cdot u) - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta\sin\varphi\cos\theta(u\cdot v)&\\ + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\cdot(u\times v)).& \end{aligned}$$

Since $u\cdot(u\times v) = 0$, the last term is zero, and since $u\cdot u = 1$ and

$$\sin\theta\sin\varphi(u\cdot v) = \cos\Omega - \cos\theta\cos\varphi,$$

we get

$$\begin{aligned} s &= \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi - \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta(\cos\Omega - \cos\theta\cos\varphi)\\ &= \cos\theta\cos\lambda\Omega + \frac{\sin\lambda\Omega}{\sin\Omega}(\sin^2\theta + \cos^2\theta)\cos\varphi - \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\cos\Omega\\ &= \frac{(\cos\lambda\Omega\sin\Omega - \sin\lambda\Omega\cos\Omega)\cos\theta}{\sin\Omega} + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi\\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\cos\theta + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi. \end{aligned}$$

The vector part of the product $q_1(q_1^{-1}q_2)^\lambda$ is given by

$$\begin{aligned} \nu = -\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\cos^2\theta\sin\varphi\, v - \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\sin\varphi(u\times v)&\\ + \cos\lambda\Omega\sin\theta\, u - \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\cos\varphi(u\times u) + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\sin\varphi(u\times v)&\\ - \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\times(u\times v)).& \end{aligned}$$

We have $u\times u = 0$, the two terms involving $u\times v$ cancel out,

$$u\times(u\times v) = (u\cdot v)u - (u\cdot u)v,$$

and $u \cdot u = 1$, so we get

$$\begin{aligned}\nu = -\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\cos^2\theta\sin\varphi\, v \\ + \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi\, v - \frac{\sin\lambda\Omega}{\sin\Omega}\sin^2\theta\sin\varphi(u\cdot v)u.\end{aligned}$$

Using

$$\sin\theta\sin\varphi(u\cdot v) = \cos\Omega - \cos\theta\cos\varphi,$$

we get

$$\begin{aligned}\nu &= -\frac{\sin\lambda\Omega}{\sin\Omega}\cos\theta\sin\theta\cos\varphi\, u + \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v \\ &\quad - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta(\cos\Omega - \cos\theta\cos\varphi)u \\ &= \cos\lambda\Omega\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v - \frac{\sin\lambda\Omega}{\sin\Omega}\sin\theta\cos\Omega\, u \\ &= \frac{(\cos\lambda\Omega\sin\Omega - \sin\lambda\Omega\cos\Omega)}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v \\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v.\end{aligned}$$

Putting the scalar part and the vector part together, we obtain

$$\begin{aligned}q_1(q_1^{-1}q_2)^\lambda &= \left[\frac{\sin(1-\lambda)\Omega}{\sin\Omega}\cos\theta + \frac{\sin\lambda\Omega}{\sin\Omega}\cos\varphi,\ \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\sin\theta\, u + \frac{\sin\lambda\Omega}{\sin\Omega}\sin\varphi\, v\right], \\ &= \frac{\sin(1-\lambda)\Omega}{\sin\Omega}[\cos\theta,\ \sin\theta\, u] + \frac{\sin\lambda\Omega}{\sin\Omega}[\cos\varphi,\ \sin\varphi\, v].\end{aligned}$$

This yields the celebrated *slerp interpolation formula*

$$Z(\lambda) = q_1(q_1^{-1}q_2)^\lambda = \frac{\sin(1-\lambda)\Omega}{\sin\Omega}q_1 + \frac{\sin\lambda\Omega}{\sin\Omega}q_2,$$

with

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi(u\cdot v).$$

16.7 Nonexistence of a "Nice" Section from $\mathbf{SO}(3)$ to $\mathbf{SU}(2)$

We conclude by discussing the problem of a consistent choice of sign for the quaternion q representing a rotation $R = \rho_q \in \mathbf{SO}(3)$. We are looking for a "nice" section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$, that is, a function s satisfying the condition

$$\rho \circ s = \mathrm{id},$$

where ρ is the surjective homomorphism $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.

Proposition 16.9. *Any section* $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ *of* ρ *is neither a homomorphism nor continuous.*

Intuitively, this means that there is no "nice and simple " way to pick the sign of the quaternion representing a rotation.

The following proof is due to Marcel Berger.

Proof. Let Γ be the subgroup of $\mathbf{SU}(2)$ consisting of all quaternions of the form $q = [a, (b, 0, 0)]$. Then, using the formula for the rotation matrix R_q corresponding to q (and the fact that $a^2 + b^2 = 1$), we get

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2a^2 - 1 & -2ab \\ 0 & 2ab & 2a^2 - 1 \end{pmatrix}.$$

Since $a^2 + b^2 = 1$, we may write $a = \cos\theta, b = \sin\theta$, and we see that

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos 2\theta & -\sin 2\theta \\ 0 & \sin 2\theta & \cos 2\theta \end{pmatrix},$$

a rotation of angle 2θ around the x-axis. Thus, both Γ and its image are isomorphic to $\mathbf{SO}(2)$, which is also isomorphic to $\mathbf{U}(1) = \{w \in \mathbb{C} \mid |w| = 1\}$. By identifying $\mathbf{i}$ and i, and identifying Γ and its image to $\mathbf{U}(1)$, if we write $w = \cos\theta + i\sin\theta \in \Gamma$, the restriction of the map ρ to Γ is given by $\rho(w) = w^2$.

We claim that any section s of ρ is not a homomorphism. Consider the restriction of s to $\mathbf{U}(1)$. Then since $\rho \circ s = \mathrm{id}$ and $\rho(w) = w^2$, for $-1 \in \rho(\Gamma) \approx \mathbf{U}(1)$, we have

$$-1 = \rho(s(-1)) = (s(-1))^2.$$

On the other hand, if s is a homomorphism, then

$$(s(-1))^2 = s((-1)^2) = s(1) = 1,$$

contradicting $(s(-1))^2 = -1$.

We also claim that s is not continuous. Assume that $s(1) = 1$, the case where $s(1) = -1$ being analogous. Then s is a bijection inverting ρ on Γ whose restriction to $\mathbf{U}(1)$ must be given by

$$s(\cos\theta + i\sin\theta) = \cos(\theta/2) + \mathbf{i}\sin(\theta/2), \quad -\pi \leq \theta < \pi.$$

If θ tends to π, that is $z = \cos\theta + i\sin\theta$ tends to -1 in the upper-half plane, then $s(z)$ tends to $\mathbf{i}$, but if θ tends to $-\pi$, that is z tends to -1 in the lower-half plane, then $s(z)$ tends to $-\mathbf{i}$, which shows that s is not continuous. $\square$

Another way (due to Jean Dieudonné) to prove that a section s of ρ is not a homomorphism is to prove that any unit quaternion is the product of two unit pure quaternions. Indeed, if $q = [a, u]$ is a unit quaternion, if we let $q_1 = [0, u_1]$, where u_1 is any unit vector orthogonal to u, then

$$q_1 q = [-u_1 \cdot u, au_1 + u_1 \times u] = [0, au_1 + u_1 \times u] = q_2$$

is a nonzero unit pure quaternion. This is because if $a \neq 0$ then $au_1 + u_1 \times u \neq 0$ (since $u_1 \times u$ is orthogonal to $au_1 \neq 0$), and if $a = 0$ then $u \neq 0$, so $u_1 \times u \neq 0$ (since u_1 is orthogonal to u). But then, $q_1^{-1} = [0, -u_1]$ is a unit pure quaternion and we have

$$q = q_1^{-1} q_2,$$

a product of two pure unit quaternions.

We also observe that for any two pure quaternions q_1, q_2, there is some unit quaternion q such that

$$q_2 = q q_1 q^{-1}.$$

This is just a restatement of the fact that the group $\mathbf{SO}(3)$ is transitive. Since the kernel of $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is $\{I, -I\}$, the subgroup $s(\mathbf{SO}(3))$ would be a normal subgroup of index 2 in $\mathbf{SU}(2)$. Then we would have a surjective homomorphism η from $\mathbf{SU}(2)$ onto the quotient group $\mathbf{SU}(2)/s(\mathbf{SO}(3))$, which is isomorphic to $\{1, -1\}$. Now, since any two pure quaternions are conjugate of each other, η would have a constant value on the unit pure quaternions. Since $\mathbf{k} = \mathbf{ij}$, we would have

$$\eta(\mathbf{k}) = \eta(\mathbf{ij}) = (\eta(\mathbf{i}))^2 = 1.$$

Consequently, η would map all pure unit quaternions to 1. But since every unit quaternion is the product of two pure quaternions, η would map every unit quaternion to 1, contradicting the fact that it is surjective onto $\{-1, 1\}$.

16.8 Summary

The main concepts and results of this chapter are listed below:

- The group $\mathbf{SU}(2)$ of unit quaternions.
- The skew field $\mathbb{H}$ of quaternions.
- Hamilton's identities.
- The (real) vector space $\mathfrak{su}(2)$ of 2×2 skew Hermitian matrices with zero trace.
- The adjoint representation of $\mathbf{SU}(2)$.

- The (real) vector space $\mathfrak{su}(2)$ of 2×2 Hermitian matrices with zero trace.
- The group homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$; $\mathrm{Ker}\,(r) = \{+I, -I\}$.
- The matrix representation R_q of the rotation r_q induced by a unit quaternion q.
- Surjectivity of the homomorphism $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.
- The exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$.
- Surjectivity of the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$.
- Finding a logarithm of a quaternion.
- Quaternion interpolation.
- Shoemake's slerp interpolation formula.
- Sections $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ of $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.

16.9 Problems

Problem 16.1. Verify the quaternion identities

$$\begin{aligned}
&\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1},\\
&\mathbf{ij} = -\mathbf{ji} = \mathbf{k},\\
&\mathbf{jk} = -\mathbf{kj} = \mathbf{i},\\
&\mathbf{ki} = -\mathbf{ik} = \mathbf{j}.
\end{aligned}$$

Problem 16.2. Check that for every quaternion $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, we have

$$XX^* = X^*X = (a^2 + b^2 + c^2 + d^2)\mathbf{1}.$$

Conclude that if $X \neq 0$, then X is invertible and its inverse is given by

$$X^{-1} = (a^2 + b^2 + c^2 + d^2)^{-1}X^*.$$

Problem 16.3. Given any two quaternions $X = a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}$ and $Y = a'\mathbf{1}+b'\mathbf{i}+c'\mathbf{j}+d'\mathbf{k}$, prove that

$$\begin{aligned}
XY = {} & (aa' - bb' - cc' - dd')\mathbf{1} + (ab' + ba' + cd' - dc')\mathbf{i}\\
& + (ac' + ca' + db' - bd')\mathbf{j} + (ad' + da' + bc' - cb')\mathbf{k}.
\end{aligned}$$

Also prove that if $X = [a, U]$ and $Y = [a', U']$, the quaternion product XY can be expressed as

$$XY = [aa' - U \cdot U',\, aU' + a'U + U \times U'].$$

Problem 16.4. Let $\mathrm{Ad}\colon \mathbf{SU}(2) \to \mathbf{GL}(\mathfrak{su}(2))$ be the map defined such that for every $q \in \mathbf{SU}(2)$,
$$\mathrm{Ad}_q(A) = qAq^*, \quad A \in \mathfrak{su}(2),$$
where q^* is the inverse of q (since $\mathbf{SU}(2)$ is a unitary group) Prove that the map Ad_q is an invertible linear map from $\mathfrak{su}(2)$ to itself and that Ad is a group homomorphism.

Problem 16.5. Prove that every Hermitian matrix with zero trace is of the form $x\sigma_3 + y\sigma_2 + z\sigma_1$, with
$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$
Check that $\mathbf{i} = i\sigma_3$, $\mathbf{j} = i\sigma_2$, and that $\mathbf{k} = i\sigma_1$.

Problem 16.6. If
$$B = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix},$$
and if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$ (with $0 \le \theta \le \pi$), then the Rodrigues formula says that
$$e^B = I + \frac{\sin\theta}{\theta} B + \frac{(1 - \cos\theta)}{\theta^2} B^2, \quad \theta \neq 0,$$
with $e^0 = I$. Check that $\mathrm{tr}(e^B) = 1 + 2\cos\theta$. Prove that the quaternion q corresponding to the rotation $R = e^B$ (with $B \neq 0$) is given by
$$q = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \frac{u_3}{\theta}\right)\right].$$

Problem 16.7. For every matrix $A \in \mathfrak{su}(2)$, with
$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$
prove that if we write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$, then
$$e^A = \cos\theta I + \frac{\sin\theta}{\theta} A, \quad \theta \neq 0,$$
and $e^0 = I$. Conclude that e^A is a unit quaternion representing the rotation of angle 2θ and axis (u_1, u_2, u_3) (or I when $\theta = k\pi$, $k \in \mathbb{Z}$).

Problem 16.8. Write a `Matlab` program implementing the method of Section 16.4 for finding a unit quaternion corresponding to a rotation matrix.

Problem 16.9. Show that there is a very simple method for producing an orthonormal frame in $\mathbb{R}^4$ whose first vector is any given nonnull vector (a, b, c, d).

Problem 16.10. Let i, j, and k, be the unit vectors of coordinates $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ in $\mathbb{R}^3$.

(1) Describe geometrically the rotations defined by the following quaternions:

$$p = (0, i), \quad q = (0, j).$$

Prove that the interpolant $Z(\lambda) = p(p^{-1}q)^{\lambda}$ is given by

$$Z(\lambda) = (0, \cos(\lambda\pi/2)i + \sin(\lambda\pi/2)j)\,.$$

Describe geometrically what this rotation is.

(2) Repeat Question (1) with the rotations defined by the quaternions

$$p = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}i\right), \quad q = (0, j).$$

Prove that the interpolant $Z(\lambda)$ is given by

$$Z(\lambda) = \left(\frac{1}{2}\cos(\lambda\pi/2),\ \frac{\sqrt{3}}{2}\cos(\lambda\pi/2)i + \sin(\lambda\pi/2)j\right).$$

Describe geometrically what this rotation is.

(3) Repeat Question (1) with the rotations defined by the quaternions

$$p = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i\right), \quad q = \left(0, \frac{1}{\sqrt{2}}(i + j)\right).$$

Prove that the interpolant $Z(\lambda)$ is given by

$$\begin{aligned} Z(\lambda) = \Bigg(&\frac{1}{\sqrt{2}}\cos(\lambda\pi/3) - \frac{1}{\sqrt{6}}\sin(\lambda\pi/3), \\ &\qquad (1/\sqrt{2}\cos(\lambda\pi/3) + 1/\sqrt{6}\sin(\lambda\pi/3))i + \frac{2}{\sqrt{6}}\sin(\lambda\pi/3)j\Bigg). \end{aligned}$$

Problem 16.11. Prove that

$$w \times (u \times v) = (w \cdot v)u - (u \cdot w)v.$$

Conclude that

$$u \times (u \times v) = (u \cdot v)u - (u \cdot u)v.$$

Chapter 17

Spectral Theorems in Euclidean and Hermitian Spaces

17.1 Introduction

The goal of this chapter is to show that there are nice normal forms for symmetric matrices, skew-symmetric matrices, orthogonal matrices, and normal matrices. The spectral theorem for symmetric matrices states that symmetric matrices have real eigenvalues and that they can be diagonalized over an orthonormal basis. The spectral theorem for Hermitian matrices states that Hermitian matrices also have real eigenvalues and that they can be diagonalized over a complex orthonormal basis. Normal real matrices can be block diagonalized over an orthonormal basis with blocks having size at most two and there are refinements of this normal form for skew-symmetric and orthogonal matrices.

The spectral result for real symmetric matrices can be used to prove two characterizations of the eigenvalues of a symmetric matrix in terms of the *Rayleigh ratio*. The first characterization is the *Rayleigh–Ritz theorem* and the second one is the *Courant–Fischer theorem*. Both results are used in optimization theory and to obtain results about perturbing the eigenvalues of a symmetric matrix.

In this chapter all vector spaces are finite-dimensional real or complex vector spaces.

17.2 Normal Linear Maps: Eigenvalues and Eigenvectors

We begin by studying normal maps, to understand the structure of their eigenvalues and eigenvectors. This section and the next three were inspired by Lang [109], Artin [7], Mac Lane and Birkhoff [118], Berger [11], and Bertin [15].

Definition 17.1. Given a Euclidean or Hermitian space E, a linear map $f\colon E \to E$ is *normal* if

$$f \circ f^* = f^* \circ f.$$

A linear map $f\colon E \to E$ is *self-adjoint* if $f = f^*$, *skew-self-adjoint* if $f = -f^*$, and *orthogonal* if $f \circ f^* = f^* \circ f = \mathrm{id}$.

Obviously, a self-adjoint, skew-self-adjoint, or orthogonal linear map is a normal linear map. Our first goal is to show that for every normal linear map $f\colon E \to E$, there is an orthonormal basis (w.r.t. $\langle -, - \rangle$) such that the matrix of f over this basis has an especially nice form: it is a block diagonal matrix in which the blocks are either one-dimensional matrices (i.e., single entries) or two-dimensional matrices of the form

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

This normal form can be further refined if f is self-adjoint, skew-self-adjoint, or orthogonal. As a first step we show that f and f^* have the same kernel when f is normal.

Proposition 17.1. *Given a Euclidean space E, if $f\colon E \to E$ is a normal linear map, then* $\mathrm{Ker}\, f = \mathrm{Ker}\, f^*$.

Proof. First let us prove that

$$\langle f(u), f(v) \rangle = \langle f^*(u), f^*(v) \rangle$$

for all $u, v \in E$. Since f^* is the adjoint of f and $f \circ f^* = f^* \circ f$, we have

$$\begin{aligned} \langle f(u), f(u) \rangle &= \langle u, (f^* \circ f)(u) \rangle, \\ &= \langle u, (f \circ f^*)(u) \rangle, \\ &= \langle f^*(u), f^*(u) \rangle. \end{aligned}$$

Since $\langle -, - \rangle$ is positive definite,

$$\begin{aligned} \langle f(u), f(u) \rangle = 0 \quad &\text{iff} \quad f(u) = 0, \\ \langle f^*(u), f^*(u) \rangle = 0 \quad &\text{iff} \quad f^*(u) = 0, \end{aligned}$$

and since

$$\langle f(u), f(u) \rangle = \langle f^*(u), f^*(u) \rangle,$$

we have

$$f(u) = 0 \quad \text{iff} \quad f^*(u) = 0.$$

Consequently, $\mathrm{Ker}\, f = \mathrm{Ker}\, f^*$. ∎

Assuming again that E is a Hermitian space, observe that Proposition 17.1 also holds. We deduce the following corollary.

Proposition 17.2. *Given a Hermitian space E, for any normal linear map $f\colon E \to E$, we have* $\mathrm{Ker}\,(f) \cap \mathrm{Im}(f) = (0)$.

Proof. Assume $v \in \mathrm{Ker}\,(f) \cap \mathrm{Im}(f)$, which means that $v = f(u)$ for some $u \in E$, and $f(v) = 0$. By Proposition 17.1, $\mathrm{Ker}\,(f) = \mathrm{Ker}\,(f^*)$, so $f(v) = 0$ implies that $f^*(v) = 0$. Consequently,

$$\begin{aligned} 0 &= \langle f^*(v), u\rangle \\ &= \langle v, f(u)\rangle \\ &= \langle v, v\rangle, \end{aligned}$$

and thus, $v = 0$. ☐

We also have the following crucial proposition relating the eigenvalues of f and f^*.

Proposition 17.3. *Given a Hermitian space E, for any normal linear map $f\colon E \to E$, a vector u is an eigenvector of f for the eigenvalue λ (in $\mathbb{C}$) iff u is an eigenvector of f^* for the eigenvalue $\overline{\lambda}$.*

Proof. First it is immediately verified that the adjoint of $f - \lambda\,\mathrm{id}$ is $f^* - \overline{\lambda}\,\mathrm{id}$. Furthermore, $f - \lambda\,\mathrm{id}$ is normal. Indeed,

$$\begin{aligned} (f - \lambda\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id})^* &= (f - \lambda\,\mathrm{id}) \circ (f^* - \overline{\lambda}\,\mathrm{id}), \\ &= f \circ f^* - \overline{\lambda} f - \lambda f^* + \lambda\overline{\lambda}\,\mathrm{id}, \\ &= f^* \circ f - \lambda f^* - \overline{\lambda} f + \overline{\lambda}\lambda\,\mathrm{id}, \\ &= (f^* - \overline{\lambda}\,\mathrm{id}) \circ (f - \lambda\,\mathrm{id}), \\ &= (f - \lambda\,\mathrm{id})^* \circ (f - \lambda\,\mathrm{id}). \end{aligned}$$

Applying Proposition 17.1 to $f - \lambda\,\mathrm{id}$, for every nonnull vector u, we see that

$$(f - \lambda\,\mathrm{id})(u) = 0 \quad \text{iff} \quad (f^* - \overline{\lambda}\,\mathrm{id})(u) = 0,$$

which is exactly the statement of the proposition. ☐

The next proposition shows a very important property of normal linear maps: eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proposition 17.4. *Given a Hermitian space E, for any normal linear map $f\colon E \to E$, if u and v are eigenvectors of f associated with the eigenvalues λ and μ (in $\mathbb{C}$) where $\lambda \neq \mu$, then $\langle u, v\rangle = 0$.*

Proof. Let us compute $\langle f(u), v\rangle$ in two different ways. Since v is an eigenvector of f for μ, by Proposition 17.3, v is also an eigenvector of f^* for $\overline{\mu}$, and we have

$$\langle f(u), v\rangle = \langle \lambda u, v\rangle = \lambda\langle u, v\rangle,$$

and

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle = \langle u, \overline{\mu} v\rangle = \mu\langle u, v\rangle,$$

where the last identity holds because of the semilinearity in the second argument. Thus

$$\lambda\langle u, v\rangle = \mu\langle u, v\rangle,$$

that is,

$$(\lambda - \mu)\langle u, v\rangle = 0,$$

which implies that $\langle u, v\rangle = 0$, since $\lambda \neq \mu$. $\square$

We can show easily that the eigenvalues of a self-adjoint linear map are real.

Proposition 17.5. *Given a Hermitian space E, all the eigenvalues of any self-adjoint linear map $f\colon E \to E$ are real.*

Proof. Let z (in $\mathbb{C}$) be an eigenvalue of f and let u be an eigenvector for z. We compute $\langle f(u), u\rangle$ in two different ways. We have

$$\langle f(u), u\rangle = \langle zu, u\rangle = z\langle u, u\rangle,$$

and since $f = f^*$, we also have

$$\langle f(u), u\rangle = \langle u, f^*(u)\rangle = \langle u, f(u)\rangle = \langle u, zu\rangle = \overline{z}\langle u, u\rangle.$$

Thus,

$$z\langle u, u\rangle = \overline{z}\langle u, u\rangle,$$

which implies that $z = \overline{z}$, since $u \neq 0$, and z is indeed real. $\square$

There is also a version of Proposition 17.5 for a (real) Euclidean space E and a self-adjoint map $f\colon E \to E$ since every real vector space E can be embedded into a complex vector space $E_{\mathbb{C}}$, and every linear map $f\colon E \to E$ can be extended to a linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$.

Definition 17.2. Given a real vector space E, let $E_{\mathbb{C}}$ be the structure $E \times E$ under the addition operation

$$(u_1,\, u_2) + (v_1,\, v_2) = (u_1 + v_1,\, u_2 + v_2),$$

and let multiplication by a complex scalar $z = x + iy$ be defined such that

$$(x + iy) \cdot (u,\, v) = (xu - yv,\, yu + xv).$$

The space $E_{\mathbb{C}}$ is called the *complexification* of E.

It is easily shown that the structure $E_{\mathbb{C}}$ is a complex vector space. It is also immediate that

$$(0, v) = i(v, 0),$$

and thus, identifying E with the subspace of $E_{\mathbb{C}}$ consisting of all vectors of the form $(u, 0)$, we can write

$$(u, v) = u + iv.$$

Observe that if $(e_1, \ldots, e_n)$ is a basis of E (a real vector space), then $(e_1, \ldots, e_n)$ is also a basis of $E_{\mathbb{C}}$ (recall that e_i is an abbreviation for $(e_i, 0)$).

A linear map $f\colon E \to E$ is extended to the linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ defined such that

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v).$$

For any basis $(e_1, \ldots, e_n)$ of E, the matrix $M(f)$ representing f over $(e_1, \ldots, e_n)$ is *identical* to the matrix $M(f_{\mathbb{C}})$ representing $f_{\mathbb{C}}$ over $(e_1, \ldots, e_n)$, where we view $(e_1, \ldots, e_n)$ as a basis of $E_{\mathbb{C}}$. As a consequence, $\det(zI - M(f)) = \det(zI - M(f_{\mathbb{C}}))$, which means that f and $f_{\mathbb{C}}$ have the *same* characteristic polynomial (which has real coefficients). We know that every polynomial of degree n with real (or complex) coefficients always has n complex roots (counted with their multiplicity), and the roots of $\det(zI - M(f_{\mathbb{C}}))$ that are real (if any) are the eigenvalues of f.

Next we need to extend the inner product on E to an inner product on $E_{\mathbb{C}}$.

The inner product $\langle -, - \rangle$ on a Euclidean space E is extended to the Hermitian positive definite form $\langle -, - \rangle_{\mathbb{C}}$ on $E_{\mathbb{C}}$ as follows:

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle_{\mathbb{C}} = \langle u_1, u_2 \rangle + \langle v_1, v_2 \rangle + i(\langle v_1, u_2 \rangle - \langle u_1, v_2 \rangle).$$

It is easily verified that $\langle -, - \rangle_{\mathbb{C}}$ is indeed a Hermitian form that is positive definite, and it is clear that $\langle -, - \rangle_{\mathbb{C}}$ agrees with $\langle -, - \rangle$ on real vectors. Then given any linear map $f\colon E \to E$, it is easily verified that the map $f_{\mathbb{C}}^*$ defined such that

$$f_{\mathbb{C}}^*(u + iv) = f^*(u) + if^*(v)$$

for all $u, v \in E$ is the adjoint of $f_{\mathbb{C}}$ w.r.t. $\langle -, - \rangle_{\mathbb{C}}$.

Proposition 17.6. *Given a Euclidean space E, if $f\colon E \to E$ is any self-adjoint linear map, then every eigenvalue λ of $f_{\mathbb{C}}$ is real and is actually an eigenvalue of f (which means that there is some real eigenvector $u \in E$ such that $f(u) = \lambda u$). Therefore, all the eigenvalues of f are real.*

Proof. Let $E_{\mathbb{C}}$ be the complexification of E, $\langle -, - \rangle_{\mathbb{C}}$ the complexification of the inner product $\langle -, - \rangle$ on E, and $f_{\mathbb{C}} \colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ the complexification of $f \colon E \to E$. By definition of $f_{\mathbb{C}}$ and $\langle -, - \rangle_{\mathbb{C}}$, if f is self-adjoint, we have

$$\begin{aligned}
\langle f_{\mathbb{C}}(u_1 + iv_1), u_2 + iv_2 \rangle_{\mathbb{C}} &= \langle f(u_1) + if(v_1), u_2 + iv_2 \rangle_{\mathbb{C}} \\
&= \langle f(u_1), u_2 \rangle + \langle f(v_1), v_2 \rangle + i(\langle u_2, f(v_1) \rangle - \langle f(u_1), v_2 \rangle) \\
&= \langle u_1, f(u_2) \rangle + \langle v_1, f(v_2) \rangle + i(\langle f(u_2), v_1 \rangle - \langle u_1, f(v_2) \rangle) \\
&= \langle u_1 + iv_1, f(u_2) + if(v_2) \rangle_{\mathbb{C}} \\
&= \langle u_1 + iv_1, f_{\mathbb{C}}(u_2 + iv_2) \rangle_{\mathbb{C}},
\end{aligned}$$

which shows that $f_{\mathbb{C}}$ is also self-adjoint with respect to $\langle -, - \rangle_{\mathbb{C}}$.

As we pointed out earlier, f and $f_{\mathbb{C}}$ have the same characteristic polynomial $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$, which is a polynomial with real coefficients. Proposition 17.5 shows that the zeros of $\det(zI - f_{\mathbb{C}}) = \det(zI - f)$ are all real, and for each real zero λ of $\det(zI - f)$, the linear map $\lambda \mathrm{id} - f$ is singular, which means that there is some nonzero $u \in E$ such that $f(u) = \lambda u$. Therefore, all the eigenvalues of f are real. $\square$

Proposition 17.7. *Given a Hermitian space E, for any linear map $f \colon E \to E$, if f is skew-self-adjoint, then f has eigenvalues that are pure imaginary or zero, and if f is unitary, then f has eigenvalues of absolute value* 1.

Proof. If f is skew-self-adjoint, $f^* = -f$, and then by the definition of the adjoint map, for any eigenvalue λ and any eigenvector u associated with λ, we have

$$\lambda \langle u, u \rangle = \langle \lambda u, u \rangle = \langle f(u), u \rangle = \langle u, f^*(u) \rangle = \langle u, -f(u) \rangle = -\langle u, \lambda u \rangle = -\overline{\lambda} \langle u, u \rangle,$$

and since $u \neq 0$ and $\langle -, - \rangle$ is positive definite, $\langle u, u \rangle \neq 0$, so

$$\lambda = -\overline{\lambda},$$

which shows that $\lambda = ir$ for some $r \in \mathbb{R}$.

If f is unitary, then f is an isometry, so for any eigenvalue λ and any eigenvector u associated with λ, we have

$$|\lambda|^2 \langle u, u \rangle = \lambda \overline{\lambda} \langle u, u \rangle = \langle \lambda u, \lambda u \rangle = \langle f(u), f(u) \rangle = \langle u, u \rangle,$$

and since $u \neq 0$, we obtain $|\lambda|^2 = 1$, which implies

$$|\lambda| = 1. \qquad \square$$

17.3 Spectral Theorem for Normal Linear Maps

Given a Euclidean space E, our next step is to show that for every linear map $f\colon E \to E$ there is some subspace W of dimension 1 or 2 such that $f(W) \subseteq W$. When $\dim(W) = 1$, the subspace W is actually an eigenspace for some real eigenvalue of f. Furthermore, when f is normal, there is a subspace W of dimension 1 or 2 such that $f(W) \subseteq W$ **and** $f^*(W) \subseteq W$. The difficulty is that the eigenvalues of f are not necessarily real. One way to get around this problem is to complexify both the vector space E and the inner product $\langle -, - \rangle$ as we did in Section 17.2.

Given any subspace W of a Euclidean space E, recall that the *orthogonal complement* $W^\perp$ of W is the subspace defined such that

$$W^\perp = \{u \in E \mid \langle u, w \rangle = 0, \text{ for all } w \in W\}.$$

Recall from Proposition 12.11 that $E = W \oplus W^\perp$ (this can be easily shown, for example, by constructing an orthonormal basis of E using the Gram–Schmidt orthonormalization procedure). The same result also holds for Hermitian spaces; see Proposition 14.13.

As a warm up for the proof of Theorem 17.12, let us prove that every self-adjoint map on a Euclidean space can be diagonalized with respect to an orthonormal basis of eigenvectors.

Theorem 17.8. *(Spectral theorem for self-adjoint linear maps on a Euclidean space) Given a Euclidean space E of dimension n, for every self-adjoint linear map $f\colon E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \ldots & \\ & \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \lambda_n \end{pmatrix},$$

with $\lambda_i \in \mathbb{R}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. From Proposition 17.6, all the eigenvalues of f are real, so pick some eigenvalue $\lambda \in \mathbb{R}$, and let w be some eigenvector for λ. By dividing w by its norm, we may assume that w is a unit vector. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. We claim that $f(W^\perp) \subseteq W^\perp$, where $W^\perp$ is the orthogonal complement of W.

Indeed, for any $v \in W^\perp$, that is, if $\langle v, w \rangle = 0$, because f is self-adjoint and $f(w) = \lambda w$, we have

$$\begin{aligned} \langle f(v), w \rangle &= \langle v, f(w) \rangle \\ &= \langle v, \lambda w \rangle \\ &= \lambda \langle v, w \rangle = 0 \end{aligned}$$

since $\langle v, w \rangle = 0$. Therefore,

$$f(W^{\perp}) \subseteq W^{\perp}.$$

Clearly, the restriction of f to $W^{\perp}$ is self-adjoint, and we conclude by applying the induction hypothesis to $W^{\perp}$ (whose dimension is $n-1$). $\square$

We now come back to normal linear maps. One of the key points in the proof of Theorem 17.8 is that we found a subspace W with the property that $f(W) \subseteq W$ implies that $f(W^{\perp}) \subseteq W^{\perp}$. In general, this does not happen, *but normal maps satisfy a stronger property which ensures that such a subspace exists*.

The following proposition provides a condition that will allow us to show that a normal linear map can be diagonalized. It actually holds for any linear map. We found the inspiration for this proposition in Berger [11].

Proposition 17.9. *Given a Hermitian space E, for any linear map $f\colon E \to E$ and any subspace W of E, if $f(W) \subseteq W$, then $f^*\big(W^{\perp}\big) \subseteq W^{\perp}$. Consequently, if $f(W) \subseteq W$ and $f^*(W) \subseteq W$, then $f\big(W^{\perp}\big) \subseteq W^{\perp}$ and $f^*\big(W^{\perp}\big) \subseteq W^{\perp}$.*

Proof. If $u \in W^{\perp}$, then

$$\langle w, u \rangle = 0 \quad \text{for all } w \in W.$$

However,

$$\langle f(w), u \rangle = \langle w, f^*(u) \rangle,$$

and $f(W) \subseteq W$ implies that $f(w) \in W$. Since $u \in W^{\perp}$, we get

$$0 = \langle f(w), u \rangle = \langle w, f^*(u) \rangle,$$

which shows that $\langle w, f^*(u) \rangle = 0$ for all $w \in W$, that is, $f^*(u) \in W^{\perp}$. Therefore, we have $f^*(W^{\perp}) \subseteq W^{\perp}$.

We just proved that if $f(W) \subseteq W$, then $f^*\big(W^{\perp}\big) \subseteq W^{\perp}$. If we also have $f^*(W) \subseteq W$, then by applying the above fact to f^*, we get $f^{**}(W^{\perp}) \subseteq W^{\perp}$, and since $f^{**} = f$, this is just $f(W^{\perp}) \subseteq W^{\perp}$, which proves the second statement of the proposition. $\square$

It is clear that the above proposition also holds for Euclidean spaces.

Although we are ready to prove that for every normal linear map f (over a Hermitian space) there is an orthonormal basis of eigenvectors (see Theorem 17.13 below), we now return to real Euclidean spaces.

Proposition 17.10. *If $f\colon E \to E$ is a linear map and $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ for the eigenvalue $z = \lambda + i\mu$, where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$, then*

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v. \tag{$*$}$$

As a consequence,

$$f_{\mathbb{C}}(u - iv) = f(u) - if(v) = (\lambda - i\mu)(u - iv),$$

which shows that $\overline{w} = u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\overline{z} = \lambda - i\mu$.

Proof. Since
$$f_{\mathbb{C}}(u+iv) = f(u) + if(v)$$
and
$$f_{\mathbb{C}}(u+iv) = (\lambda + i\mu)(u+iv) = \lambda u - \mu v + i(\mu u + \lambda v),$$
we have
$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v. \qquad \square$$

Using this fact, we can prove the following proposition.

Proposition 17.11. *Given a Euclidean space E, for any normal linear map $f\colon E \to E$, if $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$ associated with the eigenvalue $z = \lambda + i\mu$ (where $u, v \in E$ and $\lambda, \mu \in \mathbb{R}$), if $\mu \neq 0$ (i.e., z is not real) then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, which implies that u and v are linearly independent, and if W is the subspace spanned by u and v, then $f(W) = W$ and $f^*(W) = W$. Furthermore, with respect to the (orthogonal) basis (u, v), the restriction of f to W has the matrix*
$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$
If $\mu = 0$, then λ is a real eigenvalue of f, and either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by $v \neq 0$ if $u = 0$, then $f(W) \subseteq W$ and $f^*(W) \subseteq W$.

Proof. Since $w = u + iv$ is an eigenvector of $f_{\mathbb{C}}$, by definition it is nonnull, and either $u \neq 0$ or $v \neq 0$. Proposition 17.10 implies that $u - iv$ is an eigenvector of $f_{\mathbb{C}}$ for $\lambda - i\mu$. It is easy to check that $f_{\mathbb{C}}$ is normal. However, if $\mu \neq 0$, then $\lambda + i\mu \neq \lambda - i\mu$, and from Proposition 17.4, the vectors $u + iv$ and $u - iv$ are orthogonal w.r.t. $\langle -, -\rangle_{\mathbb{C}}$, that is,
$$\langle u + iv, u - iv\rangle_{\mathbb{C}} = \langle u, u\rangle - \langle v, v\rangle + 2i\langle u, v\rangle = 0.$$
Thus we get $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, and since $u \neq 0$ or $v \neq 0$, u and v are linearly independent. Since
$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v$$
and since by Proposition 17.3 $u + iv$ is an eigenvector of $f_{\mathbb{C}}^*$ for $\lambda - i\mu$, we have
$$f^*(u) = \lambda u + \mu v \quad \text{and} \quad f^*(v) = -\mu u + \lambda v,$$
and thus $f(W) = W$ and $f^*(W) = W$, where W is the subspace spanned by u and v.

When $\mu = 0$, we have
$$f(u) = \lambda u \quad \text{and} \quad f(v) = \lambda v,$$
and since $u \neq 0$ or $v \neq 0$, either u or v is an eigenvector of f for λ. If W is the subspace spanned by u if $u \neq 0$, or spanned by v if $u = 0$, it is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. Note that $\lambda = 0$ is possible, and this is why $\subseteq$ cannot be replaced by $=$. $\square$

The beginning of the proof of Proposition 17.11 actually shows that for every linear map $f\colon E \to E$ there is some subspace W such that $f(W) \subseteq W$, where W has dimension 1 or 2. In general, it doesn't seem possible to prove that $W^{\perp}$ is invariant under f. *However, this happens when f is normal.*

We can finally prove our first main theorem.

Theorem 17.12. *(Main spectral theorem) Given a Euclidean space E of dimension n, for every normal linear map $f\colon E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \ldots & \\ & A_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & A_p \end{pmatrix}$$

such that each block A_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. First, since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$ (where $\lambda, \mu \in \mathbb{R}$). Let $w = u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$ (where $u, v \in E$). We can now apply Proposition 17.11.

If $\mu = 0$, then either u or v is an eigenvector of f for $\lambda \in \mathbb{R}$. Let W be the subspace of dimension 1 spanned by $e_1 = u/\|u\|$ if $u \neq 0$, or by $e_1 = v/\|v\|$ otherwise. It is obvious that $f(W) \subseteq W$ and $f^*(W) \subseteq W$. The orthogonal $W^{\perp}$ of W has dimension $n - 1$, and by Proposition 17.9, we have $f\big(W^{\perp}\big) \subseteq W^{\perp}$. But the restriction of f to $W^{\perp}$ is also normal, and we conclude by applying the induction hypothesis to $W^{\perp}$.

If $\mu \neq 0$, then $\langle u, v\rangle = 0$ and $\langle u, u\rangle = \langle v, v\rangle$, and if W is the subspace spanned by $u/\|u\|$ and $v/\|v\|$, then $f(W) = W$ and $f^*(W) = W$. We also know that the restriction of f to W has the matrix

$$\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}$$

with respect to the basis $(u/\|u\|, v/\|v\|)$. If $\mu < 0$, we let $\lambda_1 = \lambda$, $\mu_1 = -\mu$, $e_1 = u/\|u\|$, and $e_2 = v/\|v\|$. If $\mu > 0$, we let $\lambda_1 = \lambda$, $\mu_1 = \mu$, $e_1 = v/\|v\|$, and $e_2 = u/\|u\|$. In all cases, it is easily verified that the matrix of the restriction of f to W w.r.t. the orthonormal basis (e_1, e_2) is

$$A_1 = \begin{pmatrix} \lambda_1 & -\mu_1 \\ \mu_1 & \lambda_1 \end{pmatrix},$$

where $\lambda_1, \mu_1 \in \mathbb{R}$, with $\mu_1 > 0$. However, $W^\perp$ has dimension $n-2$, and by Proposition 17.9, $f\big(W^\perp\big) \subseteq W^\perp$. Since the restriction of f to $W^\perp$ is also normal, we conclude by applying the induction hypothesis to $W^\perp$. $\square$

After this relatively hard work, we can easily obtain some nice normal forms for the matrices of self-adjoint, skew-self-adjoint, and orthogonal linear maps. However, for the sake of completeness (and since we have all the tools to so do), we go back to the case of a Hermitian space and show that normal linear maps can be diagonalized with respect to an orthonormal basis. The proof is a slight generalization of the proof of Theorem 17.6.

Theorem 17.13. *(Spectral theorem for normal linear maps on a Hermitian space) Given a Hermitian space E of dimension n, for every normal linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \ldots & \\ & \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \lambda_n \end{pmatrix},$$

where $\lambda_j \in \mathbb{C}$.

Proof. We proceed by induction on the dimension n of E as follows. If $n = 1$, the result is trivial. Assume now that $n \geq 2$. Since $\mathbb{C}$ is algebraically closed (i.e., every polynomial has a root in $\mathbb{C}$), the linear map $f\colon E \to E$ has some eigenvalue $\lambda \in \mathbb{C}$, and let w be some unit eigenvector for λ. Let W be the subspace of dimension 1 spanned by w. Clearly, $f(W) \subseteq W$. By Proposition 17.3, w is an eigenvector of f^* for $\overline{\lambda}$, and thus $f^*(W) \subseteq W$. By Proposition 17.9, we also have $f(W^\perp) \subseteq W^\perp$. The restriction of f to $W^\perp$ is still normal, and we conclude by applying the induction hypothesis to $W^\perp$ (whose dimension is $n-1$). $\square$

Theorem 17.13 implies that (complex) self-adjoint, skew-self-adjoint, and orthogonal linear maps can be diagonalized with respect to an orthonormal basis of eigenvectors. In this latter case, though, an orthogonal map is called a *unitary* map. Proposition 17.5 also shows that the eigenvalues of a self-adjoint linear map are real, and Proposition 17.7 shows that the eigenvalues of a skew self-adjoint map are pure imaginary or zero, and that the eigenvalues of a unitary map have absolute value 1.

Remark: There is a converse to Theorem 17.13, namely, if there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f, then f is normal. We leave the easy proof as an exercise.

In the next section we specialize Theorem 17.12 to self-adjoint, skew-self-adjoint, and orthogonal linear maps. Due to the additional structure, we obtain more precise normal forms.

17.4 Self-Adjoint, Skew-Self-Adjoint, and Orthogonal Linear Maps

We begin with self-adjoint maps.

Theorem 17.14. *Given a Euclidean space E of dimension n, for every self-adjoint linear map $f\colon E \to E$, there is an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors of f such that the matrix of f w.r.t. this basis is a diagonal matrix*

$$\begin{pmatrix} \lambda_1 & & \ldots & \\ & \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Proof. We already proved this; see Theorem 17.8. However, it is instructive to give a more direct method not involving the complexification of $\langle -, - \rangle$ and Proposition 17.5.

Since $\mathbb{C}$ is algebraically closed, $f_{\mathbb{C}}$ has some eigenvalue $\lambda + i\mu$, and let $u + iv$ be some eigenvector of $f_{\mathbb{C}}$ for $\lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$ and $u, v \in E$. We saw in the proof of Proposition 17.10 that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v.$$

Since $f = f^*$,

$$\langle f(u), v\rangle = \langle u, f(v)\rangle$$

for all $u, v \in E$. Applying this to

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$\langle f(u), v\rangle = \langle \lambda u - \mu v, v\rangle = \lambda\langle u, v\rangle - \mu\langle v, v\rangle$$

and

$$\langle u, f(v)\rangle = \langle u, \mu u + \lambda v\rangle = \mu\langle u, u\rangle + \lambda\langle u, v\rangle,$$

and thus we get

$$\lambda\langle u, v\rangle - \mu\langle v, v\rangle = \mu\langle u, u\rangle + \lambda\langle u, v\rangle,$$

that is,

$$\mu(\langle u, u\rangle + \langle v, v\rangle) = 0,$$

which implies $\mu = 0$, since either $u \neq 0$ or $v \neq 0$. Therefore, λ is a real eigenvalue of f.

Now going back to the proof of Theorem 17.12, only the case where $\mu = 0$ applies, and the induction shows that all the blocks are one-dimensional. $\square$

Theorem 17.14 implies that if $\lambda_1, \ldots, \lambda_p$ are the distinct real eigenvalues of f, and E_i is the eigenspace associated with λ_i, then

$$E = E_1 \oplus \cdots \oplus E_p,$$

where E_i and E_j are orthogonal for all $i \neq j$.

Remark: Another way to prove that a self-adjoint map has a real eigenvalue is to use a little bit of calculus. We learned such a proof from Herman Gluck. The idea is to consider the real-valued function $\Phi\colon E \to \mathbb{R}$ defined such that

$$\Phi(u) = \langle f(u), u \rangle$$

for every $u \in E$. This function is C^∞, and if we represent f by a matrix A over some orthonormal basis, it is easy to compute the gradient vector

$$\nabla\Phi(X) = \left(\frac{\partial\Phi}{\partial x_1}(X), \ldots, \frac{\partial\Phi}{\partial x_n}(X)\right)$$

of Φ at X. Indeed, we find that

$$\nabla\Phi(X) = (A + A^\top)X,$$

where X is a column vector of size n. But since f is self-adjoint, $A = A^\top$, and thus

$$\nabla\Phi(X) = 2AX.$$

The next step is to find the maximum of the function Φ on the sphere

$$S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

Since S^{n-1} is compact and Φ is continuous, and in fact C^∞, Φ takes a maximum at some X on S^{n-1}. But then it is well known that at an extremum X of Φ we must have

$$d\Phi_X(Y) = \langle \nabla\Phi(X), Y \rangle = 0$$

for all tangent vectors Y to S^{n-1} at X, and so $\nabla\Phi(X)$ is orthogonal to the tangent plane at X, which means that

$$\nabla\Phi(X) = \lambda X$$

for some $\lambda \in \mathbb{R}$. Since $\nabla\Phi(X) = 2AX$, we get

$$2AX = \lambda X,$$

and thus $\lambda/2$ is a real eigenvalue of A (i.e., of f).

Next we consider skew-self-adjoint maps.

Theorem 17.15. *Given a Euclidean space E of dimension n, for every skew-self-adjoint linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \ldots & \\ & A_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & A_p \end{pmatrix}$$

such that each block A_j is either 0 or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary of the form $\pm i\mu_j$ or 0.

Proof. The case where $n = 1$ is trivial. As in the proof of Theorem 17.12, $f_{\mathbb{C}}$ has some eigenvalue $z = \lambda + i\mu$, where $\lambda, \mu \in \mathbb{R}$. We claim that $\lambda = 0$. First we show that

$$\langle f(w), w\rangle = 0$$

for all $w \in E$. Indeed, since $f = -f^*$, we get

$$\langle f(w), w\rangle = \langle w, f^*(w)\rangle = \langle w, -f(w)\rangle = -\langle w, f(w)\rangle = -\langle f(w), w\rangle,$$

since $\langle -, -\rangle$ is symmetric. This implies that

$$\langle f(w), w\rangle = 0.$$

Applying this to u and v and using the fact that

$$f(u) = \lambda u - \mu v \quad \text{and} \quad f(v) = \mu u + \lambda v,$$

we get

$$0 = \langle f(u), u\rangle = \langle \lambda u - \mu v, u\rangle = \lambda\langle u, u\rangle - \mu\langle u, v\rangle$$

and

$$0 = \langle f(v), v\rangle = \langle \mu u + \lambda v, v\rangle = \mu\langle u, v\rangle + \lambda\langle v, v\rangle,$$

from which, by addition, we get

$$\lambda(\langle v, v\rangle + \langle v, v\rangle) = 0.$$

Since $u \neq 0$ or $v \neq 0$, we have $\lambda = 0$.

Then going back to the proof of Theorem 17.12, unless $\mu = 0$, the case where u and v are orthogonal and span a subspace of dimension 2 applies, and the induction shows that all the blocks are two-dimensional or reduced to 0. $\square$

Remark: One will note that if f is skew-self-adjoint, then $if_{\mathbb{C}}$ is self-adjoint w.r.t. $\langle -, -\rangle_{\mathbb{C}}$. By Proposition 17.5, the map $if_{\mathbb{C}}$ has real eigenvalues, which implies that the eigenvalues of $f_{\mathbb{C}}$ are pure imaginary or 0.

Finally we consider orthogonal linear maps.

Theorem 17.16. *Given a Euclidean space E of dimension n, for every orthogonal linear map $f\colon E \to E$ there is an orthonormal basis $(e_1, \ldots, e_n)$ such that the matrix of f w.r.t. this basis is a block diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & \ldots & \\ & A_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & A_p \end{pmatrix}$$

such that each block A_j is either 1, -1, or a two-dimensional matrix of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. In particular, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta_j \pm i\sin\theta_j$, 1, or -1.

Proof. The case where $n = 1$ is trivial. It is immediately verified that $f \circ f^* = f^* \circ f = \mathrm{id}$ implies that $f_{\mathbb{C}} \circ f_{\mathbb{C}}^* = f_{\mathbb{C}}^* \circ f_{\mathbb{C}} = \mathrm{id}$, so the map $f_{\mathbb{C}}$ is unitary. By Proposition 17.7, the eigenvalues of $f_{\mathbb{C}}$ have absolute value 1. As a consequence, the eigenvalues of $f_{\mathbb{C}}$ are of the form $\cos\theta \pm i\sin\theta$, 1, or -1. The theorem then follows immediately from Theorem 17.12, where the condition $\mu > 0$ implies that $\sin\theta_j > 0$, and thus, $0 < \theta_j < \pi$. ∎

It is obvious that we can reorder the orthonormal basis of eigenvectors given by Theorem 17.16, so that the matrix of f w.r.t. this basis is a block diagonal matrix of the form

$$\begin{pmatrix} A_1 & \ldots & & & \\ \vdots & \ddots & \vdots & & \vdots \\ & \ldots & A_r & & \\ & & & -I_q & \\ \ldots & & & & I_p \end{pmatrix}$$

where each block A_j is a two-dimensional rotation matrix $A_j \neq \pm I_2$ of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j < \pi$.

The linear map f has an eigenspace $E(1, f) = \mathrm{Ker}\,(f - \mathrm{id})$ of dimension p for the eigenvalue 1, and an eigenspace $E(-1, f) = \mathrm{Ker}\,(f + \mathrm{id})$ of dimension q for the eigenvalue -1. If $\det(f) = +1$ (f is a rotation), the dimension q of $E(-1, f)$ must be even, and the entries in $-I_q$ can be paired to form two-dimensional blocks, if we wish. In this case, every rotation in $\mathbf{SO}(n)$ has a matrix of the form

$$\begin{pmatrix} A_1 & \ldots & & \\ \vdots & \ddots & \vdots & \\ & \ldots & A_m & \\ \ldots & & & I_{n-2m} \end{pmatrix}$$

where the first m blocks A_j are of the form

$$A_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

with $0 < \theta_j \leq \pi$.

Theorem 17.16 can be used to prove a version of the Cartan–Dieudonné theorem.

Theorem 17.17. *Let E be a Euclidean space of dimension $n \geq 2$. For every isometry $f \in \mathbf{O}(E)$, if $p = \dim(E(1, f)) = \dim(\mathrm{Ker}\,(f - \mathrm{id}))$, then f is the composition of $n - p$ reflections, and $n - p$ is minimal.*

Proof. From Theorem 17.16 there are r subspaces $F_1, \ldots, F_r$, each of dimension 2, such that

$$E = E(1, f) \oplus E(-1, f) \oplus F_1 \oplus \cdots \oplus F_r,$$

and all the summands are pairwise orthogonal. Furthermore, the restriction r_i of f to each F_i is a rotation $r_i \neq \pm\mathrm{id}$. Each 2D rotation r_i can be written as the composition $r_i = s'_i \circ s_i$ of two reflections s_i and s'_i about lines in F_i (forming an angle $\theta_i/2$). We can extend s_i and s'_i to hyperplane reflections in E by making them the identity on $F_i^{\perp}$. Then

$$s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1$$

agrees with f on $F_1 \oplus \cdots \oplus F_r$ and is the identity on $E(1, f) \oplus E(-1, f)$. If $E(-1, f)$ has an orthonormal basis of eigenvectors $(v_1, \ldots, v_q)$, letting s''_j be the reflection about the hyperplane $(v_j)^{\perp}$, it is clear that

$$s''_q \circ \cdots \circ s''_1$$

agrees with f on $E(-1, f)$ and is the identity on $E(1, f) \oplus F_1 \oplus \cdots \oplus F_r$. But then

$$f = s''_q \circ \cdots \circ s''_1 \circ s'_r \circ s_r \circ \cdots \circ s'_1 \circ s_1,$$

the composition of $2r + q = n - p$ reflections.

If

$$f = s_t \circ \cdots \circ s_1,$$

for t reflections s_i, it is clear that

$$F = \bigcap_{i=1}^{t} E(1, s_i) \subseteq E(1, f),$$

where $E(1, s_i)$ is the hyperplane defining the reflection s_i. By the Grassmann relation, if we intersect $t \leq n$ hyperplanes, the dimension of their intersection is at least $n - t$. Thus, $n - t \leq p$, that is, $t \geq n - p$, and $n - p$ is the smallest number of reflections composing f. $\square$

As a corollary of Theorem 17.17, we obtain the following fact: If the dimension n of the Euclidean space E is odd, then every rotation $f \in \mathbf{SO}(E)$ admits 1 as an eigenvalue.

Proof. The characteristic polynomial $\det(XI - f)$ of f has odd degree n and has real coefficients, so it must have some real root λ. Since f is an isometry, its n eigenvalues are of the form, $+1, -1$, and $e^{\pm i\theta}$, with $0 < \theta < \pi$, so $\lambda = \pm 1$. Now the eigenvalues $e^{\pm i\theta}$ appear in conjugate pairs, and since n is odd, the number of real eigenvalues of f is odd. This implies that $+1$ is an eigenvalue of f, since otherwise -1 would be the only real eigenvalue of f, and since its multiplicity is odd, we would have $\det(f) = -1$, contradicting the fact that f is a rotation. $\square$

When $n = 3$, we obtain the result due to Euler which says that every 3D rotation R has an invariant axis D, and that restricted to the plane orthogonal to D, it is a 2D rotation. Furthermore, if (a, b, c) is a unit vector defining the axis D of the rotation R and if the angle of the rotation is θ, if B is the skew-symmetric matrix

$$B = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

then the Rodigues formula (Proposition 12.15) states that

$$R = I + \sin\theta B + (1 - \cos\theta) B^2.$$

The theorems of this section and of the previous section can be immediately translated in terms of matrices. The matrix versions of these theorems is often used in applications so we briefly present them in the section.

17.5 Normal and Other Special Matrices

First we consider real matrices. Recall the following definitions.

Definition 17.3. Given a real $m \times n$ matrix A, the *transpose $A^\top$ of A* is the $n \times m$ matrix $A^\top = (a^\top_{i\,j})$ defined such that

$$a^\top_{i\,j} = a_{j\,i}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. A real $n \times n$ matrix A is

- *normal* if
$$A\,A^\top = A^\top A,$$

- *symmetric* if
$$A^\top = A,$$

- *skew-symmetric* if
$$A^\top = -A,$$

- *orthogonal* if
$$A\,A^\top = A^\top A = I_n.$$

Recall from Proposition 12.14 that when E is a Euclidean space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f\colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then $A^\top$ is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a symmetric matrix, a skew-self-adjoint linear map has a skew-symmetric matrix, and an orthogonal linear map has an orthogonal matrix.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is orthogonal, and for any linear map $f\colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^\top AP.$$

As a consequence, Theorems 17.12 and 17.14–17.16 can be restated as follows.

Theorem 17.18. *For every normal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \ldots & \\ & D_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & D_p \end{pmatrix}$$

such that each block D_j is either a one-dimensional matrix (i.e., a real scalar) or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix},$$

where $\lambda_j, \mu_j \in \mathbb{R}$, with $\mu_j > 0$.

Theorem 17.19. *For every symmetric matrix A there is an orthogonal matrix P and a diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} \lambda_1 & & \ldots & \\ & \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \lambda_n \end{pmatrix},$$

where $\lambda_i \in \mathbb{R}$.

Theorem 17.20. *For every skew-symmetric matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \ldots & \\ & D_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & D_p \end{pmatrix}$$

such that each block D_j is either 0 or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} 0 & -\mu_j \\ \mu_j & 0 \end{pmatrix},$$

where $\mu_j \in \mathbb{R}$, with $\mu_j > 0$. In particular, the eigenvalues of A are pure imaginary of the form $\pm i\mu_j$, or 0.

Theorem 17.21. *For every orthogonal matrix A there is an orthogonal matrix P and a block diagonal matrix D such that $A = PD\,P^\top$, where D is of the form*

$$D = \begin{pmatrix} D_1 & & \ldots & \\ & D_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & D_p \end{pmatrix}$$

such that each block D_j is either 1, -1, or a two-dimensional matrix of the form

$$D_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}$$

where $0 < \theta_j < \pi$. In particular, the eigenvalues of A are of the form $\cos\theta_j \pm i\sin\theta_j$, 1, or -1.

Theorem 17.21 can be used to show that the exponential map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$ is surjective; see Gallier [72].

We now consider complex matrices.

Definition 17.4. Given a complex $m \times n$ matrix A, the *transpose $A^\top$ of A* is the $n \times m$ matrix $A^\top = \left(a^\top_{i\,j}\right)$ defined such that

$$a^\top_{i\,j} = a_{j\,i}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. The *conjugate $\overline{A}$ of A* is the $m \times n$ matrix $\overline{A} = (b_{i\,j})$ defined such that

$$b_{i\,j} = \overline{a_{i\,j}}$$

for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$. Given an $m \times n$ complex matrix A, the *adjoint A^* of A* is the matrix defined such that

$$A^* = \overline{(A^\top)} = (\overline{A})^\top.$$

A complex $n \times n$ matrix A is

- *normal* if
$$AA^* = A^*A,$$

- *Hermitian* if
$$A^* = A,$$

- *skew-Hermitian* if
$$A^* = -A,$$

- *unitary* if
$$AA^* = A^*A = I_n.$$

Recall from Proposition 14.15 that when E is a Hermitian space and $(e_1, \ldots, e_n)$ is an orthonormal basis for E, if A is the matrix of a linear map $f \colon E \to E$ w.r.t. the basis $(e_1, \ldots, e_n)$, then A^* is the matrix of the adjoint f^* of f. Consequently, a normal linear map has a normal matrix, a self-adjoint linear map has a Hermitian matrix, a skew-self-adjoint linear map has a skew-Hermitian matrix, and a unitary linear map has a unitary matrix.

Furthermore, if $(u_1, \ldots, u_n)$ is another orthonormal basis for E and P is the change of basis matrix whose columns are the components of the u_i w.r.t. the basis $(e_1, \ldots, e_n)$, then P is unitary, and for any linear map $f \colon E \to E$, if A is the matrix of f w.r.t $(e_1, \ldots, e_n)$ and B is the matrix of f w.r.t. $(u_1, \ldots, u_n)$, then

$$B = P^*AP.$$

Theorem 17.13 and Proposition 17.7 can be restated in terms of matrices as follows.

Theorem 17.22. *For every complex normal matrix A there is a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. Furthermore, if A is Hermitian, then D is a real matrix; if A is skew-Hermitian, then the entries in D are pure imaginary or zero; and if A is unitary, then the entries in D have absolute value* 1.

17.6 Rayleigh–Ritz Theorems and Eigenvalue Interlacing

A fact that is used frequently in optimization problems is that the eigenvalues of a symmetric matrix are characterized in terms of what is known as the *Rayleigh ratio*, defined by

$$R(A)(x) = \frac{x^\top Ax}{x^\top x}, \quad x \in \mathbb{R}^n, x \neq 0.$$

The following proposition is often used to prove the correctness of various optimization or approximation problems (for example PCA; see Section 23.4). It is also used to prove Proposition 17.25, which is used to justify the correctness of a method for graph-drawing (see Chapter 21).

Proposition 17.23. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_n$$

(with the maximum attained for $x = u_n$), and

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top Ax}{x^\top x} = \lambda_{n-k}$$

(with the maximum attained for $x = u_{n-k}$), where $1 \leq k \leq n-1$. Equivalently, if V_k is the subspace spanned by $(u_1, \ldots, u_k)$, then

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top A x}{x^\top x}, \quad k = 1, \ldots, n.$$

Proof. First observe that

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \max_x \{x^\top A x \mid x^\top x = 1\},$$

and similarly,

$$\max_{x \neq 0, x \in \{u_{n-k+1}, \ldots, u_n\}^\perp} \frac{x^\top A x}{x^\top x} = \max_x \left\{x^\top A x \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1, \ldots, u_n)$ be such a basis. If we write

$$x = \sum_{i=1}^n x_i u_i,$$

a simple computation shows that

$$x^\top A x = \sum_{i=1}^n \lambda_i x_i^2.$$

If $x^\top x = 1$, then $\sum_{i=1}^n x_i^2 = 1$, and since we assumed that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we get

$$x^\top A x = \sum_{i=1}^n \lambda_i x_i^2 \leq \lambda_n \left(\sum_{i=1}^n x_i^2\right) = \lambda_n.$$

Thus,

$$\max_x \left\{x^\top A x \mid x^\top x = 1\right\} \leq \lambda_n,$$

and since this maximum is achieved for $e_n = (0, 0, \ldots, 1)$, we conclude that

$$\max_x \left\{x^\top A x \mid x^\top x = 1\right\} = \lambda_n.$$

Next observe that $x \in \{u_{n-k+1}, \ldots, u_n\}^\perp$ and $x^\top x = 1$ iff $x_{n-k+1} = \cdots = x_n = 0$ and $\sum_{i=1}^{n-k} x_i^2 = 1$. Consequently, for such an x, we have

$$x^\top A x = \sum_{i=1}^{n-k} \lambda_i x_i^2 \leq \lambda_{n-k} \left(\sum_{i=1}^{n-k} x_i^2\right) = \lambda_{n-k}.$$

Thus,

$$\max_x \left\{x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1)\right\} \leq \lambda_{n-k},$$

and since this maximum is achieved for $e_{n-k} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in position $n-k$, we conclude that

$$\max_x \left\{x^\top Ax \mid (x \in \{u_{n-k+1}, \ldots, u_n\}^\perp) \wedge (x^\top x = 1)\right\} = \lambda_{n-k},$$

as claimed. □

For our purposes we need the version of Proposition 17.23 applying to min instead of max, whose proof is obtained by a trivial modification of the proof of Proposition 17.23.

Proposition 17.24. *(Rayleigh–Ritz) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \ldots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\min_{x \neq 0} \frac{x^\top Ax}{x^\top x} = \lambda_1$$

(with the minimum attained for $x = u_1$), and

$$\min_{x \neq 0, x \in \{u_1, \ldots, u_{i-1}\}^\perp} \frac{x^\top Ax}{x^\top x} = \lambda_i$$

(with the minimum attained for $x = u_i$), where $2 \leq i \leq n$. Equivalently, if $W_k = V_{k-1}^\perp$ denotes the subspace spanned by $(u_k, \ldots, u_n)$ (with $V_0 = (0)$), then

$$\lambda_k = \min_{x \neq 0, x \in W_k} \frac{x^\top Ax}{x^\top x} = \min_{x \neq 0, x \in V_{k-1}^\perp} \frac{x^\top Ax}{x^\top x}, \quad k = 1, \ldots, n.$$

Propositions 17.23 and 17.24 together are known as the *Rayleigh–Ritz theorem*.

Observe that Proposition 17.24 immediately implies that if A is a symmetric matrix, then A is positive definite iff all its eigenvalues are positive. We also prove this fact in Section 22.1; see Proposition 22.3.

As an application of Propositions 17.23 and 17.24, we prove a proposition which allows us to compare the eigenvalues of two symmetric matrices A and $B = R^\top AR$, where R is a rectangular matrix satisfying the equation $R^\top R = I$.

First we need a definition.

Definition 17.5. Given an $n \times n$ symmetric matrix A and an $m \times m$ symmetric B, with $m \leq n$, if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, then we say that the eigenvalues of B *interlace* the eigenvalues of A if

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

For example, if $n = 5$ and $m = 3$, we have

$$\begin{aligned}
\lambda_1 \leq \mu_1 \leq \lambda_3 \\
\lambda_2 \leq \mu_2 \leq \lambda_4 \\
\lambda_3 \leq \mu_3 \leq \lambda_5.
\end{aligned}$$

Proposition 17.25. *Let A be an $n \times n$ symmetric matrix, R be an $n \times m$ matrix such that $R^\top R = I$ (with $m \leq n$), and let $B = R^\top AR$ (an $m \times m$ matrix). The following properties hold:*

(a) The eigenvalues of B interlace the eigenvalues of A.

(b) If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$ are the eigenvalues of B, and if $\lambda_i = \mu_i$, then there is an eigenvector v of B with eigenvalue μ_i such that Rv is an eigenvector of A with eigenvalue λ_i.

Proof. (a) Let $(u_1, \ldots, u_n)$ be an orthonormal basis of eigenvectors for A, and let $(v_1, \ldots, v_m)$ be an orthonormal basis of eigenvectors for B. Let U_j be the subspace spanned by $(u_1, \ldots, u_j)$ and let V_j be the subspace spanned by $(v_1, \ldots, v_j)$. For any i, the subspace V_i has dimension i and the subspace $R^\top U_{i-1}$ has dimension at most $i - 1$. Therefore, there is some nonzero vector $v \in V_i \cap (R^\top U_{i-1})^\perp$, and since

$$v^\top R^\top u_j = (Rv)^\top u_j = 0, \quad j = 1, \ldots, i - 1,$$

we have $Rv \in (U_{i-1})^\perp$. By Proposition 17.24 and using the fact that $R^\top R = I$, we have

$$\lambda_i \leq \frac{(Rv)^\top ARv}{(Rv)^\top Rv} = \frac{v^\top Bv}{v^\top v}.$$

On the other hand, by Proposition 17.23,

$$\mu_i = \max_{x \neq 0, x \in \{v_{i+1}, \ldots, v_n\}^\perp} \frac{x^\top Bx}{x^\top x} = \max_{x \neq 0, x \in \{v_1, \ldots, v_i\}} \frac{x^\top Bx}{x^\top x},$$

so

$$\frac{w^\top Bw}{w^\top w} \leq \mu_i \quad \text{for all } w \in V_i,$$

and since $v \in V_i$, we have

$$\lambda_i \leq \frac{v^\top Bv}{v^\top v} \leq \mu_i, \quad i = 1, \ldots, m.$$

We can apply the same argument to the symmetric matrices $-A$ and $-B$, to conclude that

$$-\lambda_{n-m+i} \leq -\mu_i,$$

that is,

$$\mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m.$$

Therefore,

$$\lambda_i \leq \mu_i \leq \lambda_{n-m+i}, \quad i = 1, \ldots, m,$$

as desired.

(b) If $\lambda_i = \mu_i$, then

$$\lambda_i = \frac{(Rv)^\top ARv}{(Rv)^\top Rv} = \frac{v^\top Bv}{v^\top v} = \mu_i,$$

so v must be an eigenvector for B and Rv must be an eigenvector for A, both for the eigenvalue $\lambda_i = \mu_i$. $\square$

Proposition 17.25 immediately implies the *Poincaré separation theorem*. It can be used in situations, such as in quantum mechanics, where one has information about the inner products $u_i^\top A u_j$.

Proposition 17.26. *(Poincaré separation theorem) Let A be a $n \times n$ symmetric (or Hermitian) matrix, let m be some integer with $1 \leq m \leq n$, and let $(u_1, \ldots, u_m)$ be m orthonormal vectors. Let $B = (u_i^\top A u_j)$ (an $m \times m$ matrix), let $\lambda_1(A) \leq \ldots \leq \lambda_n(A)$ be the eigenvalues of A and $\lambda_1(B) \leq \ldots \leq \lambda_m(B)$ be the eigenvalues of B; then we have*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-m}(A), \quad k = 1, \ldots, m.$$

Observe that Proposition 17.25 implies that

$$\lambda_1 + \cdots + \lambda_m \leq \operatorname{tr}(R^\top AR) \leq \lambda_{n-m+1} + \cdots + \lambda_n.$$

If P_1 is the the $n \times (n-1)$ matrix obtained from the identity matrix by dropping its last column, we have $P_1^\top P_1 = I$, and the matrix $B = P_1^\top A P_1$ is the matrix obtained from A by deleting its last row and its last column. In this case the interlacing result is

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-2} \leq \lambda_{n-1} \leq \mu_{n-1} \leq \lambda_n,$$

a genuine interlacing. We obtain similar results with the matrix P_{n-m} obtained by dropping the last $n-m$ columns of the identity matrix and setting $B = P_{n-m}^\top A P_{n-m}$ (B is the $m \times m$ matrix obtained from A by deleting its last $n-m$ rows and columns). In this case we have the following interlacing inequalities known as *Cauchy interlacing theorem*:

$$\lambda_k \leq \mu_k \leq \lambda_{k+n-m}, \quad k = 1, \ldots, m. \tag{$*$}$$

17.7 The Courant–Fischer Theorem; Perturbation Results

Another useful tool to prove eigenvalue equalities is the Courant–Fischer characterization of the eigenvalues of a symmetric matrix, also known as the Min-max (and Max-min) theorem.

Theorem 17.27. *(Courant–Fischer) Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. If $\mathcal{V}_k$ denotes the set of subspaces of $\mathbb{R}^n$ of dimension k, then*

$$\begin{aligned}\lambda_k &= \max_{W \in \mathcal{V}_{n-k+1}} \min_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x} \\ \lambda_k &= \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.\end{aligned}$$

Proof. Let us consider the second equality, the proof of the first equality being similar. Let $(u_1, \ldots, u_n)$ be any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i. Observe that the space V_k spanned by $(u_1, \ldots, u_k)$ has dimension k, and by Proposition 17.23, we have

$$\lambda_k = \max_{x \neq 0, x \in V_k} \frac{x^\top A x}{x^\top x} \geq \inf_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

Therefore, we need to prove the reverse inequality; that is, we have to show that

$$\lambda_k \leq \max_{x \neq 0, x \in W} \frac{x^\top A x}{x^\top x}, \quad \text{for all} \quad W \in \mathcal{V}_k.$$

Now for any $W \in \mathcal{V}_k$, if we can prove that $W \cap V_{k-1}^\perp \neq (0)$, then for any nonzero $v \in W \cap V_{k-1}^\perp$, by Proposition 17.24 , we have

$$\lambda_k = \min_{x \neq 0, x \in V_{k-1}^\perp} \frac{x^\top A x}{x^\top x} \leq \frac{v^\top A v}{v^\top v} \leq \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x}.$$

It remains to prove that $\dim(W \cap V_{k-1}^\perp) \geq 1$. However, $\dim(V_{k-1}) = k - 1$, so $\dim(V_{k-1}^\perp) = n - k + 1$, and by hypothesis $\dim(W) = k$. By the Grassmann relation,

$$\dim(W) + \dim(V_{k-1}^\perp) = \dim(W \cap V_{k-1}^\perp) + \dim(W + V_{k-1}^\perp),$$

and since $\dim(W + V_{k-1}^\perp) \leq \dim(\mathbb{R}^n) = n$, we get

$$k + n - k + 1 \leq \dim(W \cap V_{k-1}^\perp) + n;$$

that is, $1 \leq \dim(W \cap V_{k-1}^\perp)$, as claimed. Thus we proved that

$$\lambda_k = \inf_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top A x}{x^\top x},$$

but since the inf is achieved for the subspace V_k, the equation also holds with inf replaced by min. $\square$

The Courant–Fischer theorem yields the following useful result about perturbing the eigenvalues of a symmetric matrix due to Hermann Weyl.

Proposition 17.28. *Given two $n \times n$ symmetric matrices A and $B = A + \Delta A$, if $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ are the eigenvalues of A and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ are the eigenvalues of B, then*

$$|\alpha_k - \beta_k| \leq \rho(\Delta A) \leq \|\Delta A\|_2 \,, \quad k = 1, \ldots, n.$$

Proof. Let $\mathcal{V}_k$ be defined as in the Courant–Fischer theorem and let V_k be the subspace spanned by the k eigenvectors associated with $\alpha_1, \ldots, \alpha_k$. By the Courant–Fischer theorem applied to B, we have

$$\begin{aligned}
\beta_k &= \min_{W \in \mathcal{V}_k} \max_{x \in W, x \neq 0} \frac{x^\top B x}{x^\top x} \\
&\leq \max_{x \in V_k} \frac{x^\top B x}{x^\top x} \\
&= \max_{x \in V_k} \left(\frac{x^\top A x}{x^\top x} + \frac{x^\top \Delta\, A x}{x^\top x} \right) \\
&\leq \max_{x \in V_k} \frac{x^\top A x}{x^\top x} + \max_{x \in V_k} \frac{x^\top \Delta A\, x}{x^\top x}.
\end{aligned}$$

By Proposition 17.23, we have

$$\alpha_k = \max_{x \in V_k} \frac{x^\top A x}{x^\top x},$$

so we obtain

$$\begin{aligned}
\beta_k &\leq \max_{x \in V_k} \frac{x^\top A x}{x^\top x} + \max_{x \in V_k} \frac{x^\top \Delta A\, x}{x^\top x} \\
&= \alpha_k + \max_{x \in V_k} \frac{x^\top \Delta A\, x}{x^\top x} \\
&\leq \alpha_k + \max_{x \in \mathbb{R}^n} \frac{x^\top \Delta A\, x}{x^\top x}.
\end{aligned}$$

Now by Proposition 17.23 and Proposition 9.9, we have

$$\max_{x \in \mathbb{R}^n} \frac{x^\top \Delta A\, x}{x^\top x} = \max_i \lambda_i(\Delta A) \leq \rho(\Delta A) \leq \|\Delta A\|_2 \,,$$

where $\lambda_i(\Delta A)$ denotes the ith eigenvalue of ΔA, which implies that

$$\beta_k \leq \alpha_k + \rho(\Delta A) \leq \alpha_k + \|\Delta A\|_2 \,.$$

By exchanging the roles of A and B, we also have

$$\alpha_k \leq \beta_k + \rho(\Delta A) \leq \beta_k + \|\Delta A\|_2 \,,$$

and thus,

$$|\alpha_k - \beta_k| \leq \rho(\Delta A) \leq \|\Delta A\|_2 \,, \quad k = 1, \ldots, n,$$

as claimed. ∎

Proposition 17.28 also holds for Hermitian matrices.

A pretty result of Wielandt and Hoffman asserts that

$$\sum_{k=1}^{n}(\alpha_k - \beta_k)^2 \leq \|\Delta A\|_F^2,$$

where $\| \, \|_F$ is the Frobenius norm. However, the proof is significantly harder than the above proof; see Lax [113].

The Courant–Fischer theorem can also be used to prove some famous inequalities due to Hermann Weyl. These can also be viewed as perturbation results. Given two symmetric (or Hermitian) matrices A and B, let $\lambda_i(A), \lambda_i(B)$, and $\lambda_i(A+B)$ denote the ith eigenvalue of A, B, and $A+B$, respectively, arranged in nondecreasing order.

Proposition 17.29. *(Weyl) Given two symmetric (or Hermitian) $n \times n$ matrices A and B, the following inequalities hold: For all i, j, k with $1 \leq i, j, k \leq n$:*

1. *If $i + j = k + 1$, then*
$$\lambda_i(A) + \lambda_j(B) \leq \lambda_k(A+B).$$

2. *If $i + j = k + n$, then*
$$\lambda_k(A+B) \leq \lambda_i(A) + \lambda_j(B).$$

Proof. Observe that the first set of inequalities is obtained from the second set by replacing A by $-A$ and B by $-B$, so it is enough to prove the second set of inequalities. By the Courant–Fischer theorem, there is a subspace H of dimension $n-k+1$ such that

$$\lambda_k(A+B) = \min_{x \in H, x \neq 0} \frac{x^\top (A+B)x}{x^\top x}.$$

Similarly, there exists a subspace F of dimension i and a subspace G of dimension j such that

$$\lambda_i(A) = \max_{x \in F, x \neq 0} \frac{x^\top A x}{x^\top x}, \quad \lambda_j(B) = \max_{x \in G, x \neq 0} \frac{x^\top B x}{x^\top x}.$$

We claim that $F \cap G \cap H \neq (0)$. To prove this, we use the Grassmann relation twice. First,

$$\dim(F \cap G \cap H) = \dim(F) + \dim(G \cap H) - \dim(F + (G \cap H)) \geq \dim(F) + \dim(G \cap H) - n,$$

and second,

$$\dim(G \cap H) = \dim(G) + \dim(H) - \dim(G + H) \geq \dim(G) + \dim(H) - n,$$

so

$$\dim(F \cap G \cap H) \geq \dim(F) + \dim(G) + \dim(H) - 2n.$$

However,

$$\dim(F) + \dim(G) + \dim(H) = i + j + n - k + 1$$

and $i + j = k + n$, so we have

$$\dim(F \cap G \cap H) \geq i + j + n - k + 1 - 2n = k + n + n - k + 1 - 2n = 1,$$

which shows that $F \cap G \cap H \neq (0)$. Then for any unit vector $z \in F \cap G \cap H \neq (0)$, we have

$$\lambda_k(A + B) \leq z^\top (A + B)z, \quad \lambda_i(A) \geq z^\top Az, \quad \lambda_j(B) \geq z^\top Bz,$$

establishing the desired inequality $\lambda_k(A + B) \leq \lambda_i(A) + \lambda_j(B)$. $\square$

In the special case $i = j = k$, we obtain

$$\lambda_1(A) + \lambda_1(B) \leq \lambda_1(A + B), \quad \lambda_n(A + B) \leq \lambda_n(A) + \lambda_n(B).$$

It follows that λ_1 (as a function) is concave, while λ_n (as a function) is convex.

If $i = k$ and $j = 1$, we obtain

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A + B),$$

and if $i = k$ and $j = n$, we obtain

$$\lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B),$$

and combining them, we get

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B).$$

In particular, if B is positive semidefinite, since its eigenvalues are nonnegative, we obtain the following inequality known as the *monotonicity theorem* for symmetric (or Hermitian) matrices: if A and B are symmetric (or Hermitian) and B is positive semidefinite, then

$$\lambda_k(A) \leq \lambda_k(A + B) \quad k = 1, \ldots, n.$$

The reader is referred to Horn and Johnson [95] (Chapters 4 and 7) for a very complete treatment of matrix inequalities and interlacing results, and also to Lax [113] and Serre [156].

17.8 Summary

The main concepts and results of this chapter are listed below:

- *Normal* linear maps, *self-adjoint* linear maps, *skew-self-adjoint* linear maps, and *orthogonal* linear maps.

- Properties of the eigenvalues and eigenvectors of a normal linear map.
- The *complexification* of a real vector space, of a linear map, and of a Euclidean inner product.
- The eigenvalues of a self-adjoint map in a Hermitian space are *real*.
- The eigenvalues of a self-adjoint map in a Euclidean space are *real*.
- Every self-adjoint linear map on a Euclidean space has an orthonormal basis of eigenvectors.
- Every normal linear map on a Euclidean space can be block diagonalized (blocks of size at most 2×2) with respect to an orthonormal basis of eigenvectors.
- Every normal linear map on a Hermitian space can be diagonalized with respect to an orthonormal basis of eigenvectors.
- The spectral theorems for self-adjoint, skew-self-adjoint, and orthogonal linear maps (on a Euclidean space).
- The spectral theorems for normal, symmetric, skew-symmetric, and orthogonal (real) matrices.
- The spectral theorems for normal, Hermitian, skew-Hermitian, and unitary (complex) matrices.
- The *Rayleigh ratio* and the *Rayleigh–Ritz theorem*.
- *Interlacing inequalities* and the *Cauchy interlacing theorem*.
- The *Poincaré separation theorem*.
- The *Courant–Fischer theorem*.
- Inequalities involving perturbations of the eigenvalues of a symmetric matrix.
- The *Weyl inequalities*.

17.9 Problems

Problem 17.1. Prove that the structure $E_{\mathbb{C}}$ introduced in Definition 17.2 is indeed a complex vector space.

Problem 17.2. Prove that the formula

$$\langle u_1 + iv_1, u_2 + iv_2\rangle_{\mathbb{C}} = \langle u_1, u_2\rangle + \langle v_1, v_2\rangle + i(\langle v_1, u_2\rangle - \langle u_1, v_2\rangle)$$

defines a Hermitian form on $E_{\mathbb{C}}$ that is positive definite and that $\langle -, -\rangle_{\mathbb{C}}$ agrees with $\langle -, -\rangle$ on real vectors.

Problem 17.3. Given any linear map $f\colon E \to E$, prove the map $f_{\mathbb{C}}^*$ defined such that

$$f_{\mathbb{C}}^*(u + iv) = f^*(u) + if^*(v)$$

for all $u, v \in E$ is the adjoint of $f_{\mathbb{C}}$ w.r.t. $\langle -, -\rangle_{\mathbb{C}}$.

Problem 17.4. Let A be a real symmetric $n \times n$ matrix whose eigenvalues are nonnegative. Prove that for every $p > 0$, there is a real symmetric matrix S whose eigenvalues are nonnegative such that $S^p = A$.

Problem 17.5. Let A be a real symmetric $n \times n$ matrix whose eigenvalues are positive.

(1) Prove that there is a real symmetric matrix S such that $A = e^S$.

(2) Let S be a real symmetric $n \times n$ matrix. Prove that $A = e^S$ is a real symmetric $n \times n$ matrix whose eigenvalues are positive.

Problem 17.6. Let A be a complex matrix. Prove that if A can be diagonalized with respect to an orthonormal basis, then A is normal.

Problem 17.7. Let $f\colon \mathbb{C}^n \to \mathbb{C}^n$ be a linear map.

(1) Prove that if f is diagonalizable and if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of f, then $\lambda_1^2, \ldots, \lambda_n^2$ are the eigenvalues of f^2, and if $\lambda_i^2 = \lambda_j^2$ implies that $\lambda_i = \lambda_j$, then f and f^2 have the same eigenspaces.

(2) Let f and g be two real self-adjoint linear maps $f, g\colon \mathbb{R}^n \to \mathbb{R}^n$. Prove that if f and g have nonnegative eigenvalues (f and g are positve semidefinite) and if $f^2 = g^2$, then $f = g$.

Problem 17.8. (1) Let $\mathfrak{so}(3)$ be the space of 3×3 skew symmetric matrices

$$\mathfrak{so}(3) = \left\{ \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \;\middle|\; a, b, c \in \mathbb{R} \right\}.$$

For any matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \in \mathfrak{so}(3),$$

if we let $\theta = \sqrt{a^2 + b^2 + c^2}$, recall from Section 12.7 (the Rodrigues formula) that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is given by

$$e^A = I_3 + \frac{\sin\theta}{\theta}A + \frac{(1 - \cos\theta)}{\theta^2}A^2, \quad \text{if } \theta \neq 0,$$

with $\exp(0_3) = I_3$.

(2) Prove that e^A is an orthogonal matrix of determinant $+1$, i.e., a rotation matrix.

(3) Prove that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective. For this proceed as follows: Pick any rotation matrix $R \in \mathbf{SO}(3)$;

(1) The case $R = I$ is trivial.

(2) If $R \neq I$ and $\mathrm{tr}(R) \neq -1$, then

$$\exp^{-1}(R) = \left\{ \frac{\theta}{2\sin\theta}(R - R^T) \;\middle|\; 1 + 2\cos\theta = \mathrm{tr}(R) \right\}.$$

(Recall that $\mathrm{tr}(R) = r_{1\,1} + r_{2\,2} + r_{3\,3}$, the *trace* of the matrix R).

Show that there is a unique skew-symmetric B with corresponding θ satisfying $0 < \theta < \pi$ such that $e^B = R$.

(3) If $R \neq I$ and $\mathrm{tr}(R) = -1$, then prove that the eigenvalues of R are $1, -1, -1$, that $R = R^\top$, and that $R^2 = I$. Prove that the matrix

$$S = \frac{1}{2}(R - I)$$

is a symmetric matrix whose eigenvalues are $-1, -1, 0$. Thus S can be diagonalized with respect to an orthogonal matrix Q as

$$S = Q \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} Q^\top.$$

Prove that there exists a skew symmetric matrix

$$U = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

so that

$$U^2 = S = \frac{1}{2}(R - I).$$

Observe that

$$U^2 = \begin{pmatrix} -(c^2 + d^2) & bc & bd \\ bc & -(b^2 + d^2) & cd \\ bd & cd & -(b^2 + c^2) \end{pmatrix},$$

and use this to conclude that if $U^2 = S$, then $b^2 + c^2 + d^2 = 1$. Then show that

$$\exp^{-1}(R) = \left\{ (2k+1)\pi \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},\ k \in \mathbb{Z} \right\},$$

where (b, c, d) is any unit vector such that for the corresponding skew symmetric matrix U, we have $U^2 = S$.

(4) To find a skew symmetric matrix U so that $U^2 = S = \frac{1}{2}(R - I)$ as in (3), we can solve the system

$$\begin{pmatrix} b^2 - 1 & bc & bd \\ bc & c^2 - 1 & cd \\ bd & cd & d^2 - 1 \end{pmatrix} = S.$$

We immediately get b^2, c^2, d^2, and then, since one of b, c, d is nonzero, say b, if we choose the positive square root of b^2, we can determine c and d from bc and bd.

Implement a computer program in `Matlab` to solve the above system.

Problem 17.9. It was shown in Proposition 15.15 that the exponential map is a map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$, where $\mathfrak{so}(n)$ is the vector space of real $n \times n$ skew-symmetric matrices. Use the spectral theorem to prove that the map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$ is surjective.

Problem 17.10. Let $\mathfrak{u}(n)$ be the space of (complex) $n \times n$ skew-Hermitian matrices ($B^* = -B$) and let $\mathfrak{su}(n)$ be its subspace consisting of skew-Hermitian matrice with zero trace ($\mathrm{tr}(B) = 0$).

(1) Prove that if $B \in \mathfrak{u}(n)$, then $e^B \in \mathbf{U}(n)$, and if if $B \in \mathfrak{su}(n)$, then $e^B \in \mathbf{SU}(n)$. Thus we have well-defined maps $\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$ and $\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$.

(2) Prove that the map $\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$ is surjective.

(3) Prove that the map $\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$ is surjective.

Problem 17.11. Recall that a matrix $B \in \mathrm{M}_n(\mathbb{R})$ is skew-symmetric if $B^\top = -B$. Check that the set $\mathfrak{so}(n)$ of skew-symmetric matrices is a vector space of dimension $n(n-1)/2$, and thus is isomorphic to $\mathbb{R}^{n(n-1)/2}$.

(1) Given a rotation matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

where $0 < \theta < \pi$, prove that there is a skew symmetric matrix B such that

$$R = (I - B)(I + B)^{-1}.$$

(2) Prove that the eigenvalues of a skew-symmetric matrix are either 0 or pure imaginary (that is, of the form $i\mu$ for $\mu \in \mathbb{R}$.).

Let $C\colon \mathfrak{so}(n) \to \mathrm{M}_n(\mathbb{R})$ be the function (called the *Cayley transform* of B) given by

$$C(B) = (I - B)(I + B)^{-1}.$$

Prove that if B is skew-symmetric, then $I - B$ and $I + B$ are invertible, and so C is well-defined. Prove that

$$(I + B)(I - B) = (I - B)(I + B),$$

and that

$$(I + B)(I - B)^{-1} = (I - B)^{-1}(I + B).$$

Prove that

$$(C(B))^\top C(B) = I$$

and that

$$\det C(B) = +1,$$

so that $C(B)$ is a rotation matrix. Furthermore, show that $C(B)$ does not admit -1 as an eigenvalue.

(3) Let $\mathbf{SO}(n)$ be the group of $n \times n$ rotation matrices. Prove that the map

$$C\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is bijective onto the subset of rotation matrices that do not admit -1 as an eigenvalue. Show that the inverse of this map is given by

$$B = (I + R)^{-1}(I - R) = (I - R)(I + R)^{-1},$$

where $R \in \mathbf{SO}(n)$ does not admit -1 as an eigenvalue.

Problem 17.12. Please refer back to Problem 4.6. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A (not necessarily distinct). Using Schur's theorem, A is similar to an upper triangular matrix B, that is, $A = PBP^{-1}$ with B upper triangular, and we may assume that the diagonal entries of B in descending order are $\lambda_1, \ldots, \lambda_n$.

(1) If the E_{ij} are listed according to total order given by

$$(i, j) < (h, k) \quad \text{iff} \quad \begin{cases} i = h \text{ and } j > k \\ \text{or } i < h. \end{cases}$$

prove that R_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_n, \ldots, \lambda_1, \ldots, \lambda_n, \ldots, \lambda_1}_{n^2}),$$

and that L_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_1, \ldots, \lambda_1}_{n} \ldots, \underbrace{\lambda_n, \ldots, \lambda_n}_{n}).$$

Hint. Figure out what are $R_B(E_{ij}) = E_{ij}B$ and $L_B(E_{ij}) = BE_{ij}$.

(2) Use the fact that

$$L_A = L_P \circ L_B \circ L_P^{-1}, \quad R_A = R_P^{-1} \circ R_B \circ R_P,$$

to express $\mathrm{ad}_A = L_A - R_A$ in terms of $L_B - R_B$, and conclude that the eigenvalues of ad_A are $\lambda_i - \lambda_j$, for $i = 1, \ldots, n$, and for $j = n, \ldots, 1$.

Chapter 18

Computing Eigenvalues and Eigenvectors

After the problem of solving a linear system, the problem of computing the eigenvalues and the eigenvectors of a real or complex matrix is one of most important problems of numerical linear algebra. Several methods exist, among which we mention Jacobi, Givens–Householder, divide-and-conquer, QR iteration, and Rayleigh–Ritz; see Demmel [48], Trefethen and Bau [176], Meyer [125], Serre [156], Golub and Van Loan [80], and Ciarlet [41]. Typically, better performing methods exist for special kinds of matrices, such as symmetric matrices.

In theory, given an $n \times n$ complex matrix A, if we could compute a Schur form $A = UTU^*$, where T is upper triangular and U is unitary, we would obtain the eigenvalues of A, since they are the diagonal entries in T. However, this would require finding the roots of a polynomial, but methods for doing this are known to be numerically very unstable, so this is not a practical method.

A common paradigm is to construct a sequence (P_k) of matrices such that $A_k = P_k^{-1} A P_k$ converges, in some sense, to a matrix whose eigenvalues are easily determined. For example, $A_k = P_k^{-1} A P_k$ could become upper triangular in the limit. Furthermore, P_k is typically a product of "nice" matrices, for example, orthogonal matrices.

For general matrices, that is, matrices that are not symmetric, the QR iteration algorithm, due to Rutishauser, Francis, and Kublanovskaya in the early 1960s, is one of the most efficient algorithms for computing eigenvalues. A fascinating account of the history of the QR algorithm is given in Golub and Uhlig [79]. The QR algorithm constructs a sequence of matrices (A_k), where A_{k+1} is obtained from A_k by performing a QR-decomposition $A_k = Q_k R_k$ of A_k and then setting $A_{k+1} = R_k Q_k$, the result of swapping Q_k and R_k. It is immediately verified that $A_{k+1} = Q_k^* A_k Q_k$, so A_k and A_{k+1} have *the same eigenvalues*, which are the eigenvalues of A.

The basic version of this algorithm runs into difficulties with matrices that have several eigenvalues with the same modulus (it may loop or not "converge" to an upper triangular matrix). There are ways of dealing with some of these problems, but for ease of exposition, we first present a simplified version of the QR algorithm which we call basic QR algorithm. We prove a convergence theorem for the basic QR algorithm, under the rather restrictive hypothesis that the input matrix A is diagonalizable and that its eigenvalues are nonzero and have distinct moduli. The proof shows that the part of A_k strictly below the diagonal converges to zero and that the diagonal entries of A_k converge to the eigenvalues of A.

Since the convergence of the QR method depends crucially only on the fact that the part of A_k below the diagonal goes to zero, it would be highly desirable if we could replace A by a similar matrix U^*AU easily computable from A and having lots of zero strictly below the diagonal. It turns out that there is a way to construct a matrix $H = U^*AU$ which is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Such matrices called, *Hessenberg matrices*, are discussed in Section 18.2. An $n \times n$ diagonalizable Hessenberg matrix H having the property that $h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$ (such a matrix is called *unreduced*) has the nice property that its eigenvalues are all distinct. Since every Hessenberg matrix is a block diagonal matrix of unreduced Hessenberg blocks, *it suffices to compute the eigenvalues of unreduced Hessenberg matrices.* There is a special case of particular interest: symmetric (or Hermitian) positive definite tridiagonal matrices. Such matrices must have real positive distinct eigenvalues, so the QR algorithm converges to a diagonal matrix.

In Section 18.3, we consider techniques for making the basic QR method practical and more efficient. The first step is to convert the original input matrix A to a similar matrix H in Hessenberg form, and to apply the QR algorithm to H (actually, to the unreduced blocks of H). The second and crucial ingredient to speed up convergence is to add shifts.

A shift is the following step: pick some σ_k, hopefully close to some eigenvalue of A (in general, λ_n), QR-factor $A_k - \sigma_k I$ as

$$A_k - \sigma_k I = Q_k R_k,$$

and then form

$$A_{k+1} = R_k Q_k + \sigma_k I.$$

It is easy to see that we still have $A_{k+1} = Q_k^* A_k Q_k$. A judicious choice of σ_k can speed up convergence considerably. If H is real and has pairs of complex conjugate eigenvalues, we can perform a double shift, and it can be arranged that we work in real arithmetic.

The last step for improving efficiency is to compute $A_{k+1} = Q_k^* A_k Q_k$ without even performing a QR-factorization of $A_k - \sigma_k I$. This can be done when A_k is unreduced Hessenberg. Such a method is called QR iteration with implicit shifts. There is also a version of QR iteration with implicit double shifts.

If the dimension of the matrix A is very large, we can find approximations of some of the eigenvalues of A by using a truncated version of the reduction to Hessenberg form due to Arnoldi in general and to Lanczos in the symmetric (or Hermitian) tridiagonal case. *Arnoldi iteration* is discussed in Section 18.4. If A is an $m \times m$ matrix, for $n \ll m$ (n much smaller than m) the idea is to generate the $n \times n$ Hessenberg submatrix H_n of the full Hessenberg matrix H (such that $A = UHU^*$) consisting of its first n rows and n columns; the matrix U_n consisting of the first n columns of U is also produced. The Rayleigh–Ritz method consists in computing the eigenvalues of H_n using the QR- method with shifts. These eigenvalues, called *Ritz values*, are approximations of the eigenvalues of A. Typically, extreme eigenvalues are found first.

Arnoldi iteration can also be viewed as a way of computing an orthonormal basis of a *Krylov subspace*, namely the subspace $\mathcal{K}_n(A, b)$ spanned by $(b, Ab, \ldots, A^n b)$. We can also use Arnoldi iteration to find an approximate solution of a linear equation $Ax = b$ by minimizing $\|b - Ax_n\|_2$ for all x_n is the Krylov space $\mathcal{K}_n(A, b)$. This method named GMRES is discussed in Section 18.5.

The special case where H is a symmetric (or Hermitian) tridiagonal matrix is discussed in Section 18.6. In this case, Arnoldi's algorithm becomes *Lanczos' algorithm*. It is much more efficient than Arnoldi iteration.

We close this chapter by discussing two classical methods for computing a single eigenvector and a single eigenvalue: power iteration and inverse (power) iteration; see Section 18.7.

18.1 The Basic QR Algorithm

Let A be an $n \times n$ matrix which is assumed to be diagonalizable and invertible. The basic QR algorithm makes use of two very simple steps. Starting with $A_1 = A$, we construct sequences of matrices (A_k), (Q_k) (R_k) and (P_k) as follows:

$$\begin{array}{ll} \text{Factor} & A_1 = Q_1 R_1 \\ \text{Set} & A_2 = R_1 Q_1 \\ \text{Factor} & A_2 = Q_2 R_2 \\ \text{Set} & A_3 = R_2 Q_2 \\ & \vdots \\ \text{Factor} & A_k = Q_k R_k \\ \text{Set} & A_{k+1} = R_k Q_k \\ & \vdots \end{array}$$

Thus, A_{k+1} is obtained from a QR-factorization $A_k = Q_k R_k$ of A_k by swapping Q_k and R_k. Define P_k by

$$P_k = Q_1 Q_2 \cdots Q_k.$$

Since $A_k = Q_k R_k$, we have $R_k = Q_k^* A_k$, and since $A_{k+1} = R_k Q_k$, we obtain

$$A_{k+1} = Q_k^* A_k Q_k. \qquad (*_1)$$

An obvious induction shows that

$$A_{k+1} = Q_k^* \cdots Q_1^* A_1 Q_1 \cdots Q_k = P_k^* A P_k,$$

that is

$$A_{k+1} = P_k^* A P_k. \tag{$*_2$}$$

Therefore, A_{k+1} and A are similar, so they have the same eigenvalues.

The basic QR iteration method consists in computing the sequence of matrices A_k, and in the ideal situation, to expect that A_k "converges" to an upper triangular matrix, more precisely that the part of A_k below the main diagonal goes to zero, and the diagonal entries converge to the eigenvalues of A.

This ideal situation is only achieved in rather special cases. For one thing, if A is unitary (or orthogonal in the real case), since in the QR decomposition we have $R = I$, we get $A_2 = IQ = Q = A_1$, so the method does *not* make any progress. Also, if A is a real matrix, since the A_k are also real, if A has complex eigenvalues, then the part of A_k below the main diagonal can't go to zero. Generally, the method runs into troubles whenever A has distinct eigenvalues with the same modulus.

The convergence of the sequence (A_k) is only known under some fairly restrictive hypotheses. Even under such hypotheses, this is not really genuine convergence. Indeed, it can be shown that the part of A_k below the main diagonal goes to zero, and the diagonal entries converge to the eigenvalues of A, but the part of A_k above the diagonal *may not converge*. However, for the purpose of finding the eigenvalues of A, this does not matter.

The following convergence result is proven in Ciarlet [41] (Chapter 6, Theorem 6.3.10 and Serre [156] (Chapter 13, Theorem 13.2). It is rarely applicable in practice, except for symmetric (or Hermitian) positive definite matrices, as we will see shortly.

Theorem 18.1. *Suppose the (complex) $n \times n$ matrix A is invertible, diagonalizable, and that its eigenvalues $\lambda_1, \ldots, \lambda_n$ have different moduli, so that*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0.$$

If $A = P \Lambda P^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, and if P^{-1} has an LU-factorization, then the strictly lower-triangular part of A_k converges to zero, and the diagonal of A_k converges to Λ.

Proof. We reproduce the proof in Ciarlet [41] (Chapter 6, Theorem 6.3.10). The strategy is to study the asymptotic behavior of the matrices $P_k = Q_1 Q_2 \cdots Q_k$. For this, it turns out that we need to consider the powers A^k.

Step 1. Let $\mathcal{R}_k = R_k \cdots R_2 R_1$. We claim that

$$A^k = (Q_1 Q_2 \cdots Q_k)(R_k \cdots R_2 R_1) = P_k \mathcal{R}_k. \tag{$*_3$}$$

We proceed by induction. The base case $k = 1$ is trivial. For the induction step, from $(*_2)$, we have

$$P_k A_{k+1} = A P_k.$$

Since $A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1}$, we have

$$P_{k+1}\mathcal{R}_{k+1} = P_k Q_{k+1} R_{k+1} \mathcal{R}_k = P_k A_{k+1} \mathcal{R}_k = A P_k \mathcal{R}_k = A A^k = A^{k+1}$$

establishing the induction step.

Step 2. We will express the matrix P_k as $P_k = Q\widetilde{Q}_k D_k$, in terms of a diagonal matrix D_k with unit entries, with Q and $\widetilde{Q}_k$ unitary.

Let $P = QR$, a QR-factorization of P (with R an upper triangular matrix with positive diagonal entries), and $P^{-1} = LU$, an LU-factorization of P^{-1}. Since $A = P\Lambda P^{-1}$, we have

$$A^k = P\Lambda^k P^{-1} = QR\Lambda^k LU = QR(\Lambda^k L \Lambda^{-k})\Lambda^k U. \tag{$*_4$}$$

Here, Λ^{-k} is the diagonal matrix with entries λ_i^{-k}. The reason for introducing the matrix $\Lambda^k L\Lambda^{-k}$ is that its asymptotic behavior is easy to determine. Indeed, we have

$$(\Lambda^k L \Lambda^{-k})_{ij} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i = j \\ \left(\dfrac{\lambda_i}{\lambda_j}\right)^k L_{ij} & \text{if } i > j. \end{cases}$$

The hypothesis that $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ implies that

$$\lim_{k\mapsto\infty} \Lambda^k L \Lambda^{-k} = I. \tag{$\dagger$}$$

Note that it is to obtain this limit that we made the hypothesis on the moduli of the eigenvalues. We can write

$$\Lambda^k L \Lambda^{-k} = I + F_k, \quad \text{with} \quad \lim_{k\mapsto\infty} F_k = 0,$$

and consequently, since $R(\Lambda^k L\Lambda^{-k}) = R(I + F_k) = R + RF_kR^{-1}R = (I + RF_kR^{-1})R$, we have

$$R(\Lambda^k L \Lambda^{-k}) = (I + RF_kR^{-1})R. \tag{$*_5$}$$

By Proposition 9.11(1), since $\lim_{k\mapsto\infty} F_k = 0$, and thus $\lim_{k\mapsto\infty} RF_kR^{-1} = 0$, the matrices $I + RF_kR^{-1}$ are invertible for k large enough. Consequently for k large enough, we have a QR-factorization

$$I + RF_kR^{-1} = \widetilde{Q}_k \widetilde{R}_k, \tag{$*_6$}$$

with $(\widetilde{R}_k)_{ii} > 0$ for $i = 1, \ldots, n$. Since the matrices $\widetilde{Q}_k$ are unitary, we have $\left\|\widetilde{Q}_k\right\|_2 = 1$, so the sequence $(\widetilde{Q}_k)$ is bounded. It follows that it has a convergent subsequence $(\widetilde{Q}_\ell)$ that converges to some matrix $\widetilde{Q}$, which is also unitary. Since

$$\widetilde{R}_\ell = (\widetilde{Q}_\ell)^*(I + RF_\ell R^{-1}),$$

we deduce that the subsequence $(\widetilde{R}_\ell)$ also converges to some matrix $\widetilde{R}$, which is also upper triangular with positive diagonal entries. By passing to the limit (using the subsequences), we get $\widetilde{R} = (\widetilde{Q})^*$, that is,

$$I = \widetilde{Q}\widetilde{R}.$$

By the uniqueness of a QR-decomposition (when the diagonal entries of R are positive), we get

$$\widetilde{Q} = \widetilde{R} = I.$$

Since the above reasoning applies to any subsequences of $(\widetilde{Q}_k)$ and $(\widetilde{R}_k)$, by the uniqueness of limits, we conclude that the "full" sequences $(\widetilde{Q}_k)$ and $(\widetilde{R}_k)$ converge:

$$\lim_{k \mapsto \infty} \widetilde{Q}_k = I, \quad \lim_{k \mapsto \infty} \widetilde{R}_k = I.$$

Since by $(*_4)$,

$$A^k = QR(\Lambda^k L \Lambda^{-k})\Lambda^k U,$$

by $(*_5)$,

$$R(\Lambda^k L \Lambda^{-k}) = (I + RF_k R^{-1})R,$$

and by $(*_6)$

$$I + RF_k R^{-1} = \widetilde{Q}_k \widetilde{R}_k,$$

we proved that

$$A^k = (Q\widetilde{Q}_k)(\widetilde{R}_k R \Lambda^k U). \tag{$*_7$}$$

Observe that $Q\widetilde{Q}_k$ is a unitary matrix, and $\widetilde{R}_k R\Lambda^k U$ is an upper triangular matrix, as a product of upper triangular matrices. However, some entries in Λ may be negative, so we can't claim that $\widetilde{R}_k R\Lambda^k U$ has positive diagonal entries. Nevertheless, we have another QR-decomposition of A^k,

$$A^k = (Q\widetilde{Q}_k)(\widetilde{R}_k R \Lambda^k U) = P_k \mathcal{R}_k.$$

It is easy to prove that there is diagonal matrix D_k with $|(D_k)_{ii}| = 1$ for $i = 1, \ldots, n$, such that

$$P_k = Q\widetilde{Q}_k D_k. \tag{$*_8$}$$

The existence of D_k is consequence of the following fact: If an invertible matrix B has two QR factorizations $B = Q_1 R_1 = Q_2 R_2$, then there is a diagonal matrix D with unit entries such that $Q_2 = DQ_1$.

The expression for P_k in $(*_8)$ is that which we were seeking.

Step 3. Asymptotic behavior of the matrices $A_{k+1} = P_k^* A P_k$.

Since $A = P\Lambda P^{-1} = QR\Lambda R^{-1}Q^{-1}$ and by $(*_8)$, $P_k = Q\widetilde{Q}_k D_k$, we get

$$A_{k+1} = D_k^*(\widetilde{Q}_k)^* Q^* QR\Lambda R^{-1} Q^{-1} Q\widetilde{Q}_k D_k = D_k^*(\widetilde{Q}_k)^* R\Lambda R^{-1} \widetilde{Q}_k D_k. \tag{$*_9$}$$

Since $\lim_{k\mapsto\infty} \widetilde{Q}_k = I$, we deduce that

$$\lim_{k\mapsto\infty} (\widetilde{Q}_k)^* R\Lambda R^{-1}\widetilde{Q}_k = R\Lambda R^{-1} = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

an upper triangular matrix with the eigenvalues of A on the diagonal. Since R is upper triangular, the order of the eigenvalues is preserved. If we let

$$\mathcal{D}_k = (\widetilde{Q}_k)^* R\Lambda R^{-1}\widetilde{Q}_k, \tag{$*_{10}$}$$

then by ($*_9$) we have $A_{k+1} = D_k^* \mathcal{D}_k D_k$, and since the matrices D_k are diagonal matrices, we have

$$(A_{k+1})_{jj} = (D_k^* \mathcal{D}_k D_k)_{ij} = \overline{(D_k)_{ii}}(D_k)_{jj}(\mathcal{D}_k)_{ij},$$

which implies that

$$(A_{k+1})_{ii} = (\mathcal{D}_k)_{ii}, \quad i = 1, \ldots, n, \tag{$*_{11}$}$$

since $|(D_k)_{ii}| = 1$ for $i = 1, \ldots, n$. Since $\lim_{k\mapsto\infty} \mathcal{D}_k = R\Lambda R^{-1}$, we conclude that the strictly lower-triangular part of A_{k+1} converges to zero, and the diagonal of A_{k+1} converges to Λ. $\square$

Observe that if the matrix A is real, then the hypothesis that the eigenvalues have distinct moduli implies that the eigenvalues are all real and simple.

The following `Matlab` program implements the basic QR-method using the function `qrv4` from Section 12.8.

```
function T = qreigen(A,m)
T = A;
for k = 1:m
  [Q R] = qrv4(T);
  T = R*Q;
end
end
```

Example 18.1. If we run the function `qreigen` with 100 iterations on the 8×8 symmetric matrix

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 \end{pmatrix},$$

we find the matrix

$$T = \begin{pmatrix} 5.8794 & 0.0015 & 0.0000 & -0.0000 & 0.0000 & -0.0000 & 0.0000 & -0.0000 \\ 0.0015 & 5.5321 & 0.0001 & 0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0.0001 & 5.0000 & 0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0 & 0.0000 & 4.3473 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0 & 0 & 0.0000 & 3.6527 & 0.0000 & 0.0000 & -0.0000 \\ 0 & 0 & 0 & 0 & 0.0000 & 3.0000 & 0.0000 & -0.0000 \\ 0 & 0 & 0 & 0 & 0 & 0.0000 & 2.4679 & 0.0000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0000 & 2.1206 \end{pmatrix}.$$

The diagonal entries match the eigenvalues found by running the `Matlab` function `eig(A)`.

If several eigenvalues have the same modulus, then the proof breaks down, we can no longer claim (†), namely that

$$\lim_{k \mapsto \infty} \Lambda^k L \Lambda^{-k} = I.$$

If we assume that P^{-1} has a suitable "block LU-factorization," it can be shown that the matrices A_{k+1} converge to a block upper-triangular matrix, where each block corresponds to eigenvalues having the same modulus. For example, if A is a 9×9 matrix with eigenvalues λ_i such that $|\lambda_1| = |\lambda_2| = |\lambda_3| > |\lambda_4| > |\lambda_5| = |\lambda_6| = |\lambda_7| = |\lambda_8| = |\lambda_9|$, then A_k converges to a block diagonal matrix (with three blocks, a 3×3 block, a 1×1 block, and a 5×5 block) of the form

$$\begin{pmatrix} \star & \star & \star & * & * & * & * & * & * \\ \star & \star & \star & * & * & * & * & * & * \\ \star & \star & \star & * & * & * & * & * & * \\ 0 & 0 & 0 & \star & * & * & * & * & * \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & \star & \star & \star & \star & \star \end{pmatrix}.$$

See Ciarlet [41] (Chapter 6 Section 6.3) for more details.

Under the conditions of Theorem 18.1, in particular, if A is a symmetric (or Hermitian) positive definite matrix, the eigenvectors of A can be approximated. However, when A is not a symmetric matrix, since the upper triangular part of A_k does not necessarily converge, one has to be cautious that a rigorous justification is lacking.

Suppose we apply the QR algorithm to a matrix A satisfying the hypotheses of Theorem Theorem 18.1. For k large enough, $A_{k+1} = P_k^* A P_k$ is nearly upper triangular and the diagonal entries of A_{k+1} are all distinct, so we can consider that they are the eigenvalues of A_{k+1}, and thus of A. To avoid too many subscripts, write T for the upper triangular matrix obtained by settting the entries of the part of A_{k+1} below the diagonal to 0. Then we can find the corresponding eigenvectors by solving the linear system

$$Tv = t_{ii}v,$$

and since T is upper triangular, this can be done by bottom-up elimination. We leave it as an exercise to show that the following vectors $v^i = (v_1^i, \ldots, v_n^i)$ are eigenvectors:

$$v^1 = e_1,$$

and if $i = 2, \ldots, n$, then

$$v_j^i = \begin{cases} 0 & \text{if } i+1 \leq j \leq n \\ 1 & \text{if } j = i \\ -\dfrac{t_{jj+1}v_{j+1}^i + \cdots + t_{ji}v_i^i}{t_{jj} - t_{ii}} & \text{if } i-1 \geq j \geq 1. \end{cases}$$

Then the vectors $(P_k v^1, \ldots, P_k v^n)$ are a basis of (approximate) eigenvectors for A. In the special case where T is a diagonal matrix, then $v^i = e_i$ for $i = 1, \ldots, n$ and the columns of P_k are an orthonormal basis of (approximate) eigenvectors for A.

If A is a real matrix whose eigenvalues are not all real, then there is some complex pair of eigenvalues $\lambda + i\mu$ (with $\mu \neq 0$), and the QR-algorithm cannot converge to a matrix whose strictly lower-triangular part is zero. There is a way to deal with this situation using upper Hessenberg matrices which will be discussed in the next section.

Since the convergence of the QR method depends crucially only on the fact that the part of A_k below the diagonal goes to zero, it would be highly desirable if we could replace A by a similar matrix U^*AU easily computable from A having lots of zero strictly below the diagonal. We can't expect U^*AU to be a diagonal matrix (since this would mean that A was easily diagonalized), but it turns out that there is a way to construct a matrix $H = U^*AU$ which is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Such matrices called Hessenberg matrices are discussed in the next section.

18.2 Hessenberg Matrices

Definition 18.1. An $n \times n$ matrix (real or complex) H is an *(upper) Hessenberg matrix* if it is almost triangular, except that it may have an extra nonzero diagonal below the main diagonal. Technically, $h_{jk} = 0$ for all (j, k) such that $j - k \geq 2$.

The 5×5 matrix below is an example of a Hessenberg matrix.

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

The following result can be shown.

Theorem 18.2. *Every $n \times n$ complex or real matrix A is similar to an upper Hessenberg matrix H, that is, $A = UHU^*$ for some unitary matrix U. Furthermore, U can be constructed as a product of Householder matrices (the definition is the same as in Section 13.1, except that W is a complex vector, and that the inner product is the Hermitian inner product on $\mathbb{C}^n$). If A is a real matrix, then U is an orthogonal matrix (and H is a real matrix).*

Theorem 18.2 and algorithms for converting a matrix to Hessenberg form are discussed in Trefethen and Bau [176] (Lecture 26), Demmel [48] (Section 4.4.6, in the real case), Serre [156] (Theorem 13.1), and Meyer [125] (Example 5.7.4, in the real case). The proof of correctness is not difficult and will be the object of a homework problem.

The following functions written in `Matlab` implement a function to compute a Hessenberg form of a matrix.

The function `house` constructs the normalized vector u defining the Householder reflection that zeros all but the first entries in a vector x.

```
function [uu, u] = house(x)
tol = 2*10^(-15);   % tolerance
uu = x;
p = size(x,1);
% computes l^1-norm of x(2:p,1)
n1 = sum(abs(x(2:p,1)));
if n1 <= tol
   u = zeros(p,1);  uu = u;
else
   l = sqrt(x'*x);  % l^2 norm of x
   uu(1) = x(1) + signe(x(1))*l;
   u = uu/sqrt(uu'*uu);
end
end
```

The function `signe(z)` returms -1 if $z < 0$, else $+1$.

The function `buildhouse` builds a Householder reflection from a vector uu.

```
function P = buildhouse(v,i)
% This function builds a Householder reflection
%   [I 0 ]
%   [0 PP]
%   from a Householder reflection
%   PP = I - 2uu*uu'
%   where uu = v(i:n)
```

```
%   If uu = 0 then P - I
%

n = size(v,1);
if v(i:n) == zeros(n - i + 1,1)
   P = eye(n);
else
   PP = eye(n - i + 1) - 2*v(i:n)*v(i:n)';
   P = [eye(i-1) zeros(i-1, n - i + 1); zeros(n - i + 1, i - 1) PP];
end
end
```

The function Hessenberg1 computes an upper Hessenberg matrix H and an orthogonal matrix Q such that $A = Q^\top HQ$.

```
function [H, Q] = Hessenberg1(A)
%
%   This function constructs an upper Hessenberg
%   matrix H and an orthogonal matrix Q such that
%   A = Q' H Q

n = size(A,1);
H = A;
Q = eye(n);
for i = 1:n-2
  %  H(i+1:n,i)
    [~,u] = house(H(i+1:n,i));
  %  u
    P =  buildhouse(u,1);
    Q(i+1:n,i:n) = P*Q(i+1:n,i:n);
    H(i+1:n,i:n) = H(i+1:n,i:n) - 2*u*(u')*H(i+1:n,i:n);
    H(1:n,i+1:n) = H(1:n,i+1:n) - 2*H(1:n,i+1:n)*u*(u');
end
end
```

Example 18.2. If

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix},$$

running `Hessenberg1` we find

$$H = \begin{pmatrix} 1.0000 & -5.3852 & 0 & 0 \\ -5.3852 & 15.2069 & -1.6893 & -0.0000 \\ -0.0000 & -1.6893 & -0.2069 & -0.0000 \\ 0 & -0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1.0000 & 0 & 0 & 0 \\ 0 & -0.3714 & -0.5571 & -0.7428 \\ 0 & 0.8339 & 0.1516 & -0.5307 \\ 0 & 0.4082 & -0.8165 & 0.4082 \end{pmatrix}.$$

An important property of (upper) Hessenberg matrices is that if some subdiagonal entry $H_{p+1p} = 0$, then H is of the form

$$H = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix},$$

where both H_{11} and H_{22} are upper Hessenberg matrices (with H_{11} a $p \times p$ matrix and H_{22} a $(n-p)\times(n-p)$ matrix), and the eigenvalues of H are the eigenvalues of H_{11} and H_{22}. For example, in the matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix},$$

if $h_{43} = 0$, then we have the block matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

Then the list of eigenvalues of H is the concatenation of the list of eigenvalues of H_{11} and the list of the eigenvalues of H_{22}. This is easily seen by induction on the dimension of the block H_{11}.

More generally, every upper Hessenberg matrix can be written in such a way that it has diagonal blocks that are Hessenberg blocks whose subdiagonal is not zero.

Definition 18.2. An upper Hessenberg $n \times n$ matrix H is *unreduced* if $h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$. A Hessenberg matrix which is not unreduced is said to be *reduced*.

The following is an example of an 8×8 matrix consisting of three diagonal unreduced Hessenberg blocks:

$$H = \begin{pmatrix} \star & \star & \star & * & * & * & * & * \\ \mathbf{h_{21}} & \star & \star & * & * & * & * & * \\ \mathbf{0} & \mathbf{h_{32}} & \star & * & * & * & * & * \\ 0 & 0 & 0 & \star & \star & \star & * & * \\ 0 & 0 & 0 & \mathbf{h_{54}} & \star & \star & * & * \\ 0 & 0 & 0 & \mathbf{0} & \mathbf{h_{65}} & \star & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & \star & \star \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{h_{87}} & \star \end{pmatrix}.$$

An interesting and important property of unreduced Hessenberg matrices is the following.

Proposition 18.3. *Let H be an $n \times n$ complex or real unreduced Hessenberg matrix. Then every eigenvalue of H is geometrically simple, that is,* $\dim(E_\lambda) = 1$ *for every eigenvalue λ, where E_λ is the eigenspace associated with λ. Furthermore, if H is diagonalizable, then every eigenvalue is simple, that is, H has n distinct eigenvalues.*

Proof. We follow Serre's proof [156] (Proposition 3.26). Let λ be any eigenvalue of H, let $M = \lambda I_n - H$, and let N be the $(n-1) \times (n-1)$ matrix obtained from M by deleting its first row and its last column. Since H is upper Hessenberg, N is a diagonal matrix with entries $-h_{i+1i} \neq 0$, $i = 1, \ldots, n-1$. Thus N is invertible and has rank $n-1$. But a matrix has rank greater than or equal to the rank of any of its submatrices, so $\mathrm{rank}(M) = n-1$, since M is singular. By the rank-nullity theorem, $\mathrm{rank}(\mathrm{Ker}\, N) = 1$, that is, $\dim(E_\lambda) = 1$, as claimed.

If H is diagonalizable, then the sum of the dimensions of the eigenspaces is equal to n, which implies that the eigenvalues of H are distinct. $\square$

As we said earlier, a case where Theorem 18.1 applies is the case where A is a symmetric (or Hermitian) positive definite matrix. This follows from two facts.

The first fact is that if A is Hermitian (or symmetric in the real case), then it is easy to show that the Hessenberg matrix similar to A is a Hermitian (or symmetric in real case) *tridiagonal matrix*. The conversion method is also more efficient. Here is an example of a symmetric tridiagonal matrix consisting of three unreduced blocks:

$$H = \begin{pmatrix} \alpha_1 & \beta_1 & \mathbf{0} & 0 & 0 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & \beta_2 & \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_4 & \beta_4 & \mathbf{0} & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & \alpha_5 & \beta_5 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0} & \beta_5 & \alpha_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_7 & \beta_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_7 & \alpha_8 \end{pmatrix}.$$

Thus the problem of finding the eigenvalues of a symmetric (or Hermitian) matrix reduces to the problem of finding the eigenvalues of a symmetric (resp. Hermitian) tridiagonal matrix, and this can be done much more efficiently.

The second fact is that if H is an upper Hessenberg matrix and if it is diagonalizable, then there is an invertible matrix P such that $H = P\Lambda P^{-1}$ with Λ a diagonal matrix consisting of the eigenvalues of H, such that P^{-1} has an LU-decomposition; see Serre [156] (Theorem 13.3).

As a consequence, since any symmetric (or Hermitian) tridiagonal matrix is a block diagonal matrix of unreduced symmetric (resp. Hermitian) tridiagonal matrices, by Proposition 18.3, we see that the QR algorithm applied to a tridiagonal matrix which is symmetric (or Hermitian) positive definite converges to a diagonal matrix consisting of its eigenvalues. Let us record this important fact.

Theorem 18.4. *Let H be a symmetric (or Hermitian) positive definite tridiagonal matrix. If H is unreduced, then the QR algorithm converges to a diagonal matrix consisting of the eigenvalues of H.*

Since every symmetric (or Hermitian) positive definite matrix is similar to a tridiagonal symmetric (resp. Hermitian) positive definite matrix, we deduce that we have a method for finding the eigenvalues of a symmetric (resp. Hermitian) positive definite matrix (more accurately, to find approximations as good as we want for these eigenvalues).

If A is a symmetric (or Hermitian) matrix, since its eigenvalues are real, for some $\mu > 0$ large enough (pick $\mu > \rho(A)$), $A + \mu I$ is symmetric (resp. Hermitan) positive definite, so we can apply the QR algorithm to an upper Hessenberg matrix similar to $A + \mu I$ to find its eigenvalues, and then the eigenvalues of A are obtained by subtracting μ.

The problem of finding the eigenvalues of a symmetric matrix is discussed extensively in Parlett [135], one of the best references on this topic.

The upper Hessenberg form also yields a way to handle singular matrices. First, checking the proof of Proposition 14.21 that an $n \times n$ complex matrix A (possibly singular) can be factored as $A = QR$ where Q is a unitary matrix which is a product of Householder reflections and R is upper triangular, it is easy to see that if A is upper Hessenberg, then Q is also upper Hessenberg. If H is an unreduced upper Hessenberg matrix, since Q is upper Hessenberg and R is upper triangular, we have $h_{i+1i} = q_{i+1i}r_{ii}$ for $i = 1\ldots, n-1$, and since H is unreduced, $r_{ii} \neq 0$ for $i = 1, \ldots, n-1$. Consequently H is singular iff $r_{nn} = 0$. Then the matrix RQ is a matrix whose last row consists of zero's thus we can deflate the problem by considering the $(n-1) \times (n-1)$ unreduced Hessenberg matrix obtained by deleting the last row and the last column. After finitely many steps (not larger that the multiplicity of the eigenvalue 0), there remains an invertible unreduced Hessenberg matrix. As an alternative, see Serre [156] (Chapter 13, Section 13.3.2).

As is, the QR algorithm, although very simple, is quite inefficient for several reasons. In the next section, we indicate how to make the method more efficient. This involves a lot of work and we only discuss the main ideas at a high level.

18.3 Making the QR Method More Efficient Using Shifts

To improve efficiency and cope with pairs of complex conjugate eigenvalues in the case of real matrices, the following steps are taken:

(1) Initially reduce the matrix A to upper Hessenberg form, as $A = UHU^*$. Then apply the QR-algorithm to H (actually, to its unreduced Hessenberg blocks). It is easy to see that the matrices H_k produced by the QR algorithm remain upper Hessenberg.

(2) To accelerate convergence, use *shifts*, and to deal with pairs of complex conjugate eigenvalues, use *double shifts*.

(3) Instead of computing a QR-factorization explicitly while doing a shift, perform an *implicit shift* which computes $A_{k+1} = Q_k^* A_k Q_k$ without having to compute a QR-factorization (of $A_k - \sigma_k I$), and similarly in the case of a double shift. This is the most intricate modification of the basic QR algorithm and we will not discuss it here. This method is usually referred as *bulge chasing*. Details about this technique for real matrices can be found in Demmel [48] (Section 4.4.8) and Golub and Van Loan [80] (Section 7.5). Watkins discusses the QR algorithm with shifts as a bulge chasing method in the more general case of complex matrices [187, 188].

Let us repeat an important remark made in the previous section. If we start with a matrix H in upper Hessenberg form, if at any stage of the QR algorithm we find that some subdiagonal entry $(H_k)_{p+1p} = 0$ *or is very small*, then H_k is of the form

$$H_k = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix},$$

where both H_{11} and H_{22} are upper Hessenberg matrices (with H_{11} a $p \times p$ matrix and H_{22} a $(n-p) \times (n-p)$ matrix), and the eigenvalues of H_k are the eigenvalues of H_{11} and H_{22}. For example, in the matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & h_{43} & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix},$$

if $h_{43} = 0$, then we have the block matrix

$$H = \begin{pmatrix} * & * & * & * & * \\ h_{21} & * & * & * & * \\ 0 & h_{32} & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & h_{54} & * \end{pmatrix}.$$

Then we can recursively apply the QR algorithm to H_{11} and H_{22}.

In particular, if $(H_k)_{nn-1} = 0$ or is very small, then $(H_k)_{nn}$ is a good approximation of an eigenvalue, so we can delete the last row and the last column of H_k and apply the QR algorithm to this submatrix. This process is called *deflation*. If $(H_k)_{n-1n-2} = 0$ or is very small, then the 2×2 "corner block"

$$\begin{pmatrix} (H_k)_{n-1n-1} & (H_k)_{n-1n} \\ (H_k)_{nn-1} & (H_k)_{nn} \end{pmatrix}$$

appears, and its eigenvalues can be computed immediately by solving a quadratic equation. Then we deflate H_k by deleting its last two rows and its last two columns and apply the QR algorithm to this submatrix.

Thus it would seem desirable to modify the basic QR algorithm so that the above situations arises, and this is what shifts are designed for. More precisely, under the hypotheses of Theorem 18.1, it can be shown (see Ciarlet [41], Section 6.3) that the entry $(A_k)_{ij}$ with $i > j$ converges to 0 as $|\lambda_i/\lambda_j|^k$ converges to 0. Also, if we let r_i be defined by

$$r_1 = \left|\frac{\lambda_2}{\lambda_1}\right|, \quad r_i = \max\left\{\left|\frac{\lambda_i}{\lambda_{i-1}}\right|, \left|\frac{\lambda_{i+1}}{\lambda_i}\right|\right\}, \quad 2 \leq i \leq n-1, \quad r_n = \left|\frac{\lambda_n}{\lambda_{n-1}}\right|,$$

then there is a constant C (independent of k) such that

$$|(A_k)_{ii} - \lambda_i| \leq C r_i^k, \quad 1 \leq i \leq n.$$

In particular, if H is upper Hessenberg, then the entry $(H_k)_{i+1i}$ converges to 0 as $|\lambda_{i+1}/\lambda_i|^k$ converges to 0. Thus if we pick σ_k close to λ_i, we expect that $(H_k - \sigma_k I)_{i+1i}$ converges to 0 as $|(\lambda_{i+1} - \sigma_k)/(\lambda_i - \sigma_k)|^k$ converges to 0, and this ratio is much smaller than 1 as σ_k is closer to λ_i. Typically, we apply a shift to accelerate convergence to λ_n (so $i = n-1$). In this case, both $(H_k - \sigma_k I)_{nn-1}$ and $|(H_k - \sigma_k I)_{nn} - \lambda_n|$ converge to 0 as $|(\lambda_n - \sigma_k)/(\lambda_{n-1} - \sigma_k)|^k$ converges to 0.

A *shift* is the following modified QR-steps (switching back to an arbitrary matrix A, since the shift technique applies in general). Pick some σ_k, hopefully close to some eigenvalue of A (in general, λ_n), and QR-factor $A_k - \sigma_k I$ as

$$A_k - \sigma_k I = Q_k R_k,$$

and then form

$$A_{k+1} = R_k Q_k + \sigma_k I.$$

Since

$$\begin{aligned} A_{k+1} &= R_k Q_k + \sigma_k I \\ &= Q_k^* Q_k R_k Q_k + Q_k^* Q_k \sigma_k \\ &= Q_k^* (Q_k R_k + \sigma_k I) Q_k \\ &= Q_k^* A_k Q_k, \end{aligned}$$

A_{k+1} is similar to A_k, as before. If A_k is upper Hessenberg, then it is easy to see that A_{k+1} is also upper Hessenberg.

If A is upper Hessenberg and if σ_i is exactly equal to an eigenvalue, then $A_k - \sigma_k I$ is singular, and forming the QR-factorization will detect that R_k has some diagonal entry equal to 0. Assuming that the QR-algorithm returns $(R_k)_{nn} = 0$ (if not, the argument is easily adapted), then the last row of $R_k Q_k$ is 0, so the last row of $A_{k+1} = R_k Q_k + \sigma_k I$ ends with σ_k (all other entries being zero), so we are in the case where we can deflate A_k (and σ_k is indeed an eigenvalue).

The question remains, what is a good choice for the shift σ_k?

Assuming again that H is in upper Hessenberg form, it turns out that when $(H_k)_{nn-1}$ is small enough, then a good choice for σ_k is $(H_k)_{nn}$. In fact, the rate of convergence is quadratic, which means roughly that the number of correct digits doubles at every iteration. The reason is that shifts are related to another method known as inverse iteration, and such a method converges very fast. For further explanations about this connection, see Demmel [48] (Section 4.4.4) and Trefethen and Bau [176] (Lecture 29).

One should still be cautious that the QR method with shifts does not necessarily converge, and that our convergence proof no longer applies, because instead of having the identity $A^k = P_k \mathcal{R}_k$, we have

$$(A - \sigma_k I) \cdots (A - \sigma_2 I)(A - \sigma_1 I) = P_k \mathcal{R}_k.$$

Of course, the QR algorithm loops immediately when applied to an orthogonal matrix A. This is also the case when A is symmetric but not positive definite. For example, both the QR algorithm and the QR algorithm with shifts loop on the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In the case of symmetric matrices, Wilkinson invented a shift which helps the QR algorithm with shifts to make progress. Again, looking at the lower corner of A_k, say

$$B = \begin{pmatrix} a_{n-1} & b_{n-1} \\ b_{n-1} & a_n \end{pmatrix},$$

the *Wilkinson shift* picks the eigenvalue of B closer to a_n. If we let

$$\delta = \frac{a_{n-1} - a_n}{2},$$

it is easy to see that the eigenvalues of B are given by

$$\lambda = \frac{a_n + a_{n-1}}{2} \pm \sqrt{\delta^2 + b_{n-1}^2}.$$

It follows that

$$\lambda - a_n = \delta \pm \sqrt{\delta^2 + b_{n-1}^2},$$

and from this it is easy to see that the eigenvalue closer to a_n is given by

$$\mu = a_n - \frac{\mathrm{sign}(\delta) b_{n-1}^2}{(|\delta| + \sqrt{\delta^2 + b_{n-1}^2})}.$$

If $\delta = 0$, then we pick arbitrarily one of the two eigenvalues. Observe that the Wilkinson shift applied to the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is either $+1$ or -1, and in one step, deflation occurs and the algorithm terminates successfully.

We now discuss double shifts, which are intended to deal with pairs of complex conjugate eigenvalues.

Let us assume that A is a real matrix. For any complex number σ_k with nonzero imaginary part, a *double shift* consists of the following steps:

$$\begin{aligned} A_k - \sigma_k I &= Q_k R_k \\ A_{k+1} &= R_k Q_k + \sigma_k I \\ A_{k+1} - \overline{\sigma}_k I &= Q_{k+1} R_{k+1} \\ A_{k+2} &= R_{k+1} Q_{k+1} + \overline{\sigma}_k I. \end{aligned}$$

From the computation made for a single shift, we have $A_{k+1} = Q_k^* A_k Q_k$ and $A_{k+2} = Q_{k+1}^* A_{k+1} Q_{k+1}$, so we obtain

$$A_{k+2} = Q_{k+1}^* Q_k^* A_k Q_k Q_{k+1}.$$

The matrices Q_k are complex, so we would expect that the A_k are also complex, but remarkably we can keep the products $Q_k Q_{k+1}$ real, and so the A_k also real. This is highly desirable to avoid complex arithmetic, which is more expensive.

Observe that since

$$Q_{k+1} R_{k+1} = A_{k+1} - \overline{\sigma}_k I = R_k Q_k + (\sigma_k - \overline{\sigma}_k) I,$$

we have

$$\begin{aligned} Q_k Q_{k+1} R_{k+1} R_k &= Q_k (R_k Q_k + (\sigma_k - \overline{\sigma}_k) I) R_k \\ &= Q_k R_k Q_k R_k + (\sigma_k - \overline{\sigma}_k) Q_k R_k \\ &= (A_k - \sigma_k I)^2 + (\sigma_k - \overline{\sigma}_k)(A_k - \sigma_k I) \\ &= A_k^2 - 2(\Re \sigma_k) A_k + |\sigma_k|^2 I. \end{aligned}$$

If we assume by induction that matrix A_k is real (with $k = 2\ell+1, \ell \geq 0$), then the matrix $S = A_k^2 - 2(\Re\sigma_k)A_k + |\sigma_k|^2 I$ is also real, and since Q_kQ_{k+1} is unitary and $R_{k+1}R_k$ is upper triangular, we see that

$$S = Q_kQ_{k+1}R_{k+1}R_k$$

is a QR-factorization of the real matrix S, thus Q_kQ_{k+1} and $R_{k+1}R_k$ can be chosen to be real matrices, in which case $(Q_kQ_{k+1})^*$ is also real, and thus

$$A_{k+2} = Q_{k+1}^*Q_k^*A_kQ_kQ_{k+1} = (Q_kQ_{k+1})^*A_kQ_kQ_{k+1}$$

is real. Consequently, if $A_1 = A$ is real, then $A_{2\ell+1}$ is real for all $\ell \geq 0$.

The strategy that consists in picking σ_k and $\overline{\sigma}_k$ as the complex conjugate eigenvalues of the corner block

$$\begin{pmatrix} (H_k)_{n-1n-1} & (H_k)_{n-1n} \\ (H_k)_{nn-1} & (H_k)_{nn} \end{pmatrix}$$

is called the *Francis shift* (here we are assuming that A has be reduced to upper Hessenberg form).

It should be noted that there are matrices for which neither a shift by $(H_k)_{nn}$ nor the Francis shift works. For instance, the permutation matrix

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

has eigenvalues $e^{i2\pi/3}, e^{i4\pi/3}, +1$, and neither of the above shifts apply to the matrix

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

However, a shift by 1 does work. There are other kinds of matrices for which the QR algorithm does not converge. Demmel gives the example of matrices of the form

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & h & 0 \\ 0 & -h & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

where h is small.

Algorithms implementing the QR algorithm with shifts and double shifts perform "exceptional" shifts every 10 shifts. Despite the fact that the QR algorithm has been perfected since the 1960's, it is still an open problem to find a shift strategy that ensures convergence of all matrices.

Implicit shifting is based on a result known as the *implicit Q theorem*. This theorem says that if A is reduced to upper Hessenberg form as $A = UHU^*$ and if H is unreduced

($h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$), then the columns of index $2, \ldots, n$ of U are determined by the first column of U up to sign; see Demmel [48] (Theorem 4.9) and Golub and Van Loan [80] (Theorem 7.4.2) for the proof in the case of real matrices. Actually, the proof is not difficult and will be the object of a homework exercise. In the case of a single shift, an implicit shift generates $A_{k+1} = Q_k^* A_k Q_k$ without having to compute a QR-factorization of $A_k - \sigma_k I$. For real matrices, this is done by applying a sequence of Givens rotations which perform a bulge chasing process (a Givens rotation is an orthogonal block diagonal matrix consisting of a single block which is a 2D rotation, the other diagonal entries being equal to 1). Similarly, in the case of a double shift, $A_{k+2} = (Q_k Q_{k+1})^* A_k Q_k Q_{k+1}$ is generated without having to compute the QR-factorizations of $A_k - \sigma_k I$ and $A_{k+1} - \overline{\sigma}_k I$. Again, $(Q_k Q_{k+1})^* A_k Q_k Q_{k+1}$ is generated by applying some simple orthogonal matrices which perform a bulge chasing process. See Demmel [48] (Section 4.4.8) and Golub and Van Loan [80] (Section 7.5) for further explanations regarding implicit shifting involving bulge chasing in the case of real matrices. Watkins [187, 188] discusses bulge chasing in the more general case of complex matrices.

The `Matlab` function for finding the eigenvalues and the eigenvectors of a matrix A is `eig` and is called as `[U, D] = eig(A)`. It is implemented using an optimized version of the QR-algorithm with implicit shifts.

If the dimension of the matrix A is very large, we can find approximations of some of the eigenvalues of A by using a truncated version of the reduction to Hessenberg form due to Arnoldi in general and to Lanczos in the symmetric (or Hermitian) tridiagonal case.

18.4 Krylov Subspaces; Arnoldi Iteration

In this section, we denote the dimension of the square real or complex matrix A by m rather than n, to make it easier for the reader to follow Trefethen and Bau exposition [176], which is particularly lucid.

Suppose that the $m \times m$ matrix A has been reduced to the upper Hessenberg form H, as $A = UHU^*$. For any $n \leq m$ (typically much smaller than m), consider the $(n+1) \times n$ upper left block

$$\widetilde{H}_n = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{nn-1} & h_{nn} \\ 0 & \cdots & 0 & 0 & h_{n+1n} \end{pmatrix}$$

of H, and the $n \times n$ upper Hessenberg matrix H_n obtained by deleting the last row of $\widetilde{H}_n$,

$$H_n = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{nn-1} & h_{nn} \end{pmatrix}.$$

If we denote by U_n the $m \times n$ matrix consisting of the first n columns of U, denoted $u_1, \ldots, u_n$, then the matrix consisting of the first n columns of the matrix $UH = AU$ can be expressed as

$$AU_n = U_{n+1}\widetilde{H}_n. \tag{$*_1$}$$

It follows that the nth column of this matrix can be expressed as

$$Au_n = h_{1n}u_1 + \cdots + h_{nn}u_n + h_{n+1n}u_{n+1}. \tag{$*_2$}$$

Since $(u_1, \ldots, u_n)$ form an orthonormal basis, we deduce from $(*_2)$ that

$$\langle u_j, Au_n \rangle = u_j^* A u_n = h_{jn}, \quad j = 1, \ldots, n. \tag{$*_3$}$$

Equations $(*_2)$ and $(*_3)$ show that U_{n+1} and $\widetilde{H}_n$ can be computed iteratively using the following algorithm due to Arnoldi, known as *Arnoldi iteration*:

Given an arbitrary nonzero vector $b \in \mathbb{C}^m$, let $u_1 = b/\left\|b\right\|$;
for $n = 1, 2, 3, \ldots$ **do**
  $z := Au_n$;
  for $j = 1$ **to** n **do**
    $h_{jn} := u_j^* z$;
    $z := z - h_{jn}u_j$
  endfor
  $h_{n+1n} := \|z\|$;
  if $h_{n+1n} = 0$ **quit**
  $u_{n+1} = z/h_{n+1n}$

When $h_{n+1n} = 0$, we say that we have a *breakdown* of the Arnoldi iteration.

Arnoldi iteration is an algorithm for producing the $n \times n$ Hessenberg submatrix H_n of the full Hessenberg matrix H consisting of its first n rows and n columns (the first n columns of U are also produced), not using Householder matrices.

As long as $h_{j+1j} \neq 0$ for $j = 1, \ldots, n$, Equation $(*_2)$ shows by an easy induction that u_{n+1} belong to the span of $(b, Ab, \ldots, A^n b)$, and obviously Au_n belongs to the span of $(u_1, \ldots, u_{n+1})$, and thus the following spaces are identical:

$$\mathrm{Span}(b, Ab, \ldots, A^n b) = \mathrm{Span}(u_1, \ldots, u_{n+1}).$$

The space $\mathcal{K}_n(A, b) = \mathrm{Span}(b, Ab, \ldots, A^{n-1}b)$ is called a *Krylov subspace*. We can view Arnoldi's algorithm as the construction of an orthonormal basis for $\mathcal{K}_n(A, b)$. It is a sort of Gram–Schmidt procedure.

Equation $(*_2)$ shows that if K_n is the $m \times n$ matrix whose columns are the vectors $(b, Ab, \ldots, A^{n-1}b)$, then there is a $n \times n$ upper triangular matrix R_n such that

$$K_n = U_n R_n. \tag{$*_4$}$$

The above is called a *reduced QR factorization* of K_n.

Since $(u_1, \ldots, u_n)$ is an orthonormal system, the matrix $U_n^* U_{n+1}$ is the $n \times (n+1)$ matrix consisting of the identity matrix I_n plus an extra column of 0's, so $U_n^* U_{n+1} \widetilde{H}_n = U_n^* A U_n$ is obtained by deleting the last row of $\widetilde{H}_n$, namely H_n, and so

$$U_n^* A U_n = H_n. \tag{$*_5$}$$

We summarize the above facts in the following proposition.

Proposition 18.5. *If Arnoldi iteration run on an $m \times m$ matrix A starting with a nonzero vector $b \in \mathbb{C}^m$ does not have a breakdown at stage $n \leq m$, then the following properties hold:*

(1) If K_n is the $m \times n$ Krylov matrix associated with the vectors $(b, Ab, \ldots, A^{n-1}b)$ and if U_n is the $m \times n$ matrix of orthogonal vectors produced by Arnoldi iteration, then there is a QR-factorization

$$K_n = U_n R_n,$$

for some $n \times n$ upper triangular matrix R_n.

(2) The $m \times n$ upper Hessenberg matrices H_n produced by Arnoldi iteration are the projection of A onto the Krylov space $\mathcal{K}_n(A, b)$, that is,

$$H_n = U_n^* A U_n.$$

(3) The successive iterates are related by the formula

$$AU_n = U_{n+1} \widetilde{H}_n.$$

Remark: If Arnoldi iteration has a breakdown at stage n, that is, $h_{n+1} = 0$, then we found the first unreduced block of the Hessenberg matrix H. It can be shown that the eigenvalues of H_n are eigenvalues of A. So a breakdown is actually a good thing. In this case, we can pick some new nonzero vector u_{n+1} orthogonal to the vectors $(u_1, \ldots, u_n)$ as a new starting vector and run Arnoldi iteration again. Such a vector exists since the $(n+1)$th column of U works. So repeated application of Arnoldi yields a full Hessenberg reduction of A. However, this is not what we are after, since m is very large an we are only interested in a "small" number of eigenvalues of A.

There is another aspect of Arnoldi iteration, which is that it solves an optimization problem involving polynomials of degree n. Let $\mathcal{P}^n$ denote the set of (complex) monic polynomials of degree n, that is, polynomials of the form

$$p(z) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0 \quad (c_i \in \mathbb{C}).$$

For any $m \times m$ matrix A, we write

$$p(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I.$$

The following result is proven in Trefethen and Bau [176] (Lecture 34, Theorem 34.1).

Theorem 18.6. *If Arnoldi iteration run on an $m \times m$ matrix A starting with a nonzero vector b does not have a breakdown at stage $n \leq m$, then there is a unique polynomial $p \in \mathcal{P}^n$ such that $\|p(A)b\|_2$ is minimum, namely the characteristic polynomial* $\det(zI - H_n)$ *of H_n.*

Theorem 18.6 can be viewed as the "justification" for a method to find some of the eigenvalues of A (say $n \ll m$ of them). Intuitively, the closer the roots of the characteristic polynomials of H_n are to the eigenvalues of A, the smaller $\|p(A)b\|_2$ should be, and conversely. In the extreme case where $m = n$, by the Cayley–Hamilton theorem, $p(A) = 0$ (where p is the characteristic polynomial of A), so this idea is plausible, but this is far from constituting a proof (also, b should have nonzero coordinates in all directions associated with the eigenvalues).

The method known as the *Rayleigh–Ritz method* is to run Arnoldi iteration on A and some $b \neq 0$ chosen at random for $n \ll m$ steps before or until a breakdown occurs. Then run the QR algorithm with shifts on H_n. The eigenvalues of the Hessenberg matrix H_n may then be considered as approximations of the eigenvalues of A. The eigenvalues of H_n are called *Arnoldi estimates* or *Ritz values*. One has to be cautious because H_n is a truncated version of the full Hessenberg matrix H, so not all of the Ritz values are necessarily close to eigenvalues of A. It has been observed that the eigenvalues that are found first are the *extreme* eigenvalues of A, namely those close to the boundary of the spectrum of A plotted in $\mathbb{C}$. So if A has real eigenvalues, the largest and the smallest eigenvalues appear first as Ritz values. In many problems where eigenvalues occur, the extreme eigenvalues are the one that need to be computed. Similarly, the eigenvectors of H_n may be considered as approximations of eigenvectors of A.

The `Matlab` function `eigs` is based on the computation of Ritz values. It computes the six eigenvalues of largest magnitude of a matrix A, and the call is `[V, D] = eigs(A)`. More generally, to get the top k eigenvalues, use `[V, D] = eigs(A, k)`.

In the absence of rigorous theorems about error estimates, it is hard to make the above statements more precise; see Trefethen and Bau [176] (Lecture 34) for more on this subject.

However, if A is a symmetric (or Hermitian) matrix, then H_n is a symmetric (resp. Hermitian) tridiagonal matrix and more precise results can be shown; see Demmel [48] (Chapter 7, especially Section 7.2). We will consider the symmetric (and Hermitan) case in the next section, but first we show how Arnoldi iteration can be used to find approximations for the solution of a linear system $Ax = b$ where A is invertible but of very large dimension m.

18.5 GMRES

Suppose A is an invertible $m \times m$ matrix and let b be a nonzero vector in $\mathbb{C}^m$. Let $x_0 = A^{-1}b$, the unique solution of $Ax = b$. It is not hard to show that $x_0 \in \mathcal{K}_n(A, b)$ for some $n \leq m$. In fact, there is a unique monic polynomial $p(z)$ of minimal degree $s \leq m$ such that $p(A)b = 0$, so $x_0 \in \mathcal{K}_s(A, b)$. Thus it makes sense to search for a solution of $Ax = b$ in Krylov spaces of dimension $m \leq s$. The idea is to find an approximation $x_n \in \mathcal{K}_n(A, b)$ of x_0 such that $r_n = b - Ax_n$ is minimized, that is, $\|r_n\|_2 = \|b - Ax_n\|_2$ is minimized over $x_n \in \mathcal{K}_n(A, b)$. This minimization problem can be stated as

$$\text{minimize} \quad \|r_n\|_2 = \|Ax_n - b\|_2\,, \quad x_n \in \mathcal{K}_n(A, b).$$

This is a least-squares problem, and we know how to solve it (see Section 23.1). The quantity r_n is known as the *residual* and the method which consists in minimizing $\|r_n\|_2$ is known as GMRES, for *generalized minimal residuals*.

Now since $(u_1, \ldots, u_n)$ is a basis of $\mathcal{K}_n(A, b)$ (since $n \leq s$, no breakdown occurs, except for $n = s$), we may write $x_n = U_n y$, so our minimization problem is

$$\text{minimize} \quad \|AU_n y - b\|_2\,, \quad y \in \mathbb{C}^n.$$

Since by $(*_1)$ of Section 18.4, we have $AU_n = U_{n+1}\widetilde{H}_n$, minimizing $\|AU_n y - b\|_2$ is equivalent to minimizing $\|U_{n+1}\widetilde{H}_n y - b\|_2$ over $\mathbb{C}^m$. Since $U_{n+1}\widetilde{H}_n y$ and b belong to the column space of U_{n+1}, minimizing $\|U_{n+1}\widetilde{H}_n y - b\|_2$ is equivalent to minimizing $\|\widetilde{H}_n y - U_{n+1}^* b\|_2$. However, by construction,

$$U_{n+1}^* b = \|b\|_2 e_1 \in \mathbb{C}^{n+1},$$

so our minimization problem can be stated as

$$\text{minimize} \quad \|\widetilde{H}_n y - \|b\|_2 e_1\|_2, \quad y \in \mathbb{C}^n.$$

The approximate solution of $Ax = b$ is then

$$x_n = U_n y.$$

Starting with $u_1 = b/\|b\|_2$ and with $n = 1$, the GMRES method runs $n \leq s$ Arnoldi iterations to find U_n and $\widetilde{H}_n$, and then runs a method to solve the least squares problem

$$\text{minimize} \quad \|\widetilde{H}_n y - \|b\|_2 e_1\|_2, \quad y \in \mathbb{C}^n.$$

When $\|r_n\|_2 = \|\widetilde{H}_n y - \|b\|_2 e_1\|_2$ is considered small enough, we stop and the approximate solution of $Ax = b$ is then

$$x_n = U_n y.$$

There are ways of improving efficiency of the "naive" version of GMRES that we just presented; see Trefethen and Bau [176] (Lecture 35). We now consider the case where A is a Hermitian (or symmetric) matrix.

18.6 The Hermitian Case; Lanczos Iteration

If A is an $m \times m$ symmetric or Hermitian matrix, then Arnoldi's method is simpler and much more efficient. Indeed, in this case, it is easy to see that the upper Hessenberg matrices H_n are also symmetric (Hermitian respectively), and thus tridiagonal. Also, the eigenvalues of A and H_n are real. It is convenient to write

$$H_n = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix}.$$

The recurrence $(*_2)$ of Section 18.4 becomes the three-term recurrence

$$Au_n = \beta_{n-1} u_{n-1} + \alpha_n u_n + \beta_n u_{n+1}. \qquad (*_6)$$

We also have $\alpha_n = u_n^* A u_n$, so Arnoldi's algorithm becomes the following algorithm known as *Lanczos' algorithm* (or *Lanczos iteration*). The inner loop on j from 1 to n has been eliminated and replaced by a single assignment.

Given an arbitrary nonzero vector $b \in \mathbb{C}^m$, let $u_1 = b/\|b\|$;
for $n = 1, 2, 3, \ldots$ **do**
  $z := Au_n$;
    $\alpha_n := u_n^* z$;
    $z := z - \beta_{n-1} u_{n-1} - \alpha_n u_n$
  $\beta_n := \|z\|$;
  if $\beta_n = 0$ **quit**
  $u_{n+1} = z/\beta_n$

When $\beta_n = 0$, we say that we have a *breakdown* of the Lanczos iteration.

Versions of Proposition 18.5 and Theorem 18.6 apply to Lanczos iteration.

Besides being much more efficient than Arnoldi iteration, Lanczos iteration has the advantage that the *Rayleigh–Ritz method* for finding some of the eigenvalues of A as the eigenvalues of the symmetric (respectively Hermitian) tridiagonal matrix H_n applies, but there are more methods for finding the eigenvalues of symmetric (respectively Hermitian) tridiagonal matrices. Also theorems about error estimates exist. The version of Lanczos iteration given above may run into problems in floating point arithmetic. What happens is that the vectors u_j may lose the property of being orthogonal, so it may be necessary to reorthogonalize them. For more on all this, see Demmel [48] (Chapter 7, in particular Section 7.2-7.4). The version of GMRES using Lanczos iteration is called MINRES.

We close our brief survey of methods for computing the eigenvalues and the eigenvectors of a matrix with a quick discussion of two methods known as power methods.

18.7 Power Methods

Let A be an $m \times m$ complex or real matrix. There are two power methods, both of which yield one eigenvalue and one eigenvector associated with this vector:

(1) *Power iteration.*

(2) *Inverse (power) iteration.*

Power iteration only works if the matrix A has an eigenvalue λ of largest modulus, which means that if $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A, then

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_m| \geq 0.$$

In particular, if A is a real matrix, then λ_1 must be real (since otherwise there are two complex conjugate eigenvalues of the same largest modulus). If the above condition is satisfied, then power iteration yields λ_1 and some eigenvector associated with it. The method is simple enough:

Pick some initial unit vector x^0 and compute the following sequence (x^k), where

$$x^{k+1} = \frac{Ax^k}{\|Ax^k\|}, \quad k \geq 0.$$

We would expect that (x^k) converges to an eigenvector associated with λ_1, but this is not quite correct. The following results are proven in Serre [156] (Section 13.5.1). First assume that $\lambda_1 \neq 0$.

We have

$$\lim_{k \mapsto \infty} \left\|Ax^k\right\| = |\lambda_1|.$$

If A is a complex matrix which has a unique complex eigenvalue λ_1 of largest modulus, then

$$v = \lim_{k \mapsto \infty} \left(\frac{\overline{\lambda_1}}{|\lambda_1|} \right)^k x^k$$

is a unit eigenvector of A associated with λ_1. If λ_1 is real, then

$$v = \lim_{k \mapsto \infty} x^k$$

is a unit eigenvector of A associated with λ_1. Actually some condition on x^0 is needed: x^0 must have a nonzero component in the eigenspace E associated with λ_1 (in any direct sum of $\mathbb{C}^m$ in which E is a summand).

The eigenvalue λ_1 is found as follows. If λ_1 is complex, and if $v_j \neq 0$ is any nonzero coordinate of v, then

$$\lambda_1 = \lim_{k \mapsto \infty} \frac{(Ax^k)_j}{x_j^k}.$$

If λ_1 is real, then we can define the sequence $(\lambda^{(k)})$ by

$$\lambda^{(k+1)} = (x^{k+1})^* A x^{k+1}, \quad k \geq 0,$$

and we have

$$\lambda_1 = \lim_{k \mapsto \infty} \lambda^{(k)}.$$

Indeed, in this case, since $v = \lim_{k \mapsto \infty} x^k$ and v is a unit eigenvector for λ_1, we have

$$\lim_{k \mapsto \infty} \lambda^{(k)} = \lim_{k \mapsto \infty} (x^{k+1})^* A x^{k+1} = v^* A v = \lambda_1 v^* v = \lambda_1.$$

Note that since x^{k+1} is a unit vector, $(x^{k+1})^* A x^{k+1}$ is a Rayleigh ratio.

If A is a Hermitian matrix, then the eigenvalues are real and we can say more about the rate of convergence, which is not great (only linear). For details, see Trefethen and Bau [176] (Lecture 27).

If $\lambda_1 = 0$, then there is some power $\ell < m$ such that $Ax^\ell = 0$.

The *inverse iteration method* is designed to find an eigenvector associated with an eigenvalue λ of A for which we know a good approximation μ.

Pick some initial unit vector x^0 and compute the following sequences (w^k) and (x^k), where w^{k+1} is the solution of the system

$$(A - \mu I) w^{k+1} = x^k \quad \text{equivalently} \quad w^{k+1} = (A - \mu I)^{-1} x^k, \quad k \geq 0,$$

and

$$x^{k+1} = \frac{w^{k+1}}{\|w^{k+1}\|}, \quad k \geq 0.$$

The following result is proven in Ciarlet [41] (Theorem 6.4.1).

Proposition 18.7. *Let A be an $m \times m$ diagonalizable (complex or real) matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$, and let $\lambda = \lambda_\ell$ be an arbitrary eigenvalue of A (not necessary simple). For any μ such that*

$$\mu \neq \lambda \quad \text{and} \quad |\mu - \lambda| < |\mu - \lambda_j| \quad \text{for all } j \neq \ell,$$

if x^0 does not belong to the subspace spanned by the eigenvectors associated with the eigenvalues λ_j with $j \neq \ell$, then

$$\lim_{k \mapsto \infty} \left(\frac{(\lambda - \mu)^k}{|\lambda - \mu|^k} \right) x^k = v,$$

where v is an eigenvector associated with λ. Furthermore, if both λ and μ are real, we have

$$\begin{aligned} \lim_{k \mapsto \infty} x^k &= v & & \text{if } \mu < \lambda, \\ \lim_{k \mapsto \infty} (-1)^k x^k &= v & & \text{if } \mu > \lambda. \end{aligned}$$

Also, if we define the sequence $(\lambda^{(k)})$ by

$$\lambda^{(k+1)} = (x^{k+1})^* A x^{k+1},$$

then

$$\lim_{k \mapsto \infty} \lambda^{(k+1)} = \lambda.$$

The condition of x^0 may seem quite stringent, but in practice, a vector x^0 chosen at random usually satisfies it.

If A is a Hermitian matrix, then we can say more. In particular, the inverse iteration algorithm can be modified to make use of the newly computed $\lambda^{(k+1)}$ instead of μ, and an even faster convergence is achieved. Such a method is called the *Rayleigh quotient iteration*. When it converges (which is for almost all x^0), this method eventually achieves cubic convergence, which is remarkable. Essentially, this means that the number of correct digits is tripled at every iteration. For more details, see Trefethen and Bau [176] (Lecture 27) and Demmel [48] (Section 5.3.2).

18.8 Summary

The main concepts and results of this chapter are listed below:

- QR iteration, QR algorithm.
- Upper Hessenberg matrices.
- Householder matrix.

- Unreduced and reduced Hessenberg matrices.
- Deflation.
- Shift.
- Wilkinson shift.
- Double shift.
- Francis shift.
- Implicit shifting.
- Implicit Q-theorem.
- Arnoldi iteration.
- Breakdown of Arnoldi iteration.
- Krylov subspace.
- Rayleigh–Ritz method.
- Ritz values, Arnoldi estimates.
- Residual.
- GMRES
- Lanczos iteration.
- Power iteration.
- Inverse power iteration.
- Rayleigh ratio.

18.9 Problems

Problem 18.1. Prove Theorem 18.2; see Problem 13.7.

Problem 18.2. Prove that if a matrix A is Hermitian (or real symmetric), then any Hessenberg matrix H similar to A is Hermitian tridiagonal (real symmetric tridiagonal).

Problem 18.3. For any matrix (real or complex) A, if $A = QR$ is a QR-decomposition of A using Householder reflections, prove that if A is upper Hessenberg then so is Q.

Problem 18.4. Prove that if A is upper Hessenberg, then the matrices A_k obtained by applying the QR-algorithm are also upper Hessenberg.

Problem 18.5. Prove the *implicit Q theorem*. This theorem says that if A is reduced to upper Hessenberg form as $A = UHU^*$ and if H is unreduced ($h_{i+1i} \neq 0$ for $i = 1, \ldots, n-1$), then the columns of index $2, \ldots, n$ of U are determined by the first column of U up to sign;

Problem 18.6. Read Section 7.5 of Golub and Van Loan [80] and implement their version of the QR-algorithm with shifts.

Problem 18.7. If an Arnoldi iteration has a breakdown at stage n, that is, $h_{n+1} = 0$, then we found the first unreduced block of the Hessenberg matrix H. Prove that the eigenvalues of H_n are eigenvalues of A.

Problem 18.8. Prove Theorem 18.6.

Problem 18.9. Implement GRMES and test it on some linear systems.

Problem 18.10. State and prove versions of Proposition 18.5 and Theorem 18.6 for the Lanczos iteration.

Problem 18.11. Prove the results about the power iteration method stated in Section 18.7.

Problem 18.12. Prove the results about the inverse power iteration method stated in Section 18.7.

Problem 18.13. Implement and test the power iteration method and the inverse power iteration method.

Problem 18.14. Read Lecture 27 in Trefethen and Bau [176] and implement and test the Rayleigh quotient iteration method.

Chapter 19

Variational Approximation of Boundary-Value Problems; Introduction to the Finite Elements Method

19.1 A One-Dimensional Problem: Bending of a Beam

Consider a beam of unit length supported at its ends in 0 and 1, stretched along its axis by a force P, and subjected to a transverse load $f(x)dx$ per element dx, as illustrated in Figure 19.1.

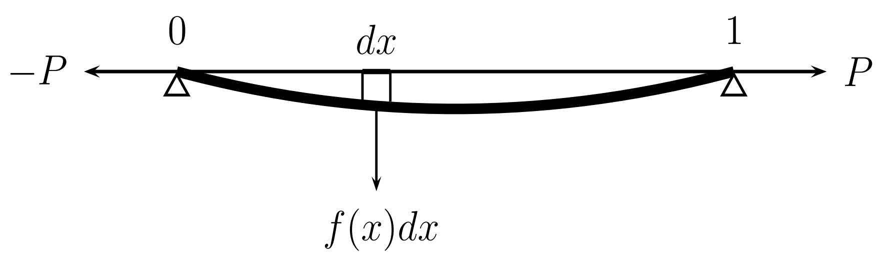


Figure 19.1: Vertical deflection of a beam

The bending moment $u(x)$ at the absissa x is the solution of a boundary problem (BP) of the form

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= \alpha \\ u(1) &= \beta, \end{aligned}$$

where $c(x) = P/(EI(x))$, where E is the Young's modulus of the material of which the beam is made and $I(x)$ is the principal moment of inertia of the cross-section of the beam at the abcissa x, and with $\alpha = \beta = 0$. For this problem, we may assume that $c(x) \geq 0$ for all $x \in [0, 1]$.

Remark: The vertical deflection $w(x)$ of the beam and the bending moment $u(x)$ are related by the equation

$$u(x) = -EI\frac{d^2w}{dx^2}.$$

If we seek a solution $u \in C^2([0, 1])$, that is, a function whose first and second derivatives exist and are continuous, then it can be shown that the problem has a unique solution (assuming c and f to be continuous functions on $[0, 1]$).

Except in very rare situations, this problem has no closed-form solution, so we are led to seek approximations of the solutions.

One one way to proceed is to use the *finite difference method*, where we discretize the problem and replace derivatives by differences. Another way is to use a variational approach. In this approach, we follow a somewhat surprising path in which we come up with a so-called "weak formulation" of the problem, by using a trick based on integrating by parts!

First, let us observe that we can always assume that $\alpha = \beta = 0$, by looking for a solution of the form $u(x) - (\alpha(1 - x) + \beta x)$. This turns out to be crucial when we integrate by parts. There are a lot of subtle mathematical details involved to make what follows rigorous, but here, we will take a "relaxed" approach.

First, we need to specify the space of "weak solutions." This will be the vector space V of continuous functions f on $[0, 1]$, with $f(0) = f(1) = 0$, and which are piecewise continuously differentiable on $[0, 1]$. This means that there is a finite number of points $x_0, \ldots, x_{N+1}$ with $x_0 = 0$ and $x_{N+1} = 1$, such that $f'(x_i)$ is undefined for $i = 1, \ldots, N$, but otherwise f' is defined and continuous on each interval (x_i, x_{i+1}) for $i = 0, \ldots, N$.[1] The space V becomes a Euclidean vector space under the inner product

$$\langle f, g\rangle_V = \int_0^1 (f(x)g(x) + f'(x)g'(x))dx,$$

for all $f, g \in V$. The associated norm is

$$\|f\|_V = \left(\int_0^1 (f(x)^2 + f'(x)^2)dx\right)^{1/2}.$$

Assume that u is a solution of our original boundary problem (BP), so that

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= 0 \\ u(1) &= 0. \end{aligned}$$

[1] We also assume that $f'(x)$ has a limit when x tends to a boundary of (x_i, x_{i+1}).

Multiply the differential equation by any arbitrary *test function* $v \in V$, obtaining

$$-u''(x)v(x) + c(x)u(x)v(x) = f(x)v(x), \tag{$*$}$$

and integrate this equation! We get

$$-\int_0^1 u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx. \tag{$\dagger$}$$

Now, the trick is to use integration by parts on the first term. Recall that

$$(u'v)' = u''v + u'v',$$

and to be careful about discontinuities, write

$$\int_0^1 u''(x)v(x)dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx.$$

Using integration by parts, we have

$$\begin{aligned}\int_{x_i}^{x_{i+1}} u''(x)v(x)dx &= \int_{x_i}^{x_{i+1}} (u'(x)v(x))'dx - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\ &= [u'(x)v(x)]_{x=x_i}^{x=x_{i+1}} - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\ &= u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx.\end{aligned}$$

It follows that

$$\begin{aligned}\int_0^1 u''(x)v(x)dx &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx \\ &= \sum_{i=0}^{N} \left(u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \right) \\ &= u'(1)v(1) - u'(0)v(0) - \int_0^1 u'(x)v'(x)dx.\end{aligned}$$

However, the test function v satisfies the boundary conditions $v(0) = v(1) = 0$ (recall that $v \in V$), so we get

$$\int_0^1 u''(x)v(x)dx = -\int_0^1 u'(x)v'(x)dx.$$

Consequently, the equation ($\dagger$) becomes

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx,$$

or

$$\int_0^1 (u'v' + cuv)dx = \int_0^1 fv dx, \quad \text{for all } v \in V. \tag{$**$}$$

Thus, it is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u, v) = \int_0^1 (u'v' + cuv)dx, \quad \text{for all } u, v \in V,$$

and the linear form $\widetilde{f}\colon V \to \mathbb{R}$ given by

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \text{for all } v \in V.$$

Then, $(**)$ becomes

$$a(u, v) = \widetilde{f}(v), \quad \text{for all } v \in V.$$

We also introduce the *energy function* J given by

$$J(v) = \frac{1}{2}a(v, v) - \widetilde{f}(v) \quad v \in V.$$

Then, we have the following theorem.

Theorem 19.1. *Let u be any solution of the boundary problem (BP).*

(1) Then we have

$$a(u, v) = \widetilde{f}(v), \quad \text{for all } v \in V, \tag{WF}$$

where

$$a(u, v) = \int_0^1 (u'v' + cuv)dx, \quad \textit{for all } u, v \in V,$$

and

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \textit{for all } v \in V.$$

(2) If $c(x) \geq 0$ for all $x \in [0, 1]$, then a function $u \in V$ is a solution of (WF) iff u minimizes $J(v)$, that is,

$$J(u) = \inf_{v \in V} J(v),$$

with

$$J(v) = \frac{1}{2}a(v, v) - \widetilde{f}(v) \quad v \in V.$$

Furthermore, u is unique.

Proof. We already proved (1).

To prove (2), first we show that

$$\|v\|_V^2 \leq 2a(v,v), \quad \text{for all } v \in V.$$

For this, it suffices to prove that

$$\|v\|_V^2 \leq 2\int_0^1 (f'(x))^2 dx, \quad \text{for all } v \in V.$$

However, by Cauchy-Schwarz for functions, for every $x \in [0,1]$, we have

$$|v(x)| = \left|\int_0^x v'(t)dt\right| \leq \int_0^1 |v'(t)|dt \leq \left(\int_0^1 |v'(t)|^2 dt\right)^{1/2},$$

and so

$$\|v\|_V^2 = \int_0^1 ((v(x))^2 + (v'(x))^2)dx \leq 2\int_0^1 (v'(x))^2 dx \leq 2a(v,v),$$

since

$$a(v,v) = \int_0^1 ((v')^2 + cv^2)dx.$$

Next, it is easy to check that

$$J(u+v) - J(u) = a(u,v) - \widetilde{f}(v) + \frac{1}{2}a(v,v), \quad \text{for all } u, v \in V.$$

Then, if u is a solution of (WF), we deduce that

$$J(u+v) - J(u) = \frac{1}{2}a(v,v) \geq \frac{1}{4}\|v\|_V \geq 0 \quad \text{for all } v \in V.$$

since $a(u,v) - \widetilde{f}(v) = 0$ for all $v \in V$. Therefore, J achieves a minimun for u.

We also have

$$J(u+\theta v) - J(u) = \theta(a(u,v) - f(v)) + \frac{\theta^2}{2}a(v,v) \quad \text{for all } \theta \in \mathbb{R},$$

and so $J(u+\theta v) - J(u) \geq 0$ for all $\theta \in \mathbb{R}$. Consequently, if J achieves a minimum for u, then $a(u,v) = \widetilde{f}(v)$, which means that u is a solution of (WF).

Finally, assuming that $c(x) \geq 0$, we claim that if $v \in V$ and $v \neq 0$, then $a(v,v) > 0$. This is because if $a(v,v) = 0$, since

$$\|v\|_V^2 \leq 2a(v,v) \quad \text{for all } v \in V,$$

we would have $\|v\|_V = 0$, that is, $v = 0$. Then, if $v \neq 0$, from

$$J(u+v) - J(u) = \frac{1}{2}a(v,v) \quad \text{for all } v \in V$$

we see that $J(u+v) > J(u)$, so the minimum u is unique $\square$

Theorem 19.1 shows that every solution u of our boundary problem (BP) is a solution (in fact, unique) of the equation (WF).

The equation (WF) is called the *weak form* or *variational equation* associated with the boundary problem. This idea to derive these equations is due to *Ritz and Galerkin*.

Now, the natural question is whether the variational equation (WF) has a solution, and whether this solution, if it exists, is also a solution of the boundary problem (it must belong to $C^2([0,1])$, which is far from obvious). Then, (BP) and (WF) would be equivalent.

Some fancy tools of analysis can be used to prove these assertions. The first difficulty is that the vector space V is not the right space of solutions, because in order for the variational problem to have a solution, it must be complete. So, we must construct a completion of the vector space V. This can be done and we get the *Sobolev space* $H_0^1(0,1)$. Then, the question of the regularity of the "weak solution" can also be tackled.

We will not worry about all this. Instead, let us find *approximations* of the problem (WF). Instead of using the infinite-dimensional vector space V, we consider *finite-dimensional* subspaces V_a (with $\dim(V_a) = n$) of V, and we consider the *discrete problem*:

Find a function $u^{(a)} \in V_a$, such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a. \tag{DWF}$$

Since V_a is finite dimensional (of dimension n), let us pick a basis of functions $(w_1, \ldots, w_n)$ in V_a, so that every function $u \in V_a$ can we written as

$$u = u_1 w_1 + \cdots + u_n w_n.$$

Then, the equation (DWF) holds iff

$$a(u, w_j) = \widetilde{f}(w_j), \quad j = 1, \ldots, n,$$

and by plugging $u_1 w_1 + \cdots + u_n w_n$ for u, we get a system of k linear equations

$$\sum_{i=1}^{n} a(w_i, w_j) u_i = \widetilde{f}(w_j), \quad 1 \leq j \leq n.$$

Because $a(v, v) \geq \frac{1}{2} \|v\|_{V_a}$, the bilinear form a is symmetric positive definite, and thus the matrix $(a(w_i, w_j))$ is symmetric positive definite, and thus invertible. Therefore, (DWF) has a solution given by a *linear system*!

From a practical point of view, we have to compute the integrals

$$a_{ij} = a(w_i, w_j) = \int_0^1 (w_i' w_j' + c w_i w_j) dx,$$

and

$$b_j = \widetilde{f}(w_j) = \int_0^1 f(x) w_j(x) dx.$$

However, if the basis functions are simple enough, this can be done "by hand." Otherwise, numerical integration methods must be used, but there are some good ones.

Let us also remark that the proof of Theorem 19.1 also shows that the unique solution of (DWF) is the unique minimizer of J over all functions in V_a. It is also possible to compare the approximate solution $u^{(a)} \in V_a$ with the exact solution $u \in V$.

Theorem 19.2. *Suppose $c(x) \geq 0$ for all $x \in [0,1]$. For every finite-dimensional subspace V_a ($\dim(V_a) = n$) of V, for every basis $(w_1, \ldots, w_n)$ of V_a, the following properties hold:*

(1) There is a unique function $u^{(a)} \in V_a$ such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a, \tag{DWF}$$

and if $u^{(a)} = u_1 w_1 + \cdots + u_n w_n$, then $\mathbf{u} = (u_1, \ldots, u_n)$ is the solution of the linear system

$$A\mathbf{u} = b, \tag{$*$}$$

with $A = (a_{ij}) = (a(w_i, w_j))$ and $b_j = \widetilde{f}(w_j)$, $1 \leq i, j \leq n$. Furthermore, the matrix $A = (a_{ij})$ is symmetric positive definite.

(2) The unique solution $u^{(a)} \in V_a$ of (DWF) is the unique minimizer of J over V_a, that is,

$$J(u^{(a)}) = \inf_{v \in V_a} J(v),$$

(3) There is a constant C independent of V_a and of the unique solution $u \in V$ of (WF), such that

$$\left\| u - u^{(a)} \right\|_V \leq C \inf_{v \in V_a} \| u - v \|_V .$$

We proved (1) and (2), but we will omit the proof of (3) which can be found in Ciarlet [41].

Let us now give examples of the subspaces V_a used in practice. They usually consist of piecewise polynomial functions.

Pick an integer $N \geq 1$ and subdivide $[0,1]$ into $N+1$ intervals $[x_i, x_{i+1}]$, where

$$x_i = hi, \quad h = \frac{1}{N+1}, \quad i = 0, \ldots, N+1.$$

We will use the following fact: every polynomial $P(x)$ of degree $2m+1$ ($m \geq 0$) is completely determined by its values as well as the values of its first m derivatives at two distinct points $\alpha, \beta \in \mathbb{R}$.

There are various ways to prove this. One way is to use the Bernstein basis, because the kth derivative of a polynomial is given by a formula in terms of its control points. For example, for $m = 1$, every degree 3 polynomial can be written as

$$P(x) = (1-x)^3\, b_0 + 3(1-x)^2 x\, b_1 + 3(1-x)x^2\, b_2 + x^3\, b_3,$$

with $b_0, b_1, b_2, b_3 \in \mathbb{R}$, and we showed that

$$\begin{aligned} P'(0) &= 3(b_1 - b_0) \\ P'(1) &= 3(b_3 - b_2). \end{aligned}$$

Given $P(0)$ and $P(1)$, we determine b_0 and b_3, and from $P'(0)$ and $P'(1)$, we determine b_1 and b_2.

In general, for a polynomial of degree m written as

$$P(x) = \sum_{j=0}^{m} b_j B_j^m(x)$$

in terms of the Bernstein basis $(B_0^m(x), \ldots, B_m^m(x))$ with

$$B_j^m(x) = \binom{m}{j}(1-x)^{m-j}x^j,$$

it can be shown that the kth derivative of P at zero is given by

$$P^{(k)}(0) = m(m-1)\cdots(m-k+1)\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\, b_i\right),$$

and there is a similar formula for $P^{(k)}(1)$.

Actually, we need to use the Bernstein basis of polynomials $B_k^m[r,s]$, where

$$B_j^m[r,s](x) = \binom{m}{j}\left(\frac{s-x}{s-r}\right)^{m-j}\left(\frac{x-r}{s-r}\right)^{j},$$

with $r < s$, in which case

$$P^{(k)}(0) = \frac{m(m-1)\cdots(m-k+1)}{(s-r)^k}\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\, b_i\right),$$

with a similar formula for $P^{(k)}(1)$. In our case, we set $r = x_i, s = x_{i+1}$.

Now, if the $2m+2$ values

$$P(0), P^{(1)}(0), \ldots, P^{(m)}(0), P(1), P^{(1)}(1), \ldots, P^{(m)}(1)$$

are given, we obtain a triangular system that determines uniquely the $2m+2$ control points $b_0, \ldots, b_{2m+1}$.

Recall that $C^m([0,1])$ denotes the set of C^m functions f on $[0,1]$, which means that $f, f^{(1)}, \ldots, f^{(m)}$ exist are are continuous on $[0,1]$.

We define the vector space V_N^m as the subspace of $C^m([0,1])$ consisting of all functions f such that

1. $f(0) = f(1) = 0$.
2. The restriction of f to $[x_i, x_{i+1}]$ is a polynomial of degree $2m+1$, for $i = 0, \ldots, N$.

Observe that the functions in V_N^0 are the piecewise affine functions f with $f(0) = f(1) = 0$; an example is shown in Figure 19.2.

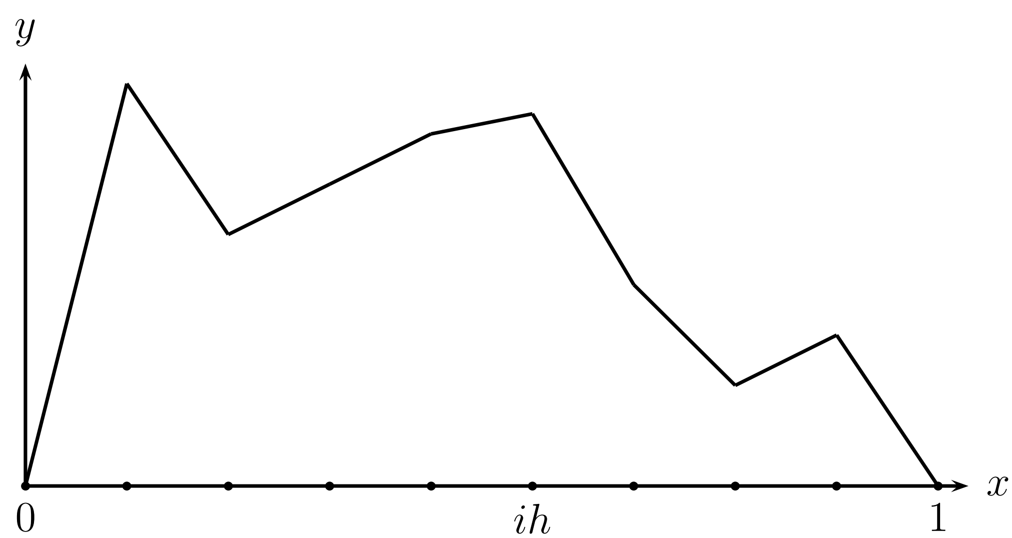


Figure 19.2: A piecewise affine function

This space has dimension N, and a basis consists of the "hat functions" w_i, where the only two nonflat parts of the graph of w_i are the line segments from $(x_{i-1}, 0)$ to $(x_i, 1)$, and from $(x_i, 1)$ to $(x_{i+1}, 0)$, for $i = 1, \ldots, N$, see Figure 19.3.

The basis functions w_i have a small support, which is good because in computing the integrals giving $a(w_i, w_j)$, we find that we get a tridiagonal matrix. They also have the nice property that every function $v \in V_N^0$ has the following expression on the basis (w_i):

$$v(x) = \sum_{i=1}^{N} v(ih) w_i(x), \quad x \in [0,1].$$

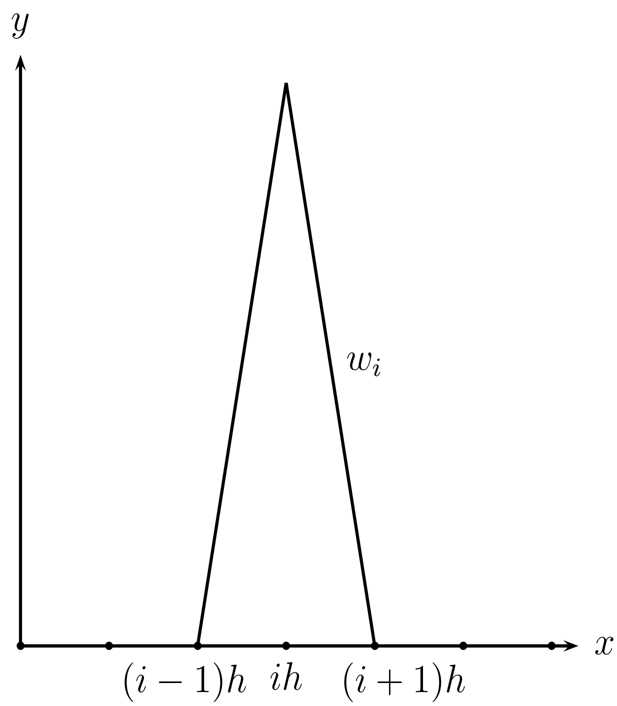


Figure 19.3: A basis "hat function"

In general, it it not hard to see that V_N^m has dimension $mN + 2(m-1)$.

Going back to our problem (the bending of a beam), assuming that c and f are constant functions, it is not hard to show that the linear system $(*)$ becomes

$$\frac{1}{h}\begin{pmatrix} 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & & \\ -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & \\ \ddots & \ddots & \ddots & & \\ & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & \\ & & & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = h\begin{pmatrix} f \\ f \\ \vdots \\ f \\ f \end{pmatrix}.$$

We can also find a basis of $2N+2$ cubic functions for V_N^1 consisting of functions with small support. This basis consists of the N functions w_i^0 and of the $N+2$ functions w_i^1

uniquely determined by the following conditions:

$$\begin{aligned} w_i^0(x_j) &= \delta_{ij}, \quad 1 \le j \le N,\ 1 \le i \le N \\ (w_i^0)'(x_j) &= 0, \quad 0 \le j \le N+1,\ 1 \le i \le N \\ w_i^1(x_j) &= 0, \quad 1 \le j \le N,\ 0 \le i \le N+1 \\ (w_i^1)'(x_j) &= \delta_{ij}, \quad 0 \le j \le N+1,\ 0 \le i \le N+1 \end{aligned}$$

with $\delta_{ij} = 1$ iff $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Some of these functions are displayed in Figure 19.4. The function w_i^0 is given explicitly by

$$w_i^0(x) = \frac{1}{h^3}(x - (i-1)h)^2((2i+1)h - 2x), \quad (i-1)h \le x \le ih,$$

$$w_i^0(x) = \frac{1}{h^3}((i+1)h - x)^2(2x - (2i-1)h), \quad ih \le x \le (i+1)h,$$

for $i = 1, \ldots, N$. The function w_j^1 is given explicitly by

$$w_j^1(x) = -\frac{1}{h^2}(ih - x)(x - (i-1)h)^2, \quad (i-1)h \le x \le ih,$$

and

$$w_j^1(x) = \frac{1}{h^2}((i+1)h - x)^2(x - ih), \quad ih \le x \le (i+1)h,$$

for $j = 0, \ldots, N+1$. Furthermore, for every function $v \in V_N^1$, we have

$$v(x) = \sum_{i=1}^{N} v(ih) w_i^0(x) + \sum_{j=0}^{N+1} v'jih) w_j^1(x), \quad x \in [0,1].$$

If we order these basis functions as

$$w_0^1, w_1^0, w_1^1, w_2^0, w_2^1, \ldots, w_N^0, w_N^1, w_{N+1}^1,$$

we find that if $c = 0$, the matrix A of the system $(*)$ is tridiagonal by blocks, where the blocks are 2×2, 2×1, or 1×2 matrices, and with single entries in the top left and bottom right corner. A different order of the basis vectors would mess up the tridiagonal block structure of A. We leave the details as an exercise.

Let us now take a quick look at a two-dimensional problem, the bending of an elastic membrane.

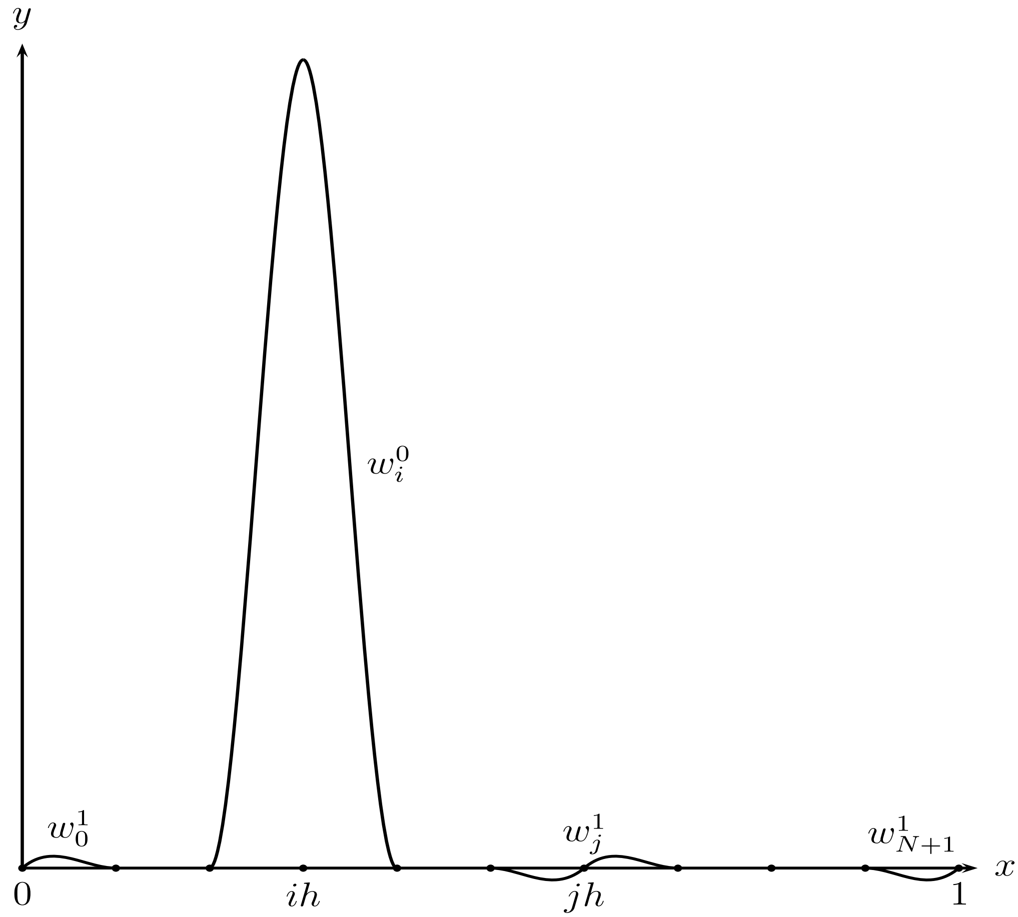


Figure 19.4: The basis functions w_i^0 and w_j^1

19.2 A Two-Dimensional Problem: An Elastic Membrane

Consider an elastic membrane attached to a round contour whose projection on the (x_1, x_2)-plane is the boundary Γ of an open, connected, bounded region Ω in the (x_1, x_2)-plane, as illustrated in Figure 19.5. In other words, we view the membrane as a surface consisting of the set of points (x, z) given by an equation of the form

$$z = u(x),$$

with $x = (x_1, x_2) \in \overline{\Omega}$, where $u\colon \overline{\Omega} \to \mathbb{R}$ is some sufficiently regular function, and we think of $u(x)$ as the vertical displacement of this membrane.

We assume that this membrane is under the action of a vertical force $\tau f(x)dx$ per surface element in the horizontal plane (where τ is the tension of the membrane). The problem is

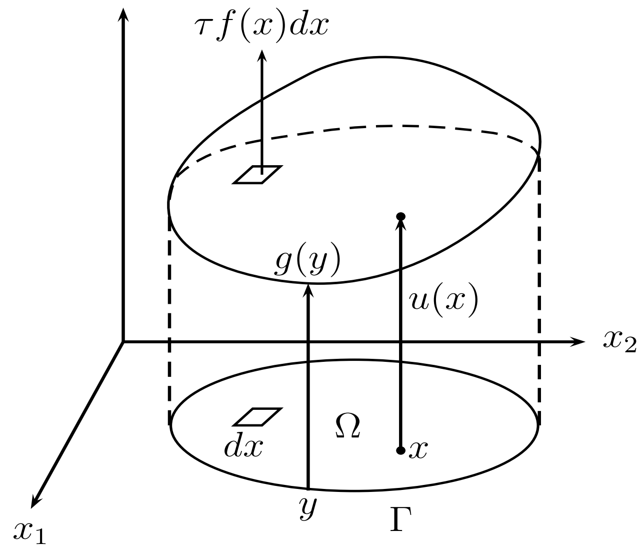


Figure 19.5: An elastic membrane

to find the vertical displacement u as a function of x, for $x \in \overline{\Omega}$. It can be shown (under some assumptions on Ω, Γ, and f), that $u(x)$ is given by a PDE with boundary condition, of the form

$$\begin{aligned} -\Delta u(x) &= f(x), \quad x \in \Omega \\ u(x) &= g(x), \quad x \in \Gamma, \end{aligned}$$

where $g\colon \Gamma \to \mathbb{R}$ represents the height of the contour of the membrane. We are looking for a function u in $C^2(\Omega) \cap C^1(\overline{\Omega})$. The operator Δ is the *Laplacian*, and it is given by

$$\Delta u(x) = \frac{\partial^2 u}{\partial x_1^2}(x) + \frac{\partial^2 u}{\partial x_2^2}(x).$$

This is an example of a *boundary problem*, since the solution u of the PDE must satisfy the condition $u(x) = g(x)$ on the boundary of the domain Ω. The above equation is known as *Poisson's equation*, and when $f = 0$ as *Laplace's equation*.

It can be proved that if the data f, g and Γ are sufficiently smooth, then the problem has a unique solution.

To get a weak formulation of the problem, first we have to make the boundary condition homogeneous, which means that $g(x) = 0$ on Γ. It turns out that g can be extended to the whole of $\overline{\Omega}$ as some sufficiently smooth function $\widehat{h}$, so we can look for a solution of the form $u - \widehat{h}$, but for simplicity, let us assume that the contour of Ω lies in a plane parallel to the

(x_1, x_2)- plane, so that $g = 0$. We let V be the subspace of $C^2(\Omega) \cap C^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ.

As before, we multiply the PDE by a test function $v \in V$, getting

$$-\Delta u(x)v(x) = f(x)v(x),$$

and we "integrate by parts." In this case, this means that we use a version of Stokes formula known as *Green's first identity*, which says that

$$\int_\Omega -\Delta u\, v\, dx = \int_\Omega (\mathrm{grad}\, u) \cdot (\mathrm{grad}\, v)\, dx - \int_\Gamma (\mathrm{grad}\, u) \cdot n\, v d\sigma$$

(where n denotes the outward pointing unit normal to the surface). Because $v = 0$ on Γ, the integral $\int_\Gamma$ drops out, and we get an equation of the form

$$a(u, v) = \widetilde{f}(v) \quad \text{for all } v \in V,$$

where a is the bilinear form given by

$$a(u, v) = \int_\Omega \left(\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx$$

and $\widetilde{f}$ is the linear form given by

$$\widetilde{f}(v) = \int_\Omega f v dx.$$

We get the same equation as in section 19.2, but over a set of functions defined on a two-dimensional domain. As before, we can choose a finite-dimensional subspace V_a of V and consider the discrete problem with respect to V_a. Again, if we pick a basis $(w_1, \ldots, w_n)$ of V_a, a vector $u = u_1 w_1 + \cdots + u_n w_n$ is a solution of the Weak Formulation of our problem iff $\mathbf{u} = (u_1, \ldots, u_n)$ is a solution of the linear system

$$A\mathbf{u} = b,$$

with $A = (a(w_i, w_j))$ and $b = (\widetilde{f}(w_j))$. However, the integrals that give the entries in A and b are much more complicated.

An approach to deal with this problem is the *method of finite elements*. The idea is to also discretize the boundary curve Γ. If we assume that Γ is a *polygonal line*, then we can *triangulate* the domain Ω, and then we consider spaces of functions which are piecewise defined on the triangles of the triangulation of Ω. The simplest functions are piecewise affine and look like tents erected above groups of triangles. Again, we can define base functions with small support, so that the matrix A is tridiagonal by blocks.

The finite element method is a vast subject and it is presented in many books of various degrees of difficulty and obscurity. Let us simply state three important requirements of the finite element method:

1. "Good" triangulations must be found. This in itself is a vast research topic. Delaunay triangulations are good candidates.

2. "Good" spaces of functions must be found; typically piecewise polynomials and splines.

3. "Good" bases consisting of functions will small support must be found, so that integrals can be easily computed and sparse banded matrices arise.

We now consider boundary problems where the solution varies with time.

19.3 Time-Dependent Boundary Problems: The Wave Equation

Consider a homogeneous string (or rope) of constant cross-section, of length L, and stretched (in a vertical plane) between its two ends which are assumed to be fixed and located along the x-axis at $x = 0$ and at $x = L$.

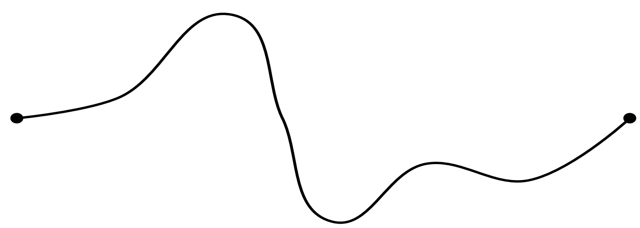


Figure 19.6: A vibrating string

The string is subjected to a transverse force $\tau f(x)dx$ per element of length dx (where τ is the tension of the string). We would like to investigate the small displacements of the string in the vertical plane, that is, how it vibrates.

Thus, we seek a function $u(x, t)$ defined for $t \geq 0$ and $x \in [0, L]$, such that $u(x, t)$ represents the vertical deformation of the string at the abscissa x and at time t.

It can be shown that u must satisfy the following PDE

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L,\ t > 0,$$

with $c = \sqrt{\tau/\rho}$, where ρ is the linear density of the string, known as the *one-dimensional wave equation*.

Furthermore, the initial shape of the string is known at $t = 0$, as well as the distribution of the initial velocities along the string; in other words, there are two functions $u_{i,0}$ and $u_{i,1}$ such that

$$\begin{aligned} u(x,0) &= u_{i,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad 0 \leq x \leq L. \end{aligned}$$

For example, if the string is simply released from its given starting position, we have $u_{i,1} = 0$. Lastly, because the ends of the string are fixed, we must have

$$u(0,t) = u(L,t) = 0, \quad t \geq 0.$$

Consequently, we look for a function $u\colon \mathbb{R}_+ \times [0,L] \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned} \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) &= f(x,t), \quad 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) &= 0, \quad t \geq 0 \quad \text{(boundary condition)}, \\ u(x,0) &= u_{i,0}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}. \end{aligned}$$

This is an example of a *time-dependent boundary-value problem*, with two *initial conditions*.

To simplify the problem, assume that $f = 0$, which amounts to neglecting the effect of gravity. In this case, our PDE becomes

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < L,\ t > 0,$$

Let us try our trick of multiplying by a test function v depending only on x, C^1 on $[0,L]$, and such that $v(0) = v(L) = 0$, and integrate by parts. We get the equation

$$\int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx - c^2 \int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = 0.$$

For the first term, we get

$$\begin{aligned} \int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx &= \int_0^L \frac{\partial^2}{\partial t^2}[u(x,t)v(x)]dx \\ &= \frac{d^2}{dt^2}\int_0^L u(x,t)v(x)dx \\ &= \frac{d^2}{dt^2}\langle u, v\rangle, \end{aligned}$$

where $\langle u, v\rangle$ is the inner product in $L^2([0, L])$. The fact that it is legitimate to move $\partial^2/\partial t^2$ outside of the integral needs to be justified rigorously, but we won't do it here.

For the second term, we get

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = -\left[\frac{\partial u}{\partial x}(x,t)v(x)\right]_{x=0}^{x=L} + \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx,$$

and because $v \in V$, we have $v(0) = v(L) = 0$, so we obtain

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx.$$

Our integrated equation becomes

$$\frac{d^2}{dt^2}\langle u, v\rangle + c^2\int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx = 0, \quad \text{for all } v \in V \quad \text{and all } t \geq 0.$$

It is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u,v) = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{\partial v}{\partial x}(x,t)dx,$$

where, for every $t \in \mathbb{R}_+$, the functions $u(x,t)$ and (v,t) belong to V. Actually, we have to replace V by the subspace of the Sobolev space $H_0^1(0, L)$ consisting of the functions such that $v(0) = v(L) = 0$. Then, the weak formulation (variational formulation) of our problem is this:

Find a function $u \in V$ such that

$$\begin{aligned}
\frac{d^2}{dt^2}\langle u, v\rangle + a(u,v) = 0, \qquad &\text{for all } v \in V \quad \text{and all } t \geq 0 \\
u(x,0) = u_{i,0}(x), \quad &0 \leq x \leq L \quad \text{(intitial condition)}, \\
\frac{\partial u}{\partial t}(x,0) = u_{i,1}(x), \quad &0 \leq x \leq L \quad \text{(intitial condition)}.
\end{aligned}$$

It can be shown that there is a positive constant $\alpha > 0$ such that

$$a(u,u) \geq \alpha \left\|u\right\|_{H_0^1}^2 \quad \text{for all } v \in V$$

(Poincaré's inequality), which shows that a is positive definite on V. The above method is known as the method of *Rayleigh-Ritz*.

A study of the above equation requires some sophisticated tools of analysis which go far beyond the scope of these notes. Let us just say that there is a countable sequence of solutions with separated variables of the form

$$u_k^{(1)} = \sin\left(\frac{k\pi x}{L}\right)\cos\left(\frac{k\pi ct}{L}\right), \quad u_k^{(2)} = \sin\left(\frac{k\pi x}{L}\right)\sin\left(\frac{k\pi ct}{L}\right), \quad k \in \mathbb{N}_+,$$

called *modes* (or *normal modes*). Complete solutions of the problem are series obtained by combining the normal modes, and they are of the form

$$u(x,t) = \sum_{k=1}^{\infty} \sin\left(\frac{k\pi x}{L}\right)\left(A_k \cos\left(\frac{k\pi ct}{L}\right) + B_k \sin\left(\frac{k\pi ct}{L}\right)\right),$$

where the coefficients A_k, B_k are determined from the Fourier series of $u_{i,0}$ and $u_{i,1}$.

We now consider discrete approximations of our problem. As before, consider a finite dimensional subspace V_a of V and assume that we have approximations $u_{a,0}$ and $u_{a,1}$ of $u_{i,0}$ and $u_{i,1}$. If we pick a basis $(w_1, \ldots, w_n)$ of V_a, then we can write our unknown function $u(x,t)$ as

$$u(x,t) = u_1(t)w_1 + \cdots + u_n(t)w_n,$$

where $u_1, \ldots, u_n$ are functions of t. Then, if we write $\mathbf{u} = (u_1, \ldots, u_n)$, the discrete version of our problem is

$$\begin{aligned} A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0, \\ u(x,0) &= u_{a,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad 0 \leq x \leq L, \end{aligned}$$

where $A = (\langle w_i, w_j \rangle)$ and $K = (a(w_i, w_j))$ are two symmetric matrices, called the *mass matrix* and the *stiffness matrix*, respectively. In fact, because a and the inner product $\langle -, - \rangle$ are positive definite, these matrices are also positive definite.

We have made some progress since we now have a system of ODE's, and we can solve it by analogy with the scalar case. So, we look for solutions of the form $\mathbf{U}\cos\omega t$ (or $\mathbf{U}\sin\omega t$), where $\mathbf{U}$ is an n-dimensional vector. We find that we should have

$$(K - \omega^2 A)\mathbf{U}\cos\omega t = 0,$$

which implies that ω must be a solution of the equation

$$K\mathbf{U} = \omega^2 A\mathbf{U}.$$

Thus, we have to find some λ such that

$$K\mathbf{U} = \lambda A\mathbf{U},$$

a problem known as a *generalized eigenvalue problem*, since the ordinary eigenvalue problem for K is

$$K\mathbf{U} = \lambda\mathbf{U}.$$

Fortunately, because A is SPD, we can reduce this generalized eigenvalue problem to a standard eigenvalue problem. A good way to do so is to use a Cholesky decomposition of A as

$$A = LL^\top,$$

where L is a lower triangular matrix (see Theorem 8.10). Because A is SPD, it is invertible, so L is also invertible, and

$$K\mathbf{U} = \lambda A\mathbf{U} = \lambda LL^\top \mathbf{U}$$

yields

$$L^{-1}K\mathbf{U} = \lambda L^\top \mathbf{U},$$

which can also be written as

$$L^{-1}K(L^\top)^{-1}L^\top \mathbf{U} = \lambda L^\top \mathbf{U}.$$

Then, if we make the change of variable

$$\mathbf{Y} = L^\top \mathbf{U},$$

using the fact $(L^\top)^{-1} = (L^{-1})^\top$, the above equation is equivalent to

$$L^{-1}K(L^{-1})^\top \mathbf{Y} = \lambda \mathbf{Y},$$

a standard eigenvalue problem for the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$. Furthermore, we know from Section 8.8 that since K is SPD and L^{-1} is invertible, the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$ is also SPD.

Consequently, $\widehat{K}$ has positive real eigenvalues $(\omega_1^2, \ldots, \omega_n^2)$ (not necessarily distinct) and it can be diagonalized with respect to an orthonormal basis of eigenvectors, say $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$. Then, since $\mathbf{Y} = L^\top \mathbf{U}$, the vectors

$$\mathbf{U}^i = (L^\top)^{-1}\mathbf{Y}^i, \quad i = 1, \ldots, n,$$

are linearly independent and are solutions of the generalized eigenvalue problem; that is,

$$K\mathbf{U}^i = \omega_i^2 A\mathbf{U}^i, \quad i = 1, \ldots, n.$$

More is true. Because the vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are orthonormal, and because $\mathbf{Y}^i = L^\top \mathbf{U}^i$, from

$$(\mathbf{Y}^i)^\top \mathbf{Y}^j = \delta_{ij},$$

we get

$$(\mathbf{U}^i)^\top LL^\top \mathbf{U}^j = \delta_{ij}, \quad 1 \leq i, j \leq n,$$

and since $A = LL^\top$, this yields

$$(\mathbf{U}^i)^\top A\mathbf{U}^j = \delta_{ij}, \quad 1 \leq i, j \leq n.$$

This suggests defining the functions $U^i \in V_a$ by

$$U^i = \sum_{k=1}^{n} \mathbf{U}^i_k w_k.$$

Then, it immediate to check that

$$a(U^i, U^j) = (\mathbf{U}^i)^\top A \mathbf{U}^j = \delta_{ij},$$

which means that the functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a. The functions $U^i \in V_a$ are called *modes* (or *modal vectors*).

As a final step, let us look again for a solution of our discrete weak formulation of the problem, this time expressing the unknown solution $u(x,t)$ over the modal basis $(U^1, \ldots, U^n)$, say

$$u = \sum_{j=1}^{n} \widetilde{u}_j(t) U^j,$$

where each $\widetilde{u}_j$ is a function of t. Because

$$u = \sum_{j=1}^{n} \widetilde{u}_j(t) U^j = \sum_{j=1}^{n} \widetilde{u}_j(t) \left(\sum_{k=1}^{n} \mathbf{U}^j_k w_k \right) = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} \widetilde{u}_j(t) \mathbf{U}^j_k \right) w_k,$$

if we write $\mathbf{u} = (u_1, \ldots, u_n)$ with $u_k = \sum_{j=1}^{n} \widetilde{u}_j(t) \mathbf{U}^j_k$ for $k = 1, \ldots, n$, we see that

$$\mathbf{u} = \sum_{j=1}^{n} \widetilde{u}_j \mathbf{U}^j,$$

so using the fact that

$$K \mathbf{U}^j = \omega_j^2 A \mathbf{U}^j, \quad j = 1, \ldots, n,$$

the equation

$$A \frac{d^2 \mathbf{u}}{dt^2} + K \mathbf{u} = 0$$

yields

$$\sum_{j=1}^{n} [(\widetilde{u}_j)'' + \omega_j^2 \widetilde{u}_j] A \mathbf{U}^j = 0.$$

Since A is invertible and since $(\mathbf{U}^1, \ldots, \mathbf{U}^n)$ are linearly independent, the vectors $(A\mathbf{U}^1, \ldots, A\mathbf{U}^n)$ are linearly independent, and consequently we get the system of n ODEs'

$$(\widetilde{u}_j)'' + \omega_j^2 \widetilde{u}_j = 0, \quad 1 \leq j \leq n.$$

Each of these equation has a well-known solution of the form

$$\widetilde{u}_j = A_j \cos \omega_j t + B_j \sin \omega_j t.$$

Therefore, the solution of our approximation problem is given by

$$u = \sum_{j=1}^{n} (A_j \cos \omega_j t + B_j \sin \omega_j t) U^j,$$

and the constants A_j, B_j are obtained from the intial conditions

$$\begin{aligned} u(x,0) &= u_{a,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad 0 \leq x \leq L, \end{aligned}$$

by expressing $u_{a,0}$ and $u_{a,1}$ on the modal basis $(U^1, \ldots, U^n)$. Furthermore, the modal functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a.

If we use the vector space V_N^0 of piecewise affine functions, we find that the matrices A and K are familar! Indeed,

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

and

$$K = \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix}.$$

To conclude this section, let us discuss briefly the wave equation for an elastic membrane, as described in Section 19.2. This time, we look for a function $u\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned} \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \Delta u(x,t) = f(x,t), \quad x \in \Omega,\ t > 0, \\ u(x,t) = 0, \quad x \in \Gamma, \quad t \geq 0 \quad \text{(boundary condition)}, \\ u(x,0) = u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)}, \\ \frac{\partial u}{\partial t}(x,0) = u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)}. \end{aligned}$$

Assuming that $f = 0$, we look for solutions in the subspace V of the Sobolev space $H_0^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ. Multiplying by a test function $v \in V$ and using Green's first identity, we get the weak formulation of our problem:

Find a function $u \in V$ such that

$$\begin{aligned} \frac{d^2}{dt^2}\langle u, v\rangle + a(u, v) &= 0, \qquad \text{for all } v \in V \quad \text{and all } t \geq 0 \\ u(x, 0) &= u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)}, \\ \frac{\partial u}{\partial t}(x, 0) &= u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)}, \end{aligned}$$

where $a \colon V \times V \to \mathbb{R}$ is the bilinear form given by

$$a(u, v) = \int_{\Omega} \left(\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} \right) dx,$$

and

$$\langle u, v\rangle = \int_{\Omega} uv dx.$$

As usual, we find approximations of our problem by using finite dimensional subspaces V_a of V. Picking some basis $(w_1, \ldots, w_n)$ of V_a, and triangulating Ω, as before, we obtain the equation

$$\begin{aligned} A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0, \\ u(x, 0) &= u_{a,0}(x), \quad x \in \Gamma, \\ \frac{\partial u}{\partial t}(x, 0) &= u_{a,1}(x), \quad x \in \Gamma, \end{aligned}$$

where $A = (\langle w_i, w_j\rangle)$ and $K = (a(w_i, w_j))$ are two symmetric positive definite matrices.

In principle, the problem is solved, but, it may be difficult to find good spaces V_a, good triangulations of Ω, and good bases of V_a, to be able to compute the matrices A and K, and to ensure that they are sparse.

Chapter 20

Graphs and Graph Laplacians; Basic Facts

In this chapter and the next we present some applications of linear algebra to graph theory. Graphs (undirected and directed) can be defined in terms of various matrices (incidence and adjacency matrices), and various connectivity properties of graphs are captured by properties of these matrices. Another very important matrix is associated with a (undirected) graph: the *graph Laplacian*. The graph Laplacian is symmetric positive definite, and its eigenvalues capture some of the properties of the underlying graph. This is a key fact that is exploited in graph clustering methods, the most powerful being the method of normalized cuts due to Shi and Malik [160]. This chapter and the next constitute an introduction to algebraic and spectral graph theory. We do not discuss normalized cuts, but we discuss graph drawings. Thorough presentations of algebraic graph theory can be found in Godsil and Royle [77] and Chung [39].

We begin with a review of basic notions of graph theory. Even though the graph Laplacian is fundamentally associated with an undirected graph, we review the definition of both directed and undirected graphs. For both directed and undirected graphs, we define the degree matrix D, the incidence matrix B, and the adjacency matrix A. Then we define a *weighted graph*. This is a pair (V, W), where V is a finite set of nodes and W is a $m \times m$ symmetric matrix with nonnegative entries and zero diagonal entries (where $m = |V|$).

For every node $v_i \in V$, the *degree* $d(v_i)$ (or d_i) of v_i is the sum of the weights of the edges adjacent to v_i:

$$d_i = d(v_i) = \sum_{j=1}^{m} w_{i\,j}.$$

The *degree matrix* is the diagonal matrix

$$D = \mathrm{diag}(d_1, \ldots, d_m).$$

The notion of degree is illustrated in Figure 20.1. Then we introduce the (unnormalized)

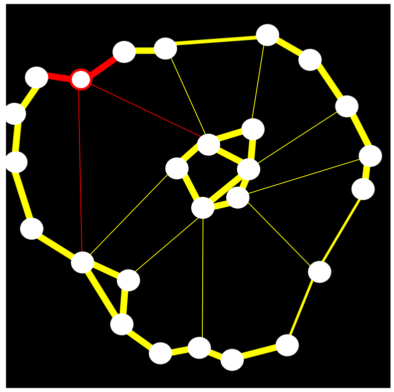


Figure 20.1: Degree of a node.

graph Laplacian L of a directed graph G in an "old-fashion" way, by showing that for any orientation of a graph G,

$$BB^\top = D - A = L$$

is an invariant. We also define the (unnormalized) *graph Laplacian* L of a weighted graph $G = (V, W)$ as $L = D - W$. We show that the notion of incidence matrix can be generalized to weighted graphs in a simple way. For any graph G^σ obtained by orienting the underlying graph of a weighted graph $G = (V, W)$, there is an incidence matrix B^σ such that

$$B^\sigma (B^\sigma)^\top = D - W = L.$$

We also prove that

$$x^\top L x = \frac{1}{2} \sum_{i,j=1}^{m} w_{i\,j}(x_i - x_j)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

Consequently, $x^\top L x$ does not depend on the diagonal entries in W, and if $w_{i\,j} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, then L is positive semidefinite. Then if W consists of nonnegative entries, the eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ of L are real and nonnegative, and there is an orthonormal basis of eigenvectors of L. We show that the number of connected components of the graph $G = (V, W)$ is equal to the dimension of the kernel of L, which is also equal to the dimension of the kernel of the transpose $(B^\sigma)^\top$ of any incidence matrix B^σ obtained by orienting the underlying graph of G.

We also define the normalized graph Laplacians L_{sym} and L_{rw}, given by

$$\begin{aligned} L_{\mathrm{sym}} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \\ L_{\mathrm{rw}} &= D^{-1} L = I - D^{-1} W, \end{aligned}$$

and prove some simple properties relating the eigenvalues and the eigenvectors of L, L_{sym} and L_{rw}. These normalized graph Laplacians show up when dealing with normalized cuts.

Next, we turn to *graph drawings* (Chapter 21). Graph drawing is a very attractive application of so-called spectral techniques, which is a fancy way of saying that that eigenvalues and eigenvectors of the graph Laplacian are used. Furthermore, it turns out that graph clustering using normalized cuts can be cast as a certain type of graph drawing.

Given an undirected graph $G = (V, E)$, with $|V| = m$, we would like to draw G in $\mathbb{R}^n$ for n (much) smaller than m. The idea is to assign a point $\rho(v_i)$ in $\mathbb{R}^n$ to the vertex $v_i \in V$, for every $v_i \in V$, and to draw a line segment between the points $\rho(v_i)$ and $\rho(v_j)$. Thus, a *graph drawing* is a function $\rho\colon V \to \mathbb{R}^n$.

We define the *matrix of a graph drawing ρ (in $\mathbb{R}^n$)* as a $m \times n$ matrix R whose ith row consists of the row vector $\rho(v_i)$ corresponding to the point representing v_i in $\mathbb{R}^n$. Typically, we want $n < m$; in fact n should be much smaller than m.

Since there are infinitely many graph drawings, it is desirable to have some criterion to decide which graph is better than another. Inspired by a physical model in which the edges are springs, it is natural to consider a representation to be better if it requires the springs to be less extended. We can formalize this by defining the *energy* of a drawing R by

$$\mathcal{E}(R) = \sum_{\{v_i,v_j\}\in E} \|\rho(v_i) - \rho(v_j)\|^2,$$

where $\rho(v_i)$ is the ith row of R and $\|\rho(v_i) - \rho(v_j)\|^2$ is the square of the Euclidean length of the line segment joining $\rho(v_i)$ and $\rho(v_j)$.

Then "good drawings" are drawings that minimize the energy function $\mathcal{E}$. Of course, the trivial representation corresponding to the zero matrix is optimum, so we need to impose extra constraints to rule out the trivial solution.

We can consider the more general situation where the springs are not necessarily identical. This can be modeled by a symmetric weight (or stiffness) matrix $W = (w_{ij})$, with $w_{ij} \geq 0$. In this case, our energy function becomes

$$\mathcal{E}(R) = \sum_{\{v_i,v_j\}\in E} w_{ij} \|\rho(v_i) - \rho(v_j)\|^2.$$

Following Godsil and Royle [77], we prove that

$$\mathcal{E}(R) = \mathrm{tr}(R^\top L R),$$

where

$$L = D - W,$$

is the familiar unnormalized Laplacian matrix associated with W, and where D is the degree matrix associated with W.

It can be shown that there is no loss in generality in assuming that the columns of R are pairwise orthogonal and that they have unit length. Such a matrix satisfies the equation

$R^\top R = I$ and the corresponding drawing is called an *orthogonal drawing*. This condition also rules out trivial drawings.

Then we prove the main theorem about graph drawings (Theorem 21.2), which essentially says that the matrix R of the desired graph drawing is constituted by the n eigenvectors of L associated with the smallest nonzero n eigenvalues of L. We give a number examples of graph drawings, many of which are borrowed or adapted from Spielman [163].

20.1 Directed Graphs, Undirected Graphs, Incidence Matrices, Adjacency Matrices, Weighted Graphs

Definition 20.1. A *directed graph* is a pair $G = (V, E)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and $E \subseteq V \times V$ is a set of ordered pairs of distinct nodes (that is, pairs $(u, v) \in V \times V$ with $u \neq v$), called *edges*. Given any edge $e = (u, v)$, we let $s(e) = u$ be the *source* of e and $t(e) = v$ be the *target* of e.

Remark: Since an edge is a pair (u, v) with $u \neq v$, self-loops are not allowed. Also, there is at most one edge from a node u to a node v. Such graphs are sometimes called *simple graphs*.

An example of a directed graph is shown in Figure 20.2.

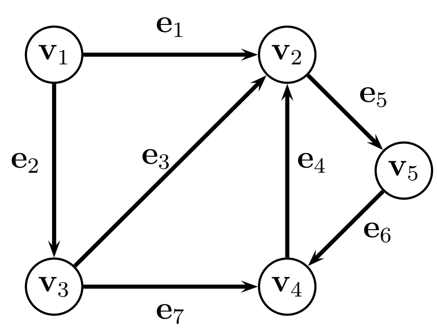


Figure 20.2: Graph G_1.

Definition 20.2. For every node $v \in V$, the *degree* $d(v)$ of v is the number of edges leaving or entering v:

$$d(v) = |\{u \in V \mid (v, u) \in E \text{ or } (u, v) \in E\}|.$$

We abbreviate $d(v_i)$ as d_i. The *degree matrix*, $D(G)$, is the diagonal matrix

$$D(G) = \mathrm{diag}(d_1, \ldots, d_m).$$

For example, for graph G_1, we have

$$D(G_1) = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

Unless confusion arises, we write D instead of $D(G)$.

Definition 20.3. Given a directed graph $G = (V, E)$, for any two nodes $u, v \in V$, a *path from u to v* is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ such that $v_0 = u$, $v_k = v$, and (v_i, v_{i+1}) is an edge in E for all i with $0 \leq i \leq k-1$. The integer k is the *length* of the path. A path is *closed* if $u = v$. The graph G is *strongly connected* if for any two distinct nodes $u, v \in V$, there is a path from u to v and there is a path from v to u.

Remark: The terminology *walk* is often used instead of *path*, the word path being reserved to the case where the nodes v_i are all distinct, except that $v_0 = v_k$ when the path is closed.

The binary relation on $V \times V$ defined so that u and v are related iff there is a path from u to v and there is a path from v to u is an equivalence relation whose equivalence classes are called the *strongly connected components* of G.

Definition 20.4. Given a directed graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, if $E = \{e_1, \ldots, e_n\}$, then the *incidence matrix* $B(G)$ of G is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +1 & \text{if } s(e_j) = v_i \\ -1 & \text{if } t(e_j) = v_i \\ 0 & \text{otherwise.} \end{cases}$$

Here is the incidence matrix of the graph G_1:

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{pmatrix}.$$

Observe that every column of an incidence matrix contains exactly two nonzero entries, $+1$ and -1. Again, unless confusion arises, we write B instead of $B(G)$.

When a directed graph has m nodes $v_1, \ldots, v_m$ and n edges $e_1, \ldots, e_n$, a vector $x \in \mathbb{R}^m$ can be viewed as a function $x \colon V \to \mathbb{R}$ assigning the value x_i to the node v_i. Under this interpretation, $\mathbb{R}^m$ is viewed as $\mathbb{R}^V$. Similarly, a vector $y \in \mathbb{R}^n$ can be viewed as a function

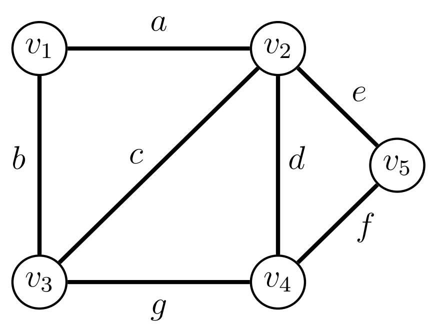


Figure 20.3: The undirected graph G_2.

in $\mathbb{R}^E$. This point of view is often useful. For example, the incidence matrix B can be interpreted as a linear map from $\mathbb{R}^E$ to $\mathbb{R}^V$, the *boundary map*, and $B^\top$ can be interpreted as a linear map from $\mathbb{R}^V$ to $\mathbb{R}^E$, the *coboundary map*.

Remark: Some authors adopt the opposite convention of sign in defining the incidence matrix, which means that their incidence matrix is $-B$.

Undirected graphs are obtained from directed graphs by forgetting the orientation of the edges.

Definition 20.5. A *graph* (or *undirected graph*) is a pair $G = (V, E)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and E is a set of two-element subsets of V (that is, subsets $\{u, v\}$, with $u, v \in V$ and $u \neq v$), called *edges*.

Remark: Since an edge is a set $\{u, v\}$, we have $u \neq v$, so self-loops are not allowed. Also, for every set of nodes $\{u, v\}$, there is at most one edge between u and v. As in the case of directed graphs, such graphs are sometimes called *simple graphs*.

An example of a graph is shown in Figure 20.3.

Definition 20.6. For every node $v \in V$, the *degree* $d(v)$ of v is the number of edges incident to v:

$$d(v) = |\{u \in V \mid \{u, v\} \in E\}|.$$

The degree matrix $D(G)$ (or simply, D) is defined as in Definition 20.2.

Definition 20.7. Given a (undirected) graph $G = (V, E)$, for any two nodes $u, v \in V$, a *path from u to v* is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ such that $v_0 = u$, $v_k = v$, and $\{v_i, v_{i+1}\}$ is an edge in E for all i with $0 \leq i \leq k - 1$. The integer k is the *length* of the path. A path is *closed* if $u = v$. The graph G is *connected* if for any two distinct nodes $u, v \in V$, there is a path from u to v.

Remark: The terminology *walk* or *chain* is often used instead of *path*, the word path being reserved to the case where the nodes v_i are all distinct, except that $v_0 = v_k$ when the path is closed.

The binary relation on $V \times V$ defined so that u and v are related iff there is a path from u to v is an equivalence relation whose equivalence classes are called the *connected components* of G.

The notion of incidence matrix for an undirected graph is not as useful as in the case of directed graphs

Definition 20.8. Given a graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, if $E = \{e_1, \ldots, e_n\}$, then the *incidence matrix* $B(G)$ of G is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +1 & \text{if } e_j = \{v_i, v_k\} \text{ for some } k \\ 0 & \text{otherwise.} \end{cases}$$

Unlike the case of directed graphs, the entries in the incidence matrix of a graph (undirected) are nonnegative. We usually write B instead of $B(G)$.

Definition 20.9. If $G = (V, E)$ is a directed or an undirected graph, given a node $u \in V$, any node $v \in V$ such that there is an edge (u, v) in the directed case or $\{u, v\}$ in the undirected case is called *adjacent to* u, and we often use the notation

$$u \sim v.$$

Observe that the binary relation $\sim$ is symmetric when G is an undirected graph, but in general it is not symmetric when G is a directed graph.

The notion of adjacency matrix is basically the same for directed or undirected graphs.

Definition 20.10. Given a directed or undirected graph $G = (V, E)$, with $V = \{v_1, \ldots, v_m\}$, the *adjacency matrix* $A(G)$ of G is the symmetric $m \times m$ matrix $(a_{i\,j})$ such that

(1) If G is directed, then

$$a_{i\,j} = \begin{cases} 1 & \text{if there is some edge } (v_i, v_j) \in E \text{ or some edge } (v_j, v_i) \in E \\ 0 & \text{otherwise.} \end{cases}$$

(2) Else if G is undirected, then

$$a_{i\,j} = \begin{cases} 1 & \text{if there is some edge } \{v_i, v_j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

As usual, unless confusion arises, we write A instead of $A(G)$. Here is the adjacency matrix of both graphs G_1 and G_2:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

If $G = (V, E)$ is an undirected graph, the adjacency matrix A of G can be viewed as a linear map from $\mathbb{R}^V$ to $\mathbb{R}^V$, such that for all $x \in \mathbb{R}^m$, we have

$$(Ax)_i = \sum_{j \sim i} x_j;$$

that is, the value of Ax at v_i is the sum of the values of x at the nodes v_j adjacent to v_i. The adjacency matrix can be viewed as a *diffusion operator*. This observation yields a geometric interpretation of what it means for a vector $x \in \mathbb{R}^m$ to be an eigenvector of A associated with some eigenvalue λ; we must have

$$\lambda x_i = \sum_{j \sim i} x_j, \quad i = 1, \ldots, m,$$

which means that the the sum of the values of x assigned to the nodes v_j adjacent to v_i is equal to λ times the value of x at v_i.

Definition 20.11. Given any undirected graph $G = (V, E)$, an *orientation* of G is a function $\sigma \colon E \to V \times V$ assigning a source and a target to every edge in E, which means that for every edge $\{u, v\} \in E$, either $\sigma(\{u, v\}) = (u, v)$ or $\sigma(\{u, v\}) = (v, u)$. The *oriented graph* G^σ obtained from G by applying the orientation σ is the directed graph $G^\sigma = (V, E^\sigma)$, with $E^\sigma = \sigma(E)$.

The following result shows how the number of connected components of an undirected graph is related to the rank of the incidence matrix of any oriented graph obtained from G.

Proposition 20.1. *Let $G = (V, E)$ be any undirected graph with m vertices, n edges, and c connected components. For any orientation σ of G, if B is the incidence matrix of the oriented graph G^σ, then $c = \dim(\mathrm{Ker}\,(B^\top))$, and B has rank $m - c$. Furthermore, the nullspace of $B^\top$ has a basis consisting of indicator vectors of the connected components of G; that is, vectors $(z_1, \ldots, z_m)$ such that $z_j = 1$ iff v_j is in the ith component K_i of G, and $z_j = 0$ otherwise.*

Proof. (After Godsil and Royle [77], Section 8.3). The fact that $\mathrm{rank}(B) = m - c$ will be proved last.

Let us prove that the kernel of $B^\top$ has dimension c. A vector $z \in \mathbb{R}^m$ belongs to the kernel of $B^\top$ iff $B^\top z = 0$ iff $z^\top B = 0$. In view of the definition of B, for every edge $\{v_i, v_j\}$

of G, the column of B corresponding to the oriented edge $\sigma(\{v_i, v_j\})$ has zero entries except for a $+1$ and a -1 in position i and position j or vice-versa, so we have

$$z_i = z_j.$$

An easy induction on the length of the path shows that if there is a path from v_i to v_j in G (unoriented), then $z_i = z_j$. Therefore, z has a constant value on any connected component of G. It follows that every vector $z \in \mathrm{Ker}\,(B^\top)$ can be written uniquely as a linear combination

$$z = \lambda_1 z^1 + \cdots + \lambda_c z^c,$$

where the vector z^i corresponds to the ith connected component K_i of G and is defined such that

$$z^i_j = \begin{cases} 1 & \text{iff } v_j \in K_i \\ 0 & \text{otherwise.} \end{cases}$$

This shows that $\dim(\mathrm{Ker}\,(B^\top)) = c$, and that $\mathrm{Ker}\,(B^\top)$ has a basis consisting of indicator vectors.

Since $B^\top$ is a $n \times m$ matrix, we have

$$m = \dim(\mathrm{Ker}\,(B^\top)) + \mathrm{rank}(B^\top),$$

and since we just proved that $\dim(\mathrm{Ker}\,(B^\top)) = c$, we obtain $\mathrm{rank}(B^\top) = m - c$. Since B and $B^\top$ have the same rank, $\mathrm{rank}(B) = m - c$, as claimed. $\square$

Definition 20.12. Following common practice, we denote by $\mathbf{1}$ the (column) vector (of dimension m) whose components are all equal to 1.

Since every column of B contains a single $+1$ and a single -1, the rows of $B^\top$ sum to zero, which can be expressed as

$$B^\top \mathbf{1} = 0.$$

According to Proposition 20.1, the graph G is connected iff B has rank $m-1$ iff the nullspace of $B^\top$ is the one-dimensional space spanned by $\mathbf{1}$.

In many applications, the notion of graph needs to be generalized to capture the intuitive idea that two nodes u and v are linked with a degree of certainty (or strength). Thus, we assign a nonnegative weight $w_{i\,j}$ to an edge $\{v_i, v_j\}$; the smaller $w_{i\,j}$ is, the weaker is the link (or similarity) between v_i and v_j, and the greater $w_{i\,j}$ is, the stronger is the link (or similarity) between v_i and v_j.

Definition 20.13. A *weighted graph* is a pair $G = (V, W)$, where $V = \{v_1, \ldots, v_m\}$ is a set of *nodes* or *vertices*, and W is a symmetric matrix called the *weight matrix*, such that $w_{i\,j} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, and $w_{i\,i} = 0$ for $i = 1, \ldots, m$. We say that a set $\{v_i, v_j\}$ is an edge iff $w_{i\,j} > 0$. The corresponding (undirected) graph (V, E) with $E = \{\{v_i, v_j\} \mid w_{i\,j} > 0\}$, is called the *underlying graph* of G.

Remark: Since $w_{i\,i} = 0$, these graphs have no self-loops. We can think of the matrix W as a generalized adjacency matrix. The case where $w_{i\,j} \in \{0, 1\}$ is equivalent to the notion of a graph as in Definition 20.5.

We can think of the weight $w_{i\,j}$ of an edge $\{v_i, v_j\}$ as a degree of similarity (or affinity) in an image, or a cost in a network. An example of a weighted graph is shown in Figure 20.4. The thickness of an edge corresponds to the magnitude of its weight.

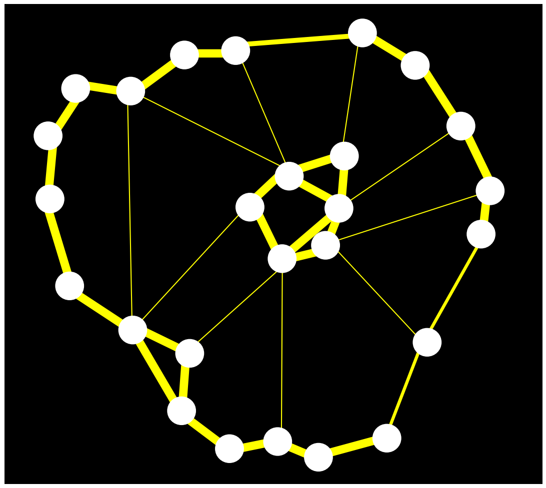


Figure 20.4: A weighted graph.

Definition 20.14. Given a weighted graph $G = (V, W)$, for every node $v_i \in V$, the *degree* $d(v_i)$ of v_i is the sum of the weights of the edges adjacent to v_i:

$$d(v_i) = \sum_{j=1}^{m} w_{i\,j}.$$

Note that in the above sum, only nodes v_j such that there is an edge $\{v_i, v_j\}$ have a nonzero contribution. Such nodes are said to be *adjacent* to v_i, and we write $v_i \sim v_j$. The degree matrix $D(G)$ (or simply, D) is defined as before, namely by $D(G) = \mathrm{diag}(d(v_1), \ldots, d(v_m))$.

The weight matrix W can be viewed as a linear map from $\mathbb{R}^V$ to itself. For all $x \in \mathbb{R}^m$, we have

$$(Wx)_i = \sum_{j \sim i} w_{ij} x_j;$$

that is, the value of Wx at v_i is the weighted sum of the values of x at the nodes v_j adjacent to v_i.

Observe that $W\mathbf{1}$ is the (column) vector $(d(v_1), \ldots, d(v_m))$ consisting of the degrees of the nodes of the graph.

We now define the most important concept of this chapter: the Laplacian matrix of a graph. Actually, as we will see, it comes in several flavors.

20.2 Laplacian Matrices of Graphs

Let us begin with directed graphs, although as we will see, graph Laplacians are fundamentally associated with undirected graph. The key proposition below shows how given an undirected graph G, for any orientation σ of G, $B^\sigma (B^\sigma)^\top$ relates to the adjacency matrix A (where B^σ is the incidence matrix of the directed graph G^σ). We reproduce the proof in Gallier [71] (see also Godsil and Royle [77]).

Proposition 20.2. *Given any undirected graph G, for any orientation σ of G, if B^σ is the incidence matrix of the directed graph G^σ, A is the adjacency matrix of G^σ, and D is the degree matrix such that $D_{i\,i} = d(v_i)$, then*

$$B^\sigma (B^\sigma)^\top = D - A.$$

Consequently, $L = B^\sigma (B^\sigma)^\top$ is independent of the orientation σ of G, and $D-A$ is symmetric and positive semidefinite; that is, the eigenvalues of $D - A$ are real and nonnegative.

Proof. The entry $B^\sigma (B^\sigma)^\top_{i\,j}$ is the inner product of the ith row b^σ_i, and the jth row b^σ_j of B^σ. If $i = j$, then as

$$b^\sigma_{i\,k} = \begin{cases} +1 & \text{if } s(e_k) = v_i \\ -1 & \text{if } t(e_k) = v_i \\ 0 & \text{otherwise} \end{cases}$$

we see that $b^\sigma_i \cdot b^\sigma_i = d(v_i)$. If $i \neq j$, then $b^\sigma_i \cdot b^\sigma_j \neq 0$ iff there is some edge e_k with $s(e_k) = v_i$ and $t(e_k) = v_j$ or vice-versa (which are mutually exclusive cases, since G^σ arises by orienting an undirected graph), in which case, $b^\sigma_i \cdot b^\sigma_j = -1$. Therefore,

$$B^\sigma (B^\sigma)^\top = D - A,$$

as claimed.

For every $x \in \mathbb{R}^m$, we have

$$x^\top L x = x^\top B^\sigma (B^\sigma)^\top x = ((B^\sigma)^\top x)^\top (B^\sigma)^\top x = \left\| (B^\sigma)^\top x \right\|_2^2 \geq 0,$$

since the Euclidean norm $\| \;\|_2$ is positive (definite). Therefore, $L = B^\sigma (B^\sigma)^\top$ is positive semidefinite. It is well-known that a real symmetric matrix is positive semidefinite iff its eigenvalues are nonnegative. $\square$

Definition 20.15. The matrix $L = B^\sigma (B^\sigma)^\top = D - A$ is called the *(unnormalized) graph Laplacian* of the graph G^σ. The *(unnormalized) graph Laplacian* of an undirected graph $G = (V, E)$ is defined by

$$L = D - A.$$

For example, the graph Laplacian of graph G_1 is

$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 4 & -1 & -1 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & -1 & 0 & -1 & 2 \end{pmatrix}.$$

Observe that each row of L sums to zero (because $(B^\sigma)^\top \mathbf{1} = 0$). Consequently, the vector $\mathbf{1}$ is in the nullspace of L.

Remarks:

1. With the unoriented version of the incidence matrix (see Definition 20.8), it can be shown that
$$BB^\top = D + A.$$

2. As pointed out by Evangelos Chatzipantazis, Proposition 20.2 in which the incidence matrix B^σ is replaced by the incidence matrix B of any *arbitrary* directed graph G does not hold. The problem is that such graphs may have both edges (v_i, v_j) and (v_j, v_i) between two distinct nodes v_i and v_j, and as a consequence, the inner product $b_i \cdot b_j = -2$ instead of -1. A simple counterexample is given by the directed graph with three vertices and four edges whose incidence matrix is given by
$$B = \begin{pmatrix} 1 & -1 & 0 & -1 \\ -1 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$
We have
$$BB^\top = \begin{pmatrix} 3 & -2 & -1 \\ -2 & 3 & -1 \\ -1 & -1 & 2 \end{pmatrix} \neq \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} = D - A.$$

The natural generalization of the notion of graph Laplacian to weighted graphs is this:

Definition 20.16. Given any weighted graph $G = (V, W)$ with $V = \{v_1, \ldots, v_m\}$, the *(unnormalized) graph Laplacian $L(G)$ of G* is defined by

$$L(G) = D(G) - W,$$

where $D(G) = \mathrm{diag}(d_1, \ldots, d_m)$ is the degree matrix of G (a diagonal matrix), with

$$d_i = \sum_{j=1}^m w_{i\,j}.$$

As usual, unless confusion arises, we write D instead of $D(G)$ and L instead of $L(G)$.

The graph Laplacian can be interpreted as a linear map from $\mathbb{R}^V$ to itself. For all $x \in \mathbb{R}^V$, we have

$$(Lx)_i = \sum_{j \sim i} w_{ij}(x_i - x_j).$$

It is clear from the equation $L = D - W$ that each row of L sums to 0, so the vector $\mathbf{1}$ is the nullspace of L, but it is less obvious that L is positive semidefinite. One way to prove it is to generalize slightly the notion of incidence matrix.

Definition 20.17. Given a weighted graph $G = (V, W)$, with $V = \{v_1, \ldots, v_m\}$, if $\{e_1, \ldots, e_n\}$ are the edges of the underlying graph of G (recall that $\{v_i, v_j\}$ is an edge of this graph iff $w_{ij} > 0$), for any oriented graph G^σ obtained by giving an orientation to the underlying graph of G, the *incidence matrix* B^σ of G^σ is the $m \times n$ matrix whose entries $b_{i\,j}$ are given by

$$b_{i\,j} = \begin{cases} +\sqrt{w_{ij}} & \text{if } s(e_j) = v_i \\ -\sqrt{w_{ij}} & \text{if } t(e_j) = v_i \\ 0 & \text{otherwise.} \end{cases}$$

For example, given the weight matrix

$$W = \begin{pmatrix} 0 & 3 & 6 & 3 \\ 3 & 0 & 0 & 3 \\ 6 & 0 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix},$$

the incidence matrix B corresponding to the orientation of the underlying graph of W where an edge (i, j) is oriented positively iff $i < j$ is

$$B = \begin{pmatrix} 1.7321 & 2.4495 & 1.7321 & 0 & 0 \\ -1.7321 & 0 & 0 & 1.7321 & 0 \\ 0 & -2.4495 & 0 & 0 & 1.7321 \\ 0 & 0 & -1.7321 & -1.7321 & -1.7321 \end{pmatrix}.$$

The reader should verify that $BB^\top = D - W$. This is true in general, see Proposition 20.3.

It is easy to see that Proposition 20.1 applies to the underlying graph of G. For any oriented graph G^σ obtained from the underlying graph of G, the rank of the incidence matrix B^σ is equal to $m - c$, where c is the number of connected components of the underlying graph of G, and we have $(B^\sigma)^\top \mathbf{1} = 0$. We also have the following version of Proposition 20.2 whose proof is immediately adapted.

Proposition 20.3. *Given any weighted graph $G = (V, W)$ with $V = \{v_1, \ldots, v_m\}$, if B^σ is the incidence matrix of any oriented graph G^σ obtained from the underlying graph of G and D is the degree matrix of G, then*

$$B^\sigma (B^\sigma)^\top = D - W = L.$$

Consequently, $B^{\sigma}(B^{\sigma})^{\top}$ is independent of the orientation of the underlying graph of G and $L = D - W$ is symmetric and positive semidefinite; that is, the eigenvalues of $L = D - W$ are real and nonnegative.

Another way to prove that L is positive semidefinite is to evaluate the quadratic form $x^{\top} L x$.

Proposition 20.4. *For any $m \times m$ symmetric matrix $W = (w_{ij})$, if we let $L = D - W$ where D is the degree matrix associated with W (that is, $d_i = \sum_{j=1}^{m} w_{ij}$), then we have*

$$x^{\top} L x = \frac{1}{2} \sum_{i,j=1}^{m} w_{i\,j} (x_i - x_j)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

Consequently, $x^{\top} L x$ does not depend on the diagonal entries in W, and if $w_{i\,j} \geq 0$ for all $i, j \in \{1, \ldots, m\}$, then L is positive semidefinite.

Proof. We have

$$\begin{aligned} x^{\top} L x &= x^{\top} D x - x^{\top} W x \\ &= \sum_{i=1}^{m} d_i x_i^2 - \sum_{i,j=1}^{m} w_{i\,j} x_i x_j \\ &= \frac{1}{2} \left(\sum_{i=1}^{m} d_i x_i^2 - 2 \sum_{i,j=1}^{m} w_{i\,j} x_i x_j + \sum_{i=1}^{m} d_i x_i^2 \right) \\ &= \frac{1}{2} \sum_{i,j=1}^{m} w_{i\,j} (x_i - x_j)^2. \end{aligned}$$

Obviously, the quantity on the right-hand side does not depend on the diagonal entries in W, and if $w_{i\,j} \geq 0$ for all i, j, then this quantity is nonnegative. ∎

Proposition 20.4 immediately implies the following facts: For any weighted graph $G = (V, W)$,

1. The eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ of L are real and nonnegative, and there is an orthonormal basis of eigenvectors of L.

2. The smallest eigenvalue λ_1 of L is equal to 0, and $\mathbf{1}$ is a corresponding eigenvector.

It turns out that the dimension of the nullspace of L (the eigenspace of 0) is equal to the number of connected components of the underlying graph of G.

Proposition 20.5. *Let $G = (V, W)$ be a weighted graph. The number c of connected components $K_1, \ldots, K_c$ of the underlying graph of G is equal to the dimension of the nullspace of L, which is equal to the multiplicity of the eigenvalue 0. Furthermore, the nullspace of L has a basis consisting of indicator vectors of the connected components of G, that is, vectors $(f_1, \ldots, f_m)$ such that $f_j = 1$ iff $v_j \in K_i$ and $f_j = 0$ otherwise.*

Proof. Since $L = BB^\top$ for the incidence matrix B associated with any oriented graph obtained from G, and since L and $B^\top$ have the same nullspace, by Proposition 20.1, the dimension of the nullspace of L is equal to the number c of connected components of G and the indicator vectors of the connected components of G form a basis of $\mathrm{Ker}\,(L)$. $\square$

Proposition 20.5 implies that if the underlying graph of G is connected, then the second eigenvalue λ_2 of L is strictly positive.

Remarkably, the eigenvalue λ_2 contains a lot of information about the graph G (assuming that $G = (V, E)$ is an undirected graph). This was first discovered by Fiedler in 1973, and for this reason, λ_2 is often referred to as the *Fiedler number*. For more on the properties of the Fiedler number, see Godsil and Royle [77] (Chapter 13) and Chung [39]. More generally, the spectrum $(0, \lambda_2, \ldots, \lambda_m)$ of L contains a lot of information about the combinatorial structure of the graph G. Leverage of this information is the object of *spectral graph theory*.

20.3 Normalized Laplacian Matrices of Graphs

It turns out that normalized variants of the graph Laplacian are needed, especially in applications to graph clustering. These variants make sense only if G has no isolated vertices.

Definition 20.18. Given a weighted graph $G = (V, W)$, a vertex $u \in V$ is *isolated* if it is not incident to any other vertex. This means that every row of W contains some strictly positive entry.

If G has no isolated vertices, then the degree matrix D contains positive entries, so it is invertible and $D^{-1/2}$ makes sense; namely

$$D^{-1/2} = \mathrm{diag}(d_1^{-1/2}, \ldots, d_m^{-1/2}),$$

and similarly for any real exponent α.

Definition 20.19. Given any weighted directed graph $G = (V, W)$ with no isolated vertex and with $V = \{v_1, \ldots, v_m\}$, the *(normalized) graph Laplacians L_{sym} and L_{rw} of G* are defined by

$$\begin{aligned} L_{\mathrm{sym}} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \\ L_{\mathrm{rw}} &= D^{-1} L = I - D^{-1} W. \end{aligned}$$

Observe that the Laplacian $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$ is a symmetric matrix (because L and $D^{-1/2}$ are symmetric) and that

$$L_{\mathrm{rw}} = D^{-1/2}L_{\mathrm{sym}}D^{1/2}.$$

The reason for the notation L_{rw} is that this matrix is closely related to a random walk on the graph G.

Example 20.1. As an example, the matrices L_{sym} and L_{rw} associated with the graph G_1 are

$$L_{\mathrm{sym}} = \begin{pmatrix} 1.0000 & -0.3536 & -0.4082 & 0 & 0 \\ -0.3536 & 1.0000 & -0.2887 & -0.2887 & -0.3536 \\ -0.4082 & -0.2887 & 1.0000 & -0.3333 & 0 \\ 0 & -0.2887 & -0.3333 & 1.0000 & -0.4082 \\ 0 & -0.3536 & 0 & -0.4082 & 1.0000 \end{pmatrix}$$

and

$$L_{\mathrm{rw}} = \begin{pmatrix} 1.0000 & -0.5000 & -0.5000 & 0 & 0 \\ -0.2500 & 1.0000 & -0.2500 & -0.2500 & -0.2500 \\ -0.3333 & -0.3333 & 1.0000 & -0.3333 & 0 \\ 0 & -0.3333 & -0.3333 & 1.0000 & -0.3333 \\ 0 & -0.5000 & 0 & -0.5000 & 1.0000 \end{pmatrix}.$$

Since the unnormalized Laplacian L can be written as $L = BB^{\top}$, where B is the incidence matrix of any oriented graph obtained from the underlying graph of $G = (V, W)$, if we let

$$B_{\mathrm{sym}} = D^{-1/2}B,$$

we get

$$L_{\mathrm{sym}} = B_{\mathrm{sym}}B_{\mathrm{sym}}^{\top}.$$

In particular, for any singular decomposition $B_{\mathrm{sym}} = U\Sigma V^{\top}$ of B_{sym} (with U an $m \times m$ orthogonal matrix, Σ a "diagonal" $m \times n$ matrix of singular values, and V an $n \times n$ orthogonal matrix), the eigenvalues of L_{sym} are the squares of the top m singular values of B_{sym}, and the vectors in U are orthonormal eigenvectors of L_{sym} with respect to these eigenvalues (the squares of the top m diagonal entries of Σ). Computing the SVD of B_{sym} generally yields more accurate results than diagonalizing L_{sym}, especially when L_{sym} has eigenvalues with high multiplicity.

There are simple relationships between the eigenvalues and the eigenvectors of L_{sym}, and L_{rw}. There is also a simple relationship with the generalized eigenvalue problem $Lx = \lambda Dx$.

Proposition 20.6. *Let $G = (V, W)$ be a weighted graph without isolated vertices. The graph Laplacians, L, L_{sym}, and L_{rw} satisfy the following properties:*

(1) The matrix L_{sym} is symmetric and positive semidefinite. In fact,

$$x^\top L_{\mathrm{sym}} x = \frac{1}{2}\sum_{i,j=1}^m w_{i\,j}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)^2 \quad \text{for all } x \in \mathbb{R}^m.$$

(2) The normalized graph Laplacians L_{sym} and L_{rw} have the same spectrum $(0 = \nu_1 \leq \nu_2 \leq \ldots \leq \nu_m)$, and a vector $u \neq 0$ is an eigenvector of L_{rw} for λ iff $D^{1/2}u$ is an eigenvector of L_{sym} for λ.

(3) The graph Laplacians L and L_{sym} are symmetric and positive semidefinite.

(4) A vector $u \neq 0$ is a solution of the generalized eigenvalue problem $Lu = \lambda Du$ iff $D^{1/2}u$ is an eigenvector of L_{sym} for the eigenvalue λ iff u is an eigenvector of L_{rw} for the eigenvalue λ.

(5) The graph Laplacians, L and L_{rw} have the same nullspace. For any vector u, we have $u \in \mathrm{Ker}\,(L)$ iff $D^{1/2}u \in \mathrm{Ker}\,(L_{\mathrm{sym}})$.

(6) The vector $\mathbf{1}$ is in the nullspace of L_{rw}, and $D^{1/2}\mathbf{1}$ is in the nullspace of L_{sym}.

(7) For every eigenvalue ν_i of the normalized graph Laplacian L_{sym}, we have $0 \leq \nu_i \leq 2$. Furthermore, $\nu_m = 2$ iff the underlying graph of G contains a nontrivial connected bipartite component.

(8) If $m \geq 2$ and if the underlying graph of G is not a complete graph,[1] then $\nu_2 \leq 1$. Furthermore the underlying graph of G is a complete graph iff $\nu_2 = \frac{m}{m-1}$.

(9) If $m \geq 2$ and if the underlying graph of G is connected, then $\nu_2 > 0$.

(10) If $m \geq 2$ and if the underlying graph of G has no isolated vertices, then $\nu_m \geq \frac{m}{m-1}$.

Proof. (1) We have $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$, and $D^{-1/2}$ is a symmetric invertible matrix (since it is an invertible diagonal matrix). It is a well-known fact of linear algebra that if B is an invertible matrix, then a matrix S is symmetric, positive semidefinite iff $BSB^\top$ is symmetric, positive semidefinite. Since L is symmetric, positive semidefinite, so is $L_{\mathrm{sym}} = D^{-1/2}LD^{-1/2}$. The formula

$$x^\top L_{\mathrm{sym}} x = \frac{1}{2}\sum_{i,j=1}^m w_{i\,j}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)^2 \quad \text{for all } x \in \mathbb{R}^m$$

follows immediately from Proposition 20.4 by replacing x by $D^{-1/2}x$, and also shows that L_{sym} is positive semidefinite.

(2) Since

$$L_{\mathrm{rw}} = D^{-1/2}L_{\mathrm{sym}}D^{1/2},$$

[1]Recall that an undirected graph is complete if for any two distinct nodes u, v, there is an edge $\{u, v\}$.

the matrices L_{sym} and L_{rw} are similar, which implies that they have the same spectrum. In fact, since $D^{1/2}$ is invertible,

$$L_{\mathrm{rw}}u = D^{-1}Lu = \lambda u$$

iff

$$D^{-1/2}Lu = \lambda D^{1/2}u$$

iff

$$D^{-1/2}LD^{-1/2}D^{1/2}u = L_{\mathrm{sym}}D^{1/2}u = \lambda D^{1/2}u,$$

which shows that a vector $u \neq 0$ is an eigenvector of L_{rw} for λ iff $D^{1/2}u$ is an eigenvector of L_{sym} for λ.

(3) We already know that L and L_{sym} are positive semidefinite.

(4) Since $D^{-1/2}$ is invertible, we have

$$Lu = \lambda Du$$

iff

$$D^{-1/2}Lu = \lambda D^{1/2}u$$

iff

$$D^{-1/2}LD^{-1/2}D^{1/2}u = L_{\mathrm{sym}}D^{1/2}u = \lambda D^{1/2}u,$$

which shows that a vector $u \neq 0$ is a solution of the generalized eigenvalue problem $Lu = \lambda Du$ iff $D^{1/2}u$ is an eigenvector of L_{sym} for the eigenvalue λ. The second part of the statement follows from (2).

(5) Since D^{-1} is invertible, we have $Lu = 0$ iff $D^{-1}Lu = L_{\mathrm{rw}}u = 0$. Similarly, since $D^{-1/2}$ is invertible, we have $Lu = 0$ iff $D^{-1/2}LD^{-1/2}D^{1/2}u = 0$ iff $D^{1/2}u \in \mathrm{Ker}\,(L_{\mathrm{sym}})$.

(6) Since $L\mathbf{1} = 0$, we get $L_{\mathrm{rw}}\mathbf{1} = D^{-1}L\mathbf{1} = 0$. That $D^{1/2}\mathbf{1}$ is in the nullspace of L_{sym} follows from (2). Properties (7)–(10) are proven in Chung [39] (Chapter 1). $\square$

The eigenvalues the matrices L_{sym} and L_{rw} from Example 20.1 are

$$0,\ 7257,\ 1.1667,\ 1.5,\ 1.6076.$$

On the other hand, the eigenvalues of the unnormalized Laplacian for G_1 are

$$0,\ 1.5858,\ 3,\ 4.4142,\ 5.$$

Remark: Observe that although the matrices L_{sym} and L_{rw} have the same spectrum, the matrix L_{rw} is generally not symmetric, whereas L_{sym} is symmetric.

A version of Proposition 20.5 also holds for the graph Laplacians L_{sym} and L_{rw}. This follows easily from the fact that Proposition 20.1 applies to the underlying graph of a weighted graph. The proof is left as an exercise.

Proposition 20.7. *Let $G = (V, W)$ be a weighted graph. The number c of connected components $K_1, \ldots, K_c$ of the underlying graph of G is equal to the dimension of the nullspace of both L_{sym} and L_{rw}, which is equal to the multiplicity of the eigenvalue 0. Furthermore, the nullspace of L_{rw} has a basis consisting of indicator vectors of the connected components of G, that is, vectors $(f_1, \ldots, f_m)$ such that $f_j = 1$ iff $v_j \in K_i$ and $f_j = 0$ otherwise. For L_{sym}, a basis of the nullpace is obtained by multiplying the above basis of the nullspace of L_{rw} by $D^{1/2}$.*

A particularly interesting application of graph Laplacians is graph clustering.

20.4 Graph Clustering Using Normalized Cuts

In order to explain this problem we need some definitions.

Definition 20.20. Given any subset of nodes $A \subseteq V$, we define the *volume* $\mathrm{vol}(A)$ of A as the sum of the weights of all edges adjacent to nodes in A:

$$\mathrm{vol}(A) = \sum_{v_i \in A} \sum_{j=1}^{m} w_{i\,j}.$$

Given any two subsets $A, B \subseteq V$ (not necessarily distinct), we define $\mathrm{links}(A, B)$ by

$$\mathrm{links}(A, B) = \sum_{v_i \in A, v_j \in B} w_{i\,j}.$$

The quantity $\mathrm{links}(A, \overline{A}) = \mathrm{links}(\overline{A}, A)$ (where $\overline{A} = V - A$ denotes the complement of A in V) measures how many links escape from A (and $\overline{A}$). We define the *cut* of A as

$$\mathrm{cut}(A) = \mathrm{links}(A, \overline{A}).$$

The notion of volume is illustrated in Figure 20.5 and the notions of cut is illustrated in Figure 20.6.

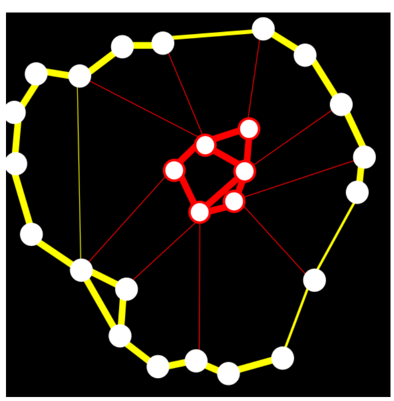


Figure 20.5: Volume of a set of nodes.

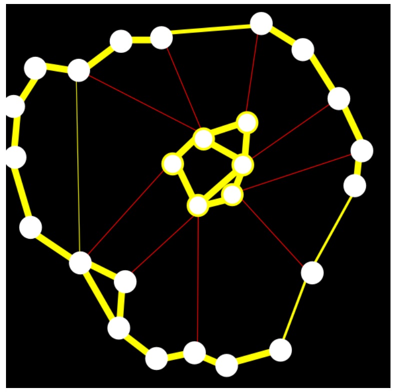


Figure 20.6: A cut involving the set of nodes in the center and the nodes on the perimeter.

The above concepts play a crucial role in the theory of normalized cuts. This beautiful and deeply original method first published in Shi and Malik [160], has now come to be a "textbook chapter" of computer vision and machine learning. It was invented by Jianbo Shi and Jitendra Malik and was the main topic of Shi's dissertation. This method was extended to $K \geq 3$ clusters by Stella Yu in her dissertation [191] and is also the subject of Yu and Shi [193].

Given a set of data, the goal of clustering is to partition the data into different groups according to their similarities. When the data is given in terms of a similarity graph G, where the weight $w_{i\,j}$ between two nodes v_i and v_j is a measure of similarity of v_i and v_j, the problem can be stated as follows: Find a partition $(A_1, \ldots, A_K)$ of the set of nodes V into different groups such that the edges between different groups have very low weight (which indicates that the points in different clusters are dissimilar), and the edges within a group have high weight (which indicates that points within the same cluster are similar).

The above graph clustering problem can be formalized as an optimization problem, using the notion of cut mentioned earlier. If we want to partition V into K clusters, we can do so by finding a partition $(A_1, \ldots, A_K)$ that minimizes the quantity

$$\mathrm{cut}(A_1, \ldots, A_K) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{cut}(A_i) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{links}(A_i, \overline{A}_i).$$

For $K = 2$, the mincut problem is a classical problem that can be solved efficiently, but in practice, it does not yield satisfactory partitions. Indeed, in many cases, the mincut solution separates one vertex from the rest of the graph. What we need is to design our cost function in such a way that it keeps the subsets A_i "reasonably large" (reasonably balanced).

An example of a weighted graph and a partition of its nodes into two clusters is shown in Figure 20.7.

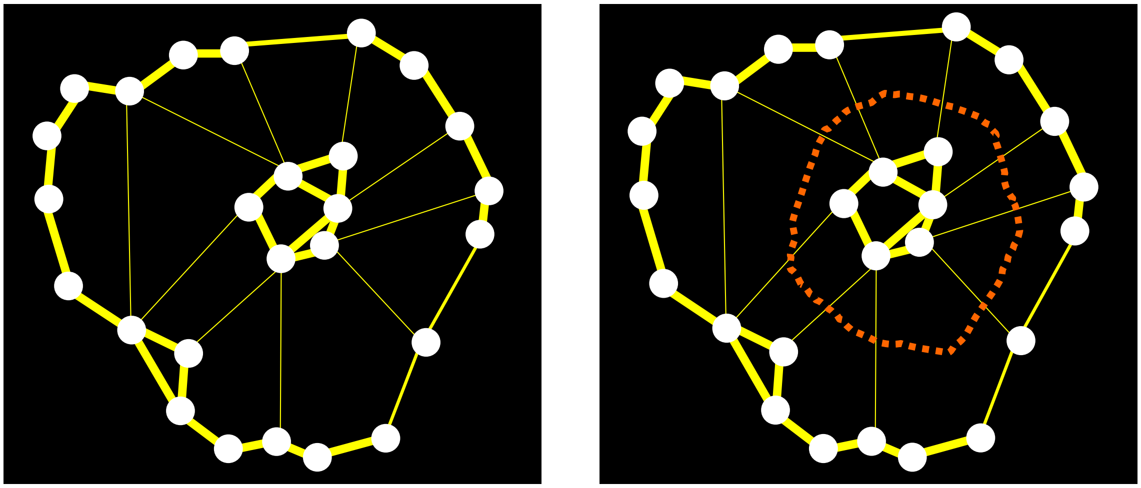


Figure 20.7: A weighted graph and its partition into two clusters.

A way to get around this problem is to normalize the cuts by dividing by some measure of each subset A_i. A solution using the volume $\mathrm{vol}(A_i)$ of A_i (for $K = 2$) was proposed and investigated in a seminal paper of Shi and Malik [160]. Subsequently, Yu (in her dissertation [191]) and Yu and Shi [193] extended the method to $K > 2$ clusters. The idea is to minimize the cost function

$$\mathrm{Ncut}(A_1, \ldots, A_K) = \sum_{i=1}^{K} \frac{\mathrm{links}(A_i, \overline{A_i})}{\mathrm{vol}(A_i)} = \sum_{i=1}^{K} \frac{\mathrm{cut}(A_i)}{\mathrm{vol}(A_i)}.$$

The next step is to express our optimization problem in matrix form, and this can be done in terms of Rayleigh ratios involving the graph Laplacian in the numerators. This theory is very beautiful, but we do not have the space to present it here. The interested reader is referred to Gallier [74].

20.5 Summary

The main concepts and results of this chapter are listed below:

- Directed graphs, undirected graphs.
- Incidence matrices, adjacency matrices.
- Weighted graphs.
- Degree matrix.
- Graph Laplacian (unnormalized).

- Normalized graph Laplacian.
- Spectral graph theory.
- Graph clustering using normalized cuts.

20.6 Problems

Problem 20.1. Find the unnormalized Laplacian of the graph representing a triangle and of the graph representing a square.

Problem 20.2. Consider the complete graph K_m on $m \geq 2$ nodes.

(1) Prove that the normalized Laplacian L_{sym} of K is

$$L_{\mathrm{sym}} = \begin{pmatrix} 1 & -1/(m-1) & \dots & -1/(m-1) & -1/(m-1) \\ -1/(m-1) & 1 & \dots & -1/(m-1) & -1/(m-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -1/(m-1) & -1/(m-1) & \dots & 1 & -1/(m-1) \\ -1/(m-1) & -1/(m-1) & \dots & -1/(m-1) & 1 \end{pmatrix}.$$

(2) Prove that the characteristic polynomial of L_{sym} is

$$\begin{vmatrix} \lambda-1 & 1/(m-1) & \dots & 1/(m-1) & 1/(m-1) \\ 1/(m-1) & \lambda-1 & \dots & 1/(m-1) & 1/(m-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1/(m-1) & 1/(m-1) & \dots & \lambda-1 & 1/(m-1) \\ 1/(m-1) & 1/(m-1) & \dots & 1/(m-1) & \lambda-1 \end{vmatrix} = \lambda \left(\lambda - \frac{m}{m-1}\right)^{m-1}.$$

Hint. First subtract the second column from the first, factor $\lambda - m/(m-1)$, and then add the first row to the second. Repeat this process. You will end up with the determinant

$$\begin{vmatrix} \lambda - 1/(m-1) & 1 \\ 1/(m-1) & \lambda - 1 \end{vmatrix}.$$

Problem 20.3. Consider the complete bipartite graph $K_{m,n}$ on $m+n \geq 3$ nodes, with edges between each of the first $m \geq 1$ nodes to each of the last $n \geq 1$ nodes. Prove that the eigenvalues of the normalized Laplacian L_{sym} of $K_{m,n}$ are 0, 1 with multiplicity $m+n-2$, and 2.

Problem 20.4. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes, and let $L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2}$ be its normalized Laplacian (with L its unnormalized Laplacian).

(1) For any $y \in \mathbb{R}^V$, consider the Rayleigh ratio

$$R = \frac{y^\top L_{\mathrm{sym}}\, y}{y^\top y}.$$

Prove that if $x = D^{-1/2}y$, then

$$R = \frac{x^\top Lx}{(D^{1/2}x)^\top D^{1/2}x} = \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_{v} d_v x(v)^2}.$$

(2) Prove that the second eigenvalue ν_2 of L_{sym} is given by

$$\nu_2 = \min_{\mathbf{1}^\top Dx=0, x\neq 0} \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_{v} d_v x(v)^2}.$$

(3) Prove that the largest eigenvalue ν_m of L_{sym} is given by

$$\nu_m = \max_{x\neq 0} \frac{\displaystyle\sum_{u\sim v}(x(u) - x(v))^2}{\displaystyle\sum_{v} d_v x(v)^2}.$$

Problem 20.5. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes. If $0 = \nu_1 \leq \nu_1 \leq \ldots \leq \nu_m$ are the eigenvalues of L_{sym}, prove the following properties:

(1) We have $\nu_1 + \nu_2 + \cdots + \nu_m = m$.

(2) We have $\nu_2 \leq m/(m-1)$, with equality holding iff $G = K_m$, the complete graph on m nodes.

(3) We have $\nu_m \geq m/(m-1)$.

(4) If G is not a complete graph, then $\nu_2 \leq 1$

Hint. If a and b are nonadjacent nodes, consider the function x given by

$$x(v) = \begin{cases} d_b & \text{if } v = a \\ -d_a & \text{if } v = b \\ 0 & \text{if } v \neq a, b, \end{cases}$$

and use Problem 20.4(2).

(5) Prove that $\nu_m \leq 2$. Prove that $\nu_m = 2$ iff the underlying graph of G contains a nontrivial connected bipartite component.

Hint. Use Problem 20.4(3).

(6) Prove that if G is connected, then $\nu_2 > 0$.

Problem 20.6. Let G be a graph with a set of nodes V with $m \geq 2$ elements, without isolated nodes. Let $\mathrm{vol}(G) = \sum_{v \in V} d_v$ and let

$$\overline{x} = \frac{\sum_v d_v x(v)}{\mathrm{vol}(G)}.$$

Prove that

$$\nu_2 = \min_{x \neq 0} \frac{\displaystyle\sum_{u \sim v} (x(u) - x(v))^2}{\displaystyle\sum_v d_v (x(v) - \overline{x})^2}.$$

Problem 20.7. Let G be a connected bipartite graph. Prove that if ν is an eigenvalue of L_{sym}, then $2 - \nu$ is also an eigenvalue of L_{sym}.

Problem 20.8. Prove Proposition 20.7.

Chapter 21

Spectral Graph Drawing

21.1 Graph Drawing and Energy Minimization

Let $G = (V, E)$ be some undirected graph. It is often desirable to draw a graph, usually in the plane but possibly in 3D, and it turns out that the graph Laplacian can be used to design surprisingly good methods. Say $|V| = m$. The idea is to assign a point $\rho(v_i)$ in $\mathbb{R}^n$ to the vertex $v_i \in V$, for every $v_i \in V$, and to draw a line segment between the points $\rho(v_i)$ and $\rho(v_j)$ iff there is an edge $\{v_i, v_j\}$.

Definition 21.1. Let $G = (V, E)$ be some undirected graph with m vertices. A *graph drawing* is a function $\rho\colon V \to \mathbb{R}^n$, for some $n \geq 1$. The *matrix of a graph drawing ρ (in $\mathbb{R}^n$)* is a $m \times n$ matrix R whose ith row consists of the row vector $\rho(v_i)$ corresponding to the point representing v_i in $\mathbb{R}^n$.

For a graph drawing to be useful we want $n \leq m$; in fact n should be much smaller than m, typically $n = 2$ or $n = 3$.

Definition 21.2. A graph drawing is *balanced* iff the sum of the entries of every column of the matrix of the graph drawing is zero, that is,

$$\mathbf{1}^\top R = 0.$$

If a graph drawing is not balanced, it can be made balanced by a suitable translation. We may also assume that the columns of R are linearly independent, since any basis of the column space also determines the drawing. Thus, from now on, we may assume that $n \leq m$.

Remark: A graph drawing $\rho\colon V \to \mathbb{R}^n$ is not required to be injective, which may result in degenerate drawings where distinct vertices are drawn as the same point. For this reason, we prefer not to use the terminology *graph embedding*, which is often used in the literature. This is because in differential geometry, an embedding always refers to an injective map. The term *graph immersion* would be more appropriate.

As explained in Godsil and Royle [77], we can imagine building a physical model of G by connecting adjacent vertices (in $\mathbb{R}^n$) by identical springs. Then it is natural to consider a representation to be better if it requires the springs to be less extended. We can formalize this by defining the *energy* of a drawing R by

$$\mathcal{E}(R) = \sum_{\{v_i,v_j\}\in E} \|\rho(v_i) - \rho(v_j)\|^2,$$

where $\rho(v_i)$ is the ith row of R and $\|\rho(v_i) - \rho(v_j)\|^2$ is the square of the Euclidean length of the line segment joining $\rho(v_i)$ and $\rho(v_j)$.

Then, "good drawings" are drawings that minimize the energy function $\mathcal{E}$. Of course, the trivial representation corresponding to the zero matrix is optimum, so we need to impose extra constraints to rule out the trivial solution.

We can consider the more general situation where the springs are not necessarily identical. This can be modeled by a symmetric weight (or stiffness) matrix $W = (w_{ij})$, with $w_{ij} \geq 0$. Then our energy function becomes

$$\mathcal{E}(R) = \sum_{\{v_i,v_j\}\in E} w_{ij} \|\rho(v_i) - \rho(v_j)\|^2.$$

It turns out that this function can be expressed in terms of the Laplacian $L = D - W$. The following proposition is shown in Godsil and Royle [77]. We give a slightly more direct proof.

Proposition 21.1. *Let $G = (V, W)$ be a weighted graph, with $|V| = m$ and W an $m \times m$ symmetric matrix, and let R be the matrix of a graph drawing ρ of G in $\mathbb{R}^n$ (a $m \times n$ matrix). If $L = D - W$ is the unnormalized Laplacian matrix associated with W, then*

$$\mathcal{E}(R) = \mathrm{tr}(R^\top L R).$$

Proof. Since $\rho(v_i)$ is the ith row of R (and $\rho(v_j)$ is the jth row of R), if we denote the kth column of R by R^k, using Proposition 20.4, we have

$$\begin{aligned}
\mathcal{E}(R) &= \sum_{\{v_i,v_j\}\in E} w_{ij} \|\rho(v_i) - \rho(v_j)\|^2 \\
&= \sum_{k=1}^n \sum_{\{v_i,v_j\}\in E} w_{ij}(R_{ik} - R_{jk})^2 \\
&= \sum_{k=1}^n \frac{1}{2} \sum_{i,j=1}^m w_{ij}(R_{ik} - R_{jk})^2 \\
&= \sum_{k=1}^n (R^k)^\top L R^k = \mathrm{tr}(R^\top L R),
\end{aligned}$$

as claimed. ∎

Since the matrix $R^\top LR$ is symmetric, it has real eigenvalues. Actually, since L is positive semidefinite, so is $R^\top LR$. Then the trace of $R^\top LR$ is equal to the sum of its positive eigenvalues, and this is the energy $\mathcal{E}(R)$ of the graph drawing.

If R is the matrix of a graph drawing in $\mathbb{R}^n$, then for any $n \times n$ invertible matrix M, the map that assigns $\rho(v_i)M$ to v_i is another graph drawing of G, and these two drawings convey the same amount of information. From this point of view, *a graph drawing is determined by the column space of R*. Therefore, it is reasonable to assume that the columns of R are pairwise orthogonal and that they have unit length. Such a matrix satisfies the equation $R^\top R = I$.

Definition 21.3. If the matrix R of a graph drawing satisfies the equation $R^\top R = I$, then the corresponding drawing is called an *orthogonal graph drawing*.

This above condition also rules out trivial drawings. The following result tells us how to find minimum energy orthogonal balanced graph drawings, provided the graph is connected. Recall that

$$L\mathbf{1} = 0,$$

as we already observed.

Theorem 21.2. *Let $G = (V, W)$ be a weighted graph with $|V| = m$. If $L = D - W$ is the (unnormalized) Laplacian of G, and if the eigenvalues of L are $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_m$, then the minimal energy of any balanced orthogonal graph drawing of G in $\mathbb{R}^n$ is equal to $\lambda_2 + \cdots + \lambda_{n+1}$ (in particular, this implies that $n < m$). The $m \times n$ matrix R consisting of any unit eigenvectors $u_2, \ldots, u_{n+1}$ associated with $\lambda_2 \leq \ldots \leq \lambda_{n+1}$ yields a balanced orthogonal graph drawing of minimal energy; it satisfies the condition $R^\top R = I$.*

Proof. We present the proof given in Godsil and Royle [77] (Section 13.4, Theorem 13.4.1). The key point is that the sum of the n smallest eigenvalues of L is a lower bound for $\mathrm{tr}(R^\top LR)$. This can be shown using a Rayleigh ratio argument; see Proposition 17.25 (the Poincaré separation theorem). Then any n eigenvectors $(u_1, \ldots, u_n)$ associated with $\lambda_1, \ldots, \lambda_n$ achieve this bound. Because the first eigenvalue of L is $\lambda_1 = 0$ and because we are assuming that $\lambda_2 > 0$, we have $u_1 = \mathbf{1}/\sqrt{m}$. Since the u_j are pairwise orthogonal for $i = 2, \ldots, n$ and since u_i is orthogonal to $u_1 = \mathbf{1}/\sqrt{m}$, the entries in u_i add up to 0. Consequently, for any ℓ with $2 \leq \ell \leq n$, by deleting u_1 and using $(u_2, \ldots, u_\ell)$, we obtain a balanced orthogonal graph drawing in $\mathbb{R}^{\ell-1}$ with the same energy as the orthogonal graph drawing in $\mathbb{R}^\ell$ using $(u_1, u_2, \ldots, u_\ell)$. Conversely, from any balanced orthogonal drawing in $\mathbb{R}^{\ell-1}$ using $(u_2, \ldots, u_\ell)$, we obtain an orthogonal graph drawing in $\mathbb{R}^\ell$ using $(u_1, u_2, \ldots, u_\ell)$ with the same energy. Therefore, the minimum energy of a balanced orthogonal graph drawing in $\mathbb{R}^n$ is equal to the minimum energy of an orthogonal graph drawing in $\mathbb{R}^{n+1}$, and this minimum is $\lambda_2 + \cdots + \lambda_{n+1}$. $\square$

Since $\mathbf{1}$ spans the nullspace of L, using u_1 (which belongs to $\mathrm{Ker}\, L$) as one of the vectors in R would have the effect that all points representing vertices of G would have the same first coordinate. This would mean that the drawing lives in a hyperplane in $\mathbb{R}^n$, which is undesirable, especially when $n = 2$, where all vertices would be collinear. This is why we omit the first eigenvector u_1.

Observe that for any orthogonal $n \times n$ matrix Q, since

$$\mathrm{tr}(R^\top L R) = \mathrm{tr}(Q^\top R^\top L R Q),$$

the matrix RQ also yields a minimum orthogonal graph drawing. This amounts to applying the rigid motion $Q^\top$ to the rows of R.

In summary, if $\lambda_2 > 0$, an automatic method for drawing a graph in $\mathbb{R}^2$ is this:

1. Compute the two smallest nonzero eigenvalues $\lambda_2 \leq \lambda_3$ of the graph Laplacian L (it is possible that $\lambda_3 = \lambda_2$ if λ_2 is a multiple eigenvalue);
2. Compute two unit eigenvectors u_2, u_3 associated with λ_2 and λ_3, and let $R = [u_2\, u_3]$ be the $m \times 2$ matrix having u_2 and u_3 as columns.
3. Place vertex v_i at the point whose coordinates is the ith row of R, that is, (R_{i1}, R_{i2}).

This method generally gives pleasing results, but beware that there is no guarantee that distinct nodes are assigned distinct images since R can have identical rows. This does not seem to happen often in practice.

21.2 Examples of Graph Drawings

We now give a number of examples using `Matlab`. Some of these are borrowed or adapted from Spielman [163].

Example 1. Consider the graph with four nodes whose adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

We use the following program to compute u_2 and u_3:

```
A = [0 1 1 0; 1 0 0 1; 1 0 0 1; 0 1 1 0];
D = diag(sum(A));
L = D - A;
[v, e] = eigs(L);
gplot(A, v(:,[3 2]))
hold on;
gplot(A, v(:,[3 2]),'o')
```

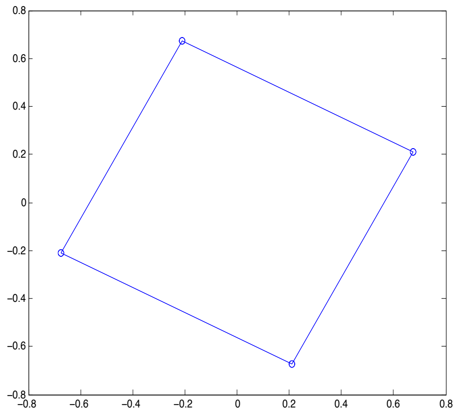


Figure 21.1: Drawing of the graph from Example 1.

The graph of Example 1 is shown in Figure 21.1. The function `eigs(L)` computes the six largest eigenvalues of L in decreasing order, and corresponding eigenvectors. It turns out that $\lambda_2 = \lambda_3 = 2$ is a double eigenvalue.

Example 2. Consider the graph G_2 shown in Figure 20.3 given by the adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

We use the following program to compute u_2 and u_3:

```
A = [0 1 1 0 0; 1 0 1 1 1; 1 1 0 1 0; 0 1 1 0 1; 0 1 0 1 0];
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on
gplot(A, v(:, [2 3]),'o')
```

The function `eig(L)` (with no `s` at the end) computes the eigenvalues of L in increasing order. The result of drawing the graph is shown in Figure 21.2. Note that node v_2 is assigned to the point $(0, 0)$, so the difference between this drawing and the drawing in Figure 20.3 is that the drawing of Figure 21.2 is not convex.

Example 3. Consider the ring graph defined by the adjacency matrix A given in the `Matlab` program shown below:

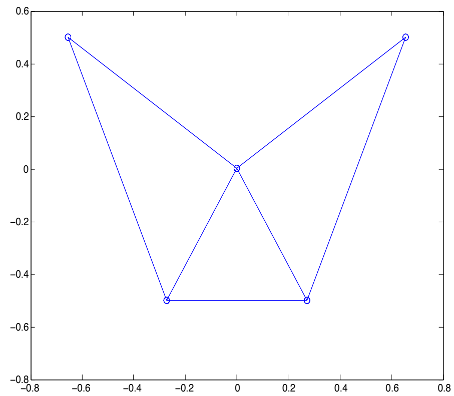


Figure 21.2: Drawing of the graph from Example 2.

```
A = diag(ones(1, 11),1);
A = A + A';
A(1, 12) = 1; A(12, 1) = 1;
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on
gplot(A, v(:, [2 3]),'o')
```

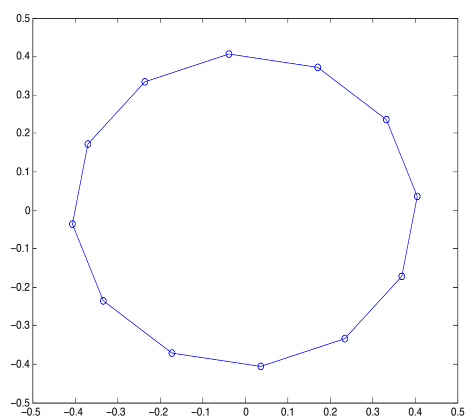


Figure 21.3: Drawing of the graph from Example 3.

Observe that we get a very nice ring; see Figure 21.3. Again $\lambda_2 = 0.2679$ is a double eigenvalue (and so are the next pairs of eigenvalues, except the last, $\lambda_{12} = 4$).

Example 4. In this example adapted from Spielman, we generate 20 randomly chosen points in the unit square, compute their Delaunay triangulation, then the adjacency matrix of the corresponding graph, and finally draw the graph using the second and third eigenvalues of the Laplacian.

```
A = zeros(20,20);
xy = rand(20, 2);
trigs = delaunay(xy(:,1), xy(:,2));
elemtrig = ones(3) - eye(3);
for i = 1:length(trigs),
 A(trigs(i,:),trigs(i,:)) = elemtrig;
end
A = double(A >0);
gplot(A,xy)
D = diag(sum(A));
L = D - A;
[v, e] = eigs(L, 3, 'sm');
figure(2)
gplot(A, v(:, [2 1]))
hold on
gplot(A, v(:, [2 1]),'o')
```

The Delaunay triangulation of the set of 20 points and the drawing of the corresponding graph are shown in Figure 21.4. The graph drawing on the right looks nicer than the graph on the left but is is no longer planar.

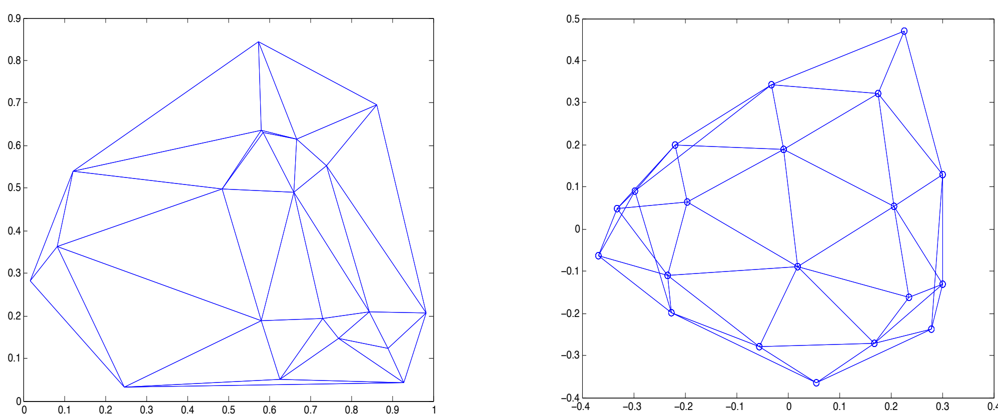


Figure 21.4: Delaunay triangulation (left) and drawing of the graph from Example 4 (right).

Example 5. Our last example, also borrowed from Spielman [163], corresponds to the skeleton of the "Buckyball," a geodesic dome invented by the architect Richard Buckminster

Fuller (1895–1983). The Montréal Biosphère is an example of a geodesic dome designed by Buckminster Fuller.

```
A = full(bucky);
D = diag(sum(A));
L = D - A;
[v, e] = eig(L);
gplot(A, v(:, [2 3]))
hold on;
gplot(A,v(:, [2 3]), 'o')
```

Figure 21.5 shows a graph drawing of the Buckyball. This picture seems a bit squashed for two reasons. First, it is really a 3-dimensional graph; second, $\lambda_2 = 0.2434$ is a triple eigenvalue. (Actually, the Laplacian of L has many multiple eigenvalues.) What we should really do is to plot this graph in $\mathbb{R}^3$ using three orthonormal eigenvectors associated with λ_2.

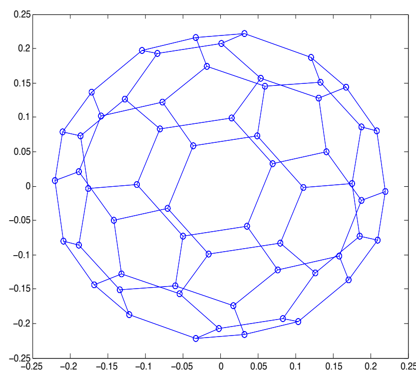


Figure 21.5: Drawing of the graph of the Buckyball.

A 3D picture of the graph of the Buckyball is produced by the following `Matlab` program, and its image is shown in Figure 21.6. It looks better!

```
[x, y] = gplot(A, v(:, [2 3]));
[x, z] = gplot(A, v(:, [2 4]));
plot3(x,y,z)
```

21.3 Summary

The main concepts and results of this chapter are listed below:

- Graph drawing.

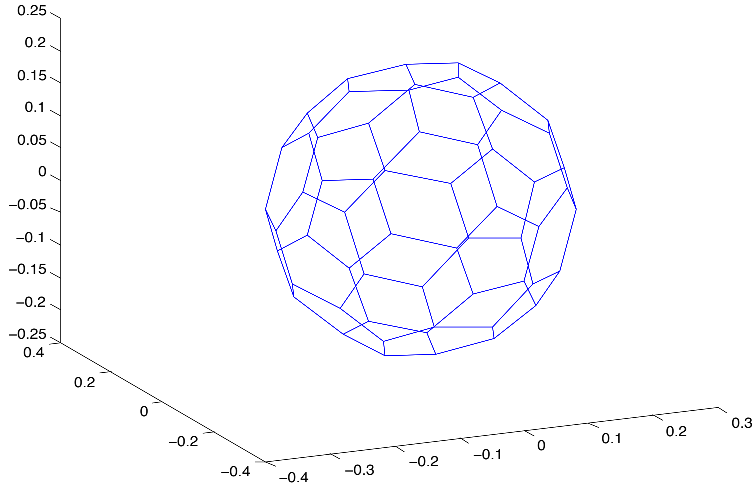


Figure 21.6: Drawing of the graph of the Buckyball in $\mathbb{R}^3$.

- Matrix of a graph drawing.
- Balanced graph drawing.
- Energy $\mathcal{E}(R)$ of a graph drawing.
- Orthogonal graph drawing.
- Delaunay triangulation.
- Buckyball.

Chapter 22

Singular Value Decomposition and Polar Form

22.1 Properties of $f^* \circ f$

In this section we assume that we are dealing with real Euclidean spaces. Let $f\colon E \to E$ be any linear map. In general, it may not be possible to diagonalize f. We show that every linear map can be diagonalized if we are willing to use *two* orthonormal bases. This is the celebrated *singular value decomposition (SVD)*. A close cousin of the SVD is the *polar form* of a linear map, which shows how a linear map can be decomposed into its purely rotational component (perhaps with a flip) and its purely stretching part.

The key observation is that $f^* \circ f$ is self-adjoint since

$$\langle (f^* \circ f)(u), v\rangle = \langle f(u), f(v)\rangle = \langle u, (f^* \circ f)(v)\rangle.$$

Similarly, $f \circ f^*$ is self-adjoint.

The fact that $f^* \circ f$ and $f \circ f^*$ are self-adjoint is very important, because by Theorem 17.8, it implies that *$f^* \circ f$ and $f \circ f^*$ can be diagonalized and that they have real eigenvalues.* In fact, these *eigenvalues are all nonnegative* as shown in the following proposition.

Proposition 22.1. *The eigenvalues of $f^* \circ f$ and $f \circ f^*$ are nonnegative.*

Proof. If u is an eigenvector of $f^* \circ f$ for the eigenvalue λ, then

$$\langle (f^* \circ f)(u), u\rangle = \langle f(u), f(u)\rangle$$

and

$$\langle (f^* \circ f)(u), u\rangle = \lambda\langle u, u\rangle,$$

and thus

$$\lambda\langle u, u\rangle = \langle f(u), f(u)\rangle,$$

which implies that $\lambda \geq 0$, since $\langle -, -\rangle$ is positive definite. A similar proof applies to $f \circ f^*$. $\square$

Thus, the eigenvalues of $f^* \circ f$ are of the form $\sigma_1^2, \ldots, \sigma_r^2$ or 0, where $\sigma_i > 0$, and similarly for $f \circ f^*$.

The above considerations also apply to any linear map $f\colon E \to F$ between two Euclidean spaces $(E, \langle -, - \rangle_1)$ and $(F, \langle -, - \rangle_2)$. Recall that the adjoint $f^*\colon F \to E$ of f is the unique linear map f^* such that

$$\langle f(u), v\rangle_2 = \langle u, f^*(v)\rangle_1, \quad \text{for all } u \in E \text{ and all } v \in F.$$

Then $f^* \circ f$ and $f \circ f^*$ are self-adjoint (the proof is the same as in the previous case), and the eigenvalues of $f^* \circ f$ and $f \circ f^*$ are nonnegative.

Proof. If λ is an eigenvalue of $f^* \circ f$ and u ($\neq 0$) is a corresponding eigenvector, we have

$$\langle (f^* \circ f)(u), u\rangle_1 = \langle f(u), f(u)\rangle_2,$$

and also

$$\langle (f^* \circ f)(u), u\rangle_1 = \lambda \langle u, u\rangle_1,$$

so

$$\lambda \langle u, u\rangle_1, = \langle f(u), f(u)\rangle_2,$$

which implies that $\lambda \geq 0$. A similar proof applies to $f \circ f^*$. ☐

The situation is even better, since we will show shortly that $f^* \circ f$ and $f \circ f^*$ have the same nonzero eigenvalues.

Remark: Given any two linear maps $f\colon E \to F$ and $g\colon F \to E$, where $\dim(E) = n$ and $\dim(F) = m$, it can be shown that

$$\lambda^m \det(\lambda\, I_n - g \circ f) = \lambda^n \det(\lambda\, I_m - f \circ g),$$

and thus $g \circ f$ and $f \circ g$ always have the same nonzero eigenvalues; see Problem 15.14.

Definition 22.1. Given any linear map $f\colon E \to F$, the square roots $\sigma_i > 0$ of the positive eigenvalues of $f^* \circ f$ (and $f \circ f^*$) are called the *singular values of f*.

Definition 22.2. A self-adjoint linear map $f\colon E \to E$ whose eigenvalues are nonnegative is called *positive semidefinite* (or *positive*), and if f is also invertible, f is said to be *positive definite*. In the latter case, every eigenvalue of f is strictly positive.

The following proposition shows that the conditions on the eigenvalues of a self-adjoint linear map used to define the notion of a positive definite linear map is equivalent to the condition used in Definition 8.4. A similar but weaker condition is equivalent to the notion of self-adjoint positive semidefinite linear map.

Proposition 22.2. *Let $f\colon E \to E$ be a self-adjoint linear map, where E is a Euclidean space of finite dimension with inner product $\langle -.- \rangle$.*

(1) The eigenvalues of f are strictly positive iff

$$\langle f(u), u\rangle > 0 \quad \textit{for all } u \neq 0.$$

(2) The eigenvalues of f are nonnegative iff

$$\langle f(u), u\rangle \geq 0 \quad \textit{for all } u \neq 0.$$

Proof. Since f is self-adjoint, by the spectral theorem (Theorem 17.8), f has real eigenvalues $\lambda_1, \ldots, \lambda_n$, and there is some orthonormal basis $(e_1, \ldots, e_n)$, where e_i is an eigenvector for λ_i. With respect to this basis, every vector $u \in E$ can be written in a unique way as $u = \sum_{i=1}^n x_i u_i$ for some $x_i \in \mathbb{R}$. Since each e_i is eigenvector associated with $\lambda_i \in \mathbb{R}$, we have

$$f\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^n x_i f(e_i) = \sum_{i=1}^n \lambda_i x_i e_i,$$

and using the bilinearity of the inner product, we have

$$\begin{aligned}\langle f(u), u\rangle &= \left\langle f\left(\sum_{i=1}^n x_i e_i\right), \sum_{j=1}^n x_j e_j\right\rangle \\ &= \left\langle \sum_{i=1}^n \lambda_i x_i e_i, \sum_{j=1}^n x_j e_j\right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i x_i x_j \langle e_i, e_j\rangle,\end{aligned}$$

and since $(e_1, \ldots, e_n)$, is an orthonormal basis, we obtain

$$\langle f(u), u\rangle = \sum_{i=1}^n \lambda_i x_i^2. \tag{$\dagger$}$$

(1) If $\lambda_i > 0$ for $i = 1, \ldots, n$, for any $u \neq 0$, we have $x_i \neq 0$ for some i, so $\langle f(u), u\rangle = \sum_{i=1}^n \lambda_i x_i^2 > 0$.

Conversely, if $\langle f(u), u\rangle > 0$ for all $u \neq 0$, by picking $u = e_i$, we get

$$\langle f(e_i), e_i\rangle = \langle \lambda_i e_i, e_i\rangle = \lambda_i \langle e_i, e_i\rangle = \lambda_i,$$

so $\lambda_i > 0$ for $i = 1, \ldots, n$.

(2) If $\lambda_i \geq 0$ for $i = 1, \ldots, n$, for any $u \neq 0$, then $\langle f(u), u\rangle = \sum_{i=1}^n \lambda_i x_i^2 \geq 0$.

Conversely, if $\langle f(u), u\rangle \geq 0$ for all $u \neq 0$, by picking $u = e_i$, we get

$$\langle f(e_i), e_i\rangle = \langle \lambda_i e_i, e_i\rangle = \lambda_i \langle e_i, e_i\rangle = \lambda_i,$$

so $\lambda_i \geq 0$ for $i = 1, \ldots, n$. ∎

Proposition 22.2 also holds for self-adjoint linear maps on a complex vector space with a Hermitian inner product. The proof is essentially the same and is left as an exercise to the reader.

The version of Proposition 22.2 for matrices follows immediately.

Proposition 22.3. *Let A be a real $n \times n$ symmetric matrix.*

(1) The eigenvalues of A are strictly positive iff

$$u^\top A u > 0 \quad \textit{for all } u \neq 0.$$

(2) The eigenvalues of A are nonnegative iff

$$u^\top A u \geq 0 \quad \textit{for all } u \neq 0.$$

It is important to note that Proposition 22.3 is false for nonsymmetric matrices.

Example 22.1. The matrix

$$A = \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}$$

has the positive eigenvalues $(1, 1)$, but

$$\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} -3 \\ -1 \end{pmatrix} = -2.$$

Example 22.2. The matrix

$$A = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$$

has the complex eigenvalues $1 + 2i, 1 - 2i$, and yet

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} x - 2y \\ 2x + y \end{pmatrix} = x^2 + y^2,$$

so $u^\top A u > 0$ for all $u \neq 0$.

Since $u^\top A u$ is a scalar, if A is a skew symmetric matrix ($A^\top = -A$), then we see that

$$u^\top A u = 0 \quad \text{for all } u \in \mathbb{R}.$$

Therefore, if A is a real $n \times n$ matrix then

$$u^\top A u = u^\top H(A) u \quad \text{for all } u \in \mathbb{R},$$

where $H(A) = (1/2)(A + A^\top)$ is the symmetric part of A. This explains why the notion of a positive definite matrix is only interesting for symmetric matrices. But but one should also be aware that even if a nonsymmetric matrix A has "well-behaved" eigenvalues, its symmetric part $H(A)$ may *not* be positive definite.

Example 22.3. The matrix

$$A = \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}$$

of Example 22.1 has positive eigenvalues $(1, 1)$, but its symmetric part

$$H(A) = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

is *not* positive definite, since its eigenvalues are $-1, 3$.

Beware that if A is a complex skew-Hermitian matrix, which means that $A^* = -A$, then

$$(u^* A u)^* = -u^* A u,$$

but this only implies that *the real part of* $u^* A u$ *is zero*. So for any arbitrary complex square matrix A, in general,

$$u^* A u \neq u^* H(A) u,$$

where $H(A) = (1/2)(A + A^*)$.

If $f \colon E \to F$ is any linear map, we just showed that $f^* \circ f$ and $f \circ f^*$ are positive semidefinite self-adjoint linear maps. This fact has the remarkable consequence that every linear map has two important decompositions:

1. The polar form.

2. The singular value decomposition (SVD).

The wonderful thing about the singular value decomposition is that there exist two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ such that, with respect to these bases, f is a diagonal matrix consisting of the singular values of f or 0. Thus, in some sense, f can always be diagonalized with respect to *two* orthonormal bases. The SVD is also a useful tool for solving overdetermined linear systems in the least squares sense and for data analysis, as we show later on.

First we show some useful relationships between the kernels and the images of f, f^*, $f^* \circ f$, and $f \circ f^*$. Recall that if $f \colon E \to F$ is a linear map, the *image* $\mathrm{Im}\, f$ *of* f is the subspace $f(E)$ of F, and the *rank of* f is the dimension $\dim(\mathrm{Im}\, f)$ of its image. Also recall that (Theorem 6.16)

$$\dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f) = \dim(E),$$

and that (Propositions 12.11 and 14.13) for every subspace W of E,

$$\dim(W) + \dim(W^{\perp}) = \dim(E).$$

Proposition 22.4. *Given any two Euclidean spaces E and F, where E has dimension n and F has dimension m, for any linear map $f\colon E \to F$, we have*

$$\begin{aligned}
\mathrm{Ker}\, f &= \mathrm{Ker}\,(f^* \circ f),\\
\mathrm{Ker}\, f^* &= \mathrm{Ker}\,(f \circ f^*),\\
\mathrm{Ker}\, f &= (\mathrm{Im}\, f^*)^{\perp},\\
\mathrm{Ker}\, f^* &= (\mathrm{Im}\, f)^{\perp},\\
\dim(\mathrm{Im}\, f) &= \dim(\mathrm{Im}\, f^*),
\end{aligned}$$

and f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank.

Proof. To simplify the notation, we will denote the inner products on E and F by the same symbol $\langle -, - \rangle$ (to avoid subscripts). If $f(u) = 0$, then $(f^* \circ f)(u) = f^*(f(u)) = f^*(0) = 0$, and so $\mathrm{Ker}\, f \subseteq \mathrm{Ker}\,(f^* \circ f)$. By definition of f^*, we have

$$\langle f(u), f(u)\rangle = \langle (f^* \circ f)(u), u\rangle$$

for all $u \in E$. If $(f^* \circ f)(u) = 0$, since $\langle -, - \rangle$ is positive definite, we must have $f(u) = 0$, and so $\mathrm{Ker}\,(f^* \circ f) \subseteq \mathrm{Ker}\, f$. Therefore,

$$\mathrm{Ker}\, f = \mathrm{Ker}\,(f^* \circ f).$$

The proof that $\mathrm{Ker}\, f^* = \mathrm{Ker}\,(f \circ f^*)$ is similar.

By definition of f^*, we have

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle \quad \text{for all } u \in E \text{ and all } v \in F. \tag{$*$}$$

This immediately implies that

$$\mathrm{Ker}\, f = (\mathrm{Im}\, f^*)^{\perp} \quad \text{and} \quad \mathrm{Ker}\, f^* = (\mathrm{Im}\, f)^{\perp}.$$

Let us explain why $\mathrm{Ker}\, f = (\mathrm{Im}\, f^*)^{\perp}$, the proof of the other equation being similar.

Because the inner product is positive definite, for every $u \in E$, we have

- $u \in \mathrm{Ker}\, f$
- iff $f(u) = 0$
- iff $\langle f(u), v\rangle = 0$ for all v,
- by $(*)$ iff $\langle u, f^*(v)\rangle = 0$ for all v,
- iff $u \in (\mathrm{Im}\, f^*)^{\perp}$.

Since

$$\dim(\mathrm{Im}\, f) = n - \dim(\mathrm{Ker}\, f)$$

and

$$\dim(\mathrm{Im}\, f^*) = n - \dim((\mathrm{Im}\, f^*)^\perp),$$

from

$$\mathrm{Ker}\, f = (\mathrm{Im}\, f^*)^\perp$$

we also have

$$\dim(\mathrm{Ker}\, f) = \dim((\mathrm{Im}\, f^*)^\perp),$$

from which we obtain

$$\dim(\mathrm{Im}\, f) = \dim(\mathrm{Im}\, f^*).$$

Since

$$\dim(\mathrm{Ker}\,(f^* \circ f)) + \dim(\mathrm{Im}\,(f^* \circ f)) = \dim(E),$$

$\mathrm{Ker}\,(f^* \circ f) = \mathrm{Ker}\, f$ and $\mathrm{Ker}\, f = (\mathrm{Im}\, f^*)^\perp$, we get

$$\dim((\mathrm{Im}\, f^*)^\perp) + \dim(\mathrm{Im}\,(f^* \circ f)) = \dim(E).$$

Since

$$\dim((\mathrm{Im}\, f^*)^\perp) + \dim(\mathrm{Im}\, f^*) = \dim(E),$$

we deduce that

$$\dim(\mathrm{Im}\, f^*) = \dim(\mathrm{Im}\,(f^* \circ f)).$$

A similar proof shows that

$$\dim(\mathrm{Im}\, f) = \dim(\mathrm{Im}\,(f \circ f^*)).$$

Consequently, f, f^*, $f^* \circ f$, and $f \circ f^*$ have the same rank. $\square$

22.2 Singular Value Decomposition for Square Matrices

We will now prove that every square matrix has an SVD. Stronger results can be obtained if we first consider the polar form and then derive the SVD from it (there are uniqueness properties of the polar decomposition). For our purposes, uniqueness results are not as important so we content ourselves with existence results, whose proofs are simpler. Readers interested in a more general treatment are referred to Gallier [72].

The early history of the singular value decomposition is described in a fascinating paper by Stewart [165]. The SVD is due to Beltrami and Camille Jordan independently (1873, 1874). Gauss is the grandfather of all this, for his work on least squares (1809, 1823) (but Legendre also published a paper on least squares!). Then come Sylvester, Schmidt, and

Hermann Weyl. Sylvester's work was apparently "opaque." He gave a computational method to find an SVD. Schmidt's work really has to do with integral equations and symmetric and asymmetric kernels (1907). Weyl's work has to do with perturbation theory (1912). Autonne came up with the polar decomposition (1902, 1915). Eckart and Young extended SVD to rectangular matrices (1936, 1939).

Theorem 22.5. *(Singular value decomposition) For every real $n \times n$ matrix A there are two orthogonal matrices U and V and a diagonal matrix D such that $A = VDU^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & & \dots & \\ & \sigma_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & \sigma_n \end{pmatrix},$$

where $\sigma_1, \dots, \sigma_r$ are the singular values of A, i.e., the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $A\,A^\top$, and $\sigma_{r+1} = \cdots = \sigma_n = 0$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $A\,A^\top$.

Proof. Since $A^\top A$ is a symmetric matrix, in fact, a positive semidefinite matrix, there exists an orthogonal matrix U such that

$$A^\top A = UD^2U^\top,$$

with $D = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$, where $\sigma_1^2, \dots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A; that is, $\sigma_1, \dots, \sigma_r$ are the singular values of A. It follows that

$$U^\top A^\top AU = (AU)^\top AU = D^2,$$

and if we let f_j be the jth column of AU for $j = 1, \dots, n$, then we have

$$\langle f_i, f_j\rangle = \sigma_i^2\delta_{ij}, \quad 1 \le i, j \le r$$

and

$$f_j = 0, \quad r + 1 \le j \le n.$$

If we define $(v_1, \dots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \le j \le r,$$

then we have

$$\langle v_i, v_j\rangle = \delta_{ij}, \quad 1 \le i, j \le r,$$

so complete $(v_1, \dots, v_r)$ into an orthonormal basis $(v_1, \dots, v_r, v_{r+1}, \dots, v_n)$ (for example, using Gram–Schmidt). Now since $f_j = \sigma_j v_j$ for $j = 1 \dots, r$, we have

$$\langle v_i, f_j\rangle = \sigma_j\langle v_i, v_j\rangle = \sigma_j\delta_{i,j}, \quad 1 \le i \le n,\ 1 \le j \le r$$

and since $f_j = 0$ for $j = r+1, \ldots, n$,

$$\langle v_i, f_j \rangle = 0 \quad 1 \leq i \leq n,\ r+1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_n$, then V is orthogonal and the above equations prove that

$$V^\top A U = D,$$

which yields $A = VDU^\top$, as required.

The equation $A = VDU^\top$ implies that

$$A^\top A = UD^2U^\top, \quad AA^\top = VD^2V^\top,$$

which shows that $A^\top A$ and $AA^\top$ have the same eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $AA^\top$. $\square$

Example 22.4. Here is a simple example of how to use the proof of Theorem 22.5 to obtain an SVD decomposition. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$. Then $A^\top = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, $A^\top A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, and $AA^\top = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$. A simple calculation shows that the eigenvalues of $A^\top A$ are 2 and 0, and for the eigenvalue 2, a unit eigenvector is $\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, while a unit eigenvector for the eigenvalue 0 is $\begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$. Observe that the singular values are $\sigma_1 = \sqrt{2}$ and $\sigma_2 = 0$. Furthermore, $U = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = U^\top$. To determine V, the proof of Theorem 22.5 tells us to first calculate

$$AU = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix},$$

and then set

$$v_1 = (1/\sqrt{2}) \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Once v_1 is determined, since $\sigma_2 = 0$, we have the freedom to choose v_2 such that (v_1, v_2) forms an orthonormal basis for $\mathbb{R}^2$. Naturally, we chose $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and set $V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The columns of V are unit eigenvectors of $AA^\top$, but finding V by computing unit eigenvectors of $AA^\top$ does not guarantee that these vectors are consistent with U so that $A = V\Sigma U^\top$. Thus one has to use AU instead. We leave it to the reader to check that

$$A = V \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} U^\top.$$

Theorem 22.5 suggests the following definition.

Definition 22.3. A triple (U, D, V) such that $A = V D U^\top$, where U and V are orthogonal and D is a diagonal matrix whose entries are nonnegative (it is positive semidefinite) is called a *singular value decomposition (SVD) of* A. If $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, it is customary to assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

The `Matlab` command for computing an SVD $A = VDU^\top$ of a matrix A is
`[V, D, U] = svd(A)`.

The proof of Theorem 22.5 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, where $(u_1, \ldots, u_n)$ are eigenvectors of $A^\top A$ and $(v_1, \ldots, v_n)$ are eigenvectors of $AA^\top$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\mathrm{Im}\, A^\top$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\mathrm{Ker}\, A$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\mathrm{Im}\, A$, and $(v_{r+1}, \ldots, v_n)$ is an orthonormal basis of $\mathrm{Ker}\, A^\top$.

Using a remark made in Chapter 4, if we denote the columns of U by $u_1, \ldots, u_n$ and the columns of V by $v_1, \ldots, v_n$, then we can write

$$A = V D U^\top = \sigma_1 v_1 u_1^\top + \cdots + \sigma_r v_r u_r^\top,$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. As a consequence, if r is a lot smaller than n (we write $r \ll n$), we see that A can be reconstructed from U and V using a much smaller number of elements. This idea will be used to provide "low-rank" approximations of a matrix. The idea is to keep only the k top singular values for some suitable $k \ll r$ for which $\sigma_{k+1}, \ldots, \sigma_r$ are very small.

Remarks:

(1) In Strang [170] the matrices U, V, D are denoted by $U = Q_2$, $V = Q_1$, and $D = \Sigma$, and an SVD is written as $A = Q_1 \Sigma Q_2^\top$. This has the advantage that Q_1 comes before Q_2 in $A = Q_1 \Sigma Q_2^\top$. This has the disadvantage that A maps the columns of Q_2 (eigenvectors of $A^\top A$) to multiples of the columns of Q_1 (eigenvectors of $A\, A^\top$).

(2) Algorithms for actually computing the SVD of a matrix are presented in Golub and Van Loan [80], Demmel [48], and Trefethen and Bau [176], where the SVD and its applications are also discussed quite extensively.

(3) If A is a symmetric matrix, then in general, there is no SVD $V \Sigma U^\top$ of A with $V = U$. However, if A is positive semidefinite, then the eigenvalues of A are nonnegative, and so the nonzero eigenvalues of A are equal to the singular values of A and SVDs of A are of the form

$$A = V \Sigma V^\top.$$

(4) The SVD also applies to complex matrices. In this case, for every complex $n \times n$ matrix A, there are two unitary matrices U and V and a diagonal matrix D such that

$$A = V D U^*,$$

where D is a diagonal matrix consisting of real entries $\sigma_1, \ldots, \sigma_n$, where $\sigma_1 \geq \cdots \geq \sigma_r$ are the singular values of A, i.e., the positive square roots of the nonzero eigenvalues of $A^* A$ and $A\, A^*$, and $\sigma_{r+1} = \ldots = \sigma_n = 0$.

22.3 Polar Form for Square Matrices

A notion closely related to the SVD is the polar form of a matrix.

Definition 22.4. A pair (R, S) such that $A = RS$ with R orthogonal and S symmetric positive semidefinite is called a *polar decomposition of A*.

Theorem 22.5 implies that for every real $n \times n$ matrix A, there is some orthogonal matrix R and some positive semidefinite symmetric matrix S such that

$$A = RS.$$

This is easy to show and we will prove it below. Furthermore, R, S are unique if A is invertible, but this is harder to prove; see Problem 22.9.

For example, the matrix

$$A = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

is both orthogonal and symmetric, and $A = RS$ with $R = A$ and $S = I$, which implies that some of the eigenvalues of A are negative.

Remark: In the complex case, the polar decomposition states that for every complex $n \times n$ matrix A, there is some unitary matrix U and some positive semidefinite Hermitian matrix H such that

$$A = UH.$$

It is easy to go from the polar form to the SVD, and conversely.

Given an SVD $A = V D U^{\top}$, let $R = V U^{\top}$ and $S = U D U^{\top}$. It is clear that R is orthogonal and that S is positive semidefinite symmetric, and

$$RS = V U^{\top} U D U^{\top} = V D U^{\top} = A.$$

Example 22.5. Recall from Example 22.4 that $A = VDU^{\top}$ where $V = I_2$ and

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}, \qquad D = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Set $R = VU^{\top} = U$ and

$$S = UDU^{\top} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Since $S = \frac{1}{\sqrt{2}} A^{\top} A$, S has eigenvalues $\sqrt{2}$ and 0. We leave it to the reader to check that $A = RS$.

Going the other way, given a polar decomposition $A = R_1 S$, where R_1 is orthogonal and S is positive semidefinite symmetric, there is an orthogonal matrix R_2 and a positive semidefinite diagonal matrix D such that $S = R_2 D\, R_2^\top$, and thus

$$A = R_1 R_2 D\, R_2^\top = V D\, U^\top,$$

where $V = R_1 R_2$ and $U = R_2$ are orthogonal.

Example 22.6. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ and $A = R_1 S$, where $R_1 = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$ and $S = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$. This is the polar decomposition of Example 22.5. Observe that

$$S = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = R_2 D R_2^\top.$$

Set $U = R_2$ and $V = R_1 R_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ to obtain the SVD decomposition of Example 22.4.

The eigenvalues and the singular values of a matrix are typically not related in any obvious way. For example, the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 2 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix}$$

has the eigenvalue 1 with multiplicity n, but its singular values, $\sigma_1 \geq \cdots \geq \sigma_n$, which are the positive square roots of the eigenvalues of the matrix $B = A^\top A$ with

$$B = \begin{pmatrix} 1 & 2 & 0 & 0 & \dots & 0 & 0 \\ 2 & 5 & 2 & 0 & \dots & 0 & 0 \\ 0 & 2 & 5 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & 5 & 2 & 0 \\ 0 & 0 & \dots & 0 & 2 & 5 & 2 \\ 0 & 0 & \dots & 0 & 0 & 2 & 5 \end{pmatrix}$$

have a wide spread, since

$$\frac{\sigma_1}{\sigma_n} = \mathrm{cond}_2(A) \geq 2^{n-1}.$$

If A is a complex $n \times n$ matrix, the eigenvalues $\lambda_1, \ldots, \lambda_n$ and the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of A are not unrelated, since

$$\sigma_1^2 \cdots \sigma_n^2 = \det(A^* A) = |\det(A)|^2$$

and

$$|\lambda_1| \cdots |\lambda_n| = |\det(A)|,$$

so we have

$$|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n.$$

More generally, Hermann Weyl proved the following remarkable theorem:

Theorem 22.6. *(Weyl's inequalities, 1949) For any complex $n \times n$ matrix, A, if $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A and $\sigma_1, \ldots, \sigma_n \in \mathbb{R}_+$ are the singular values of A, listed so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$ and $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, then*

$$\begin{aligned} &|\lambda_1| \cdots |\lambda_n| = \sigma_1 \cdots \sigma_n \quad \textit{and} \\ &|\lambda_1| \cdots |\lambda_k| \leq \sigma_1 \cdots \sigma_k, \quad \textit{for} \quad k = 1, \ldots, n-1. \end{aligned}$$

A proof of Theorem 22.6 can be found in Horn and Johnson [96], Chapter 3, Section 3.3, where more inequalities relating the eigenvalues and the singular values of a matrix are given.

Theorem 22.5 can be easily extended to rectangular $m \times n$ matrices, as we show in the next section. For various versions of the SVD for rectangular matrices, see Strang [170] Golub and Van Loan [80], Demmel [48], and Trefethen and Bau [176].

22.4 Singular Value Decomposition for Rectangular Matrices

Here is the generalization of Theorem 22.5 to rectangular matrices.

Theorem 22.7. *(Singular value decomposition) For every real $m \times n$ matrix A, there are two orthogonal matrices U $(n \times n)$ and V $(m \times m)$ and a diagonal $m \times n$ matrix D such that $A = VDU^\top$, where D is of the form*

$$D = \begin{pmatrix} \sigma_1 & & \ldots & \\ & \sigma_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \sigma_n \\ 0 & \vdots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \ldots & 0 \end{pmatrix} \quad \textit{or} \;\; D = \begin{pmatrix} \sigma_1 & & \ldots & & 0 & \ldots & 0 \\ & \sigma_2 & \ldots & & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \ldots & \sigma_m & 0 & \ldots & 0 \end{pmatrix},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of A, i.e. the (positive) square roots of the nonzero eigenvalues of $A^\top A$ and $A\,A^\top$, and $\sigma_{r+1} = \ldots = \sigma_p = 0$, where $p = \min(m, n)$. The columns of U are eigenvectors of $A^\top A$, and the columns of V are eigenvectors of $A\,A^\top$.

Proof. As in the proof of Theorem 22.5, since $A^\top A$ is symmetric positive semidefinite, there exists an $n \times n$ orthogonal matrix U such that

$$A^\top A = U\Sigma^2 U^\top,$$

with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of $A^\top A$, and where r is the rank of A. Observe that $r \leq \min\{m, n\}$, and AU is an $m \times n$ matrix. It follows that

$$U^\top A^\top AU = (AU)^\top AU = \Sigma^2,$$

and if we let $f_j \in \mathbb{R}^m$ be the jth column of AU for $j = 1, \ldots, n$, then we have

$$\langle f_i, f_j \rangle = \sigma_i^2 \delta_{ij}, \quad 1 \leq i, j \leq r$$

and

$$f_j = 0, \quad r + 1 \leq j \leq n.$$

If we define $(v_1, \ldots, v_r)$ by

$$v_j = \sigma_j^{-1} f_j, \quad 1 \leq j \leq r,$$

then we have

$$\langle v_i, v_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq r,$$

so complete $(v_1, \ldots, v_r)$ into an orthonormal basis $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_m)$ (for example, using Gram–Schmidt).

Now since $f_j = \sigma_j v_j$ for $j = 1 \ldots, r$, we have

$$\langle v_i, f_j \rangle = \sigma_j \langle v_i, v_j \rangle = \sigma_j \delta_{i,j}, \quad 1 \leq i \leq m,\ 1 \leq j \leq r$$

and since $f_j = 0$ for $j = r + 1, \ldots, n$, we have

$$\langle v_i, f_j \rangle = 0 \quad 1 \leq i \leq m,\ r + 1 \leq j \leq n.$$

If V is the matrix whose columns are $v_1, \ldots, v_m$, then V is an $m \times m$ orthogonal matrix and if $m \geq n$, we let

$$D = \begin{pmatrix} \Sigma \\ 0_{m-n} \end{pmatrix} = \begin{pmatrix} \sigma_1 & & \ldots & \\ & \sigma_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \sigma_n \\ 0 & \vdots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & \ldots & 0 \end{pmatrix},$$

else if $n \geq m$, then we let

$$D = \begin{pmatrix} \sigma_1 & & \dots & & 0 & \dots & 0 \\ & \sigma_2 & \dots & & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ & & \dots & \sigma_m & 0 & \dots & 0 \end{pmatrix}.$$

In either case, the above equations prove that

$$V^\top AU = D,$$

which yields $A = VDU^\top$, as required.

The equation $A = VDU^\top$ implies that

$$A^\top A = UD^\top DU^\top = U\mathrm{diag}(\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{n-r})U^\top$$

and

$$AA^\top = VDD^\top V^\top = V\mathrm{diag}(\sigma_1^2, \dots, \sigma_r^2, \underbrace{0, \dots, 0}_{m-r})V^\top,$$

which shows that $A^\top A$ and $AA^\top$ have the same nonzero eigenvalues, that the columns of U are eigenvectors of $A^\top A$, and that the columns of V are eigenvectors of $AA^\top$. $\square$

A triple (U, D, V) such that $A = VDU^\top$ is called a *singular value decomposition (SVD) of A*. If $D = \mathrm{diag}(\sigma_1, \dots, \sigma_p)$ (with $p = \min(m, n)$), it is customary to assume that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$.

Example 22.7. Let $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$. Then $A^\top = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ $A^\top A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, and $AA^\top = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. The reader should verify that $A^\top A = U\Sigma^2 U^\top$ where $\Sigma^2 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ and $U = U^\top = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$. Since $AU = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, set $v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, and complete an orthonormal basis for $\mathbb{R}^3$ by assigning $v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, and $v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Thus $V = I_3$, and the reader should verify that $A = VDU^\top$, where $D = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$.

Even though the matrix D is an $m \times n$ rectangular matrix, since its only nonzero entries are on the descending diagonal, we still say that D is a diagonal matrix.

The `Matlab` command for computing an SVD $A = VDU^\top$ of a matrix A is `[V, D, U] = svd(A)`. Beware that `Matlab` uses the convention that the SVD of a matrix A is written as $A = UDV^\top$, and so the call for this version of the SVD is `[U, D, V] = svd(A)`.

If we view A as the representation of a linear map $f\colon E \to F$, where $\dim(E) = n$ and $\dim(F) = m$, the proof of Theorem 22.7 shows that there are two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ for E and F, respectively, where $(u_1, \ldots, u_n)$ are eigenvectors of $f^* \circ f$ and $(v_1, \ldots, v_m)$ are eigenvectors of $f \circ f^*$. Furthermore, $(u_1, \ldots, u_r)$ is an orthonormal basis of $\mathrm{Im}\, f^*$, $(u_{r+1}, \ldots, u_n)$ is an orthonormal basis of $\mathrm{Ker}\, f$, $(v_1, \ldots, v_r)$ is an orthonormal basis of $\mathrm{Im}\, f$, and $(v_{r+1}, \ldots, v_m)$ is an orthonormal basis of $\mathrm{Ker}\, f^*$.

The SVD of matrices can be used to define the pseudo-inverse of a rectangular matrix; we will do so in Chapter 23. The reader may also consult Strang [170], Demmel [48], Trefethen and Bau [176], and Golub and Van Loan [80].

One of the spectral theorems states that a symmetric matrix can be diagonalized by an orthogonal matrix. There are several numerical methods to compute the eigenvalues of a symmetric matrix A. One method consists in *tridiagonalizing* A, which means that there exists some orthogonal matrix P and some symmetric tridiagonal matrix T such that $A = PTP^\top$. In fact, this can be done using Householder transformations; see Theorem 18.2. It is then possible to compute the eigenvalues of T using a bisection method based on Sturm sequences. One can also use Jacobi's method. For details, see Golub and Van Loan [80], Chapter 8, Demmel [48], Trefethen and Bau [176], Lecture 26, Ciarlet [41], and Chapter 18. Computing the SVD of a matrix A is more involved. Most methods begin by finding orthogonal matrices U and V and a *bidiagonal* matrix B such that $A = VBU^\top$; see Problem 13.8 and Problem 22.3. This can also be done using Householder transformations. Observe that $B^\top B$ is symmetric tridiagonal. Thus, in principle, the previous method to diagonalize a symmetric tridiagonal matrix can be applied. However, it is unwise to compute $B^\top B$ explicitly, and more subtle methods are used for this last step; the matrix of Problem 22.1 can be used, and see Problem 22.3. Again, see Golub and Van Loan [80], Chapter 8, Demmel [48], and Trefethen and Bau [176], Lecture 31.

The polar form has applications in continuum mechanics. Indeed, in any deformation it is important to separate stretching from rotation. This is exactly what QS achieves. The orthogonal part Q corresponds to rotation (perhaps with an additional reflection), and the symmetric matrix S to stretching (or compression). The real eigenvalues $\sigma_1, \ldots, \sigma_r$ of S are the stretch factors (or compression factors) (see Marsden and Hughes [120]). The fact that S can be diagonalized by an orthogonal matrix corresponds to a natural choice of axes, the principal axes.

The SVD has applications to data compression, for instance in image processing. The idea is to retain only singular values whose magnitudes are significant enough. The SVD can also be used to determine the rank of a matrix when other methods such as Gaussian elimination produce very small pivots. One of the main applications of the SVD is the computation of the pseudo-inverse. Pseudo-inverses are the key to the solution of various optimization problems, in particular the method of least squares. This topic is discussed in the next chapter (Chapter 23). Applications of the material of this chapter can be found in Strang [170, 169]; Ciarlet [41]; Golub and Van Loan [80], which contains many other references; Demmel [48]; and Trefethen and Bau [176].

22.5 Ky Fan Norms and Schatten Norms

The singular values of a matrix can be used to define various norms on matrices which have found recent applications in quantum information theory and in spectral graph theory. Following Horn and Johnson [96] (Section 3.4) we can make the following definitions:

Definition 22.5. For any matrix $A \in \mathrm{M}_{m,n}(\mathbb{C})$, let $q = \min\{m, n\}$, and if $\sigma_1 \geq \cdots \geq \sigma_q$ are the singular values of A, for any k with $1 \leq k \leq q$, let

$$N_k(A) = \sigma_1 + \cdots + \sigma_k,$$

called the *Ky Fan k-norm* of A.

More generally, for any $p \geq 1$ and any k with $1 \leq k \leq q$, let

$$N_{k;p}(A) = (\sigma_1^p + \cdots + \sigma_k^p)^{1/p},$$

called the *Ky Fan p-k-norm* of A. When $k = q$, $N_{q;p}$ is also called the *Schatten p-norm*.

Observe that when $k = 1$, $N_1(A) = \sigma_1$, and the Ky Fan norm N_1 is simply the *spectral norm* from Chapter 9, which is the subordinate matrix norm associated with the Euclidean norm. When $k = q$, the Ky Fan norm N_q is given by

$$N_q(A) = \sigma_1 + \cdots + \sigma_q = \mathrm{tr}((A^*A)^{1/2})$$

and is called the *trace norm* or *nuclear norm*. When $p = 2$ and $k = q$, the Ky Fan $N_{q;2}$ norm is given by

$$N_{k;2}(A) = (\sigma_1^2 + \cdots + \sigma_q^2)^{1/2} = \sqrt{\mathrm{tr}(A^*A)} = \|A\|_F\,,$$

which is the *Frobenius norm* of A.

It can be shown that N_k and $N_{k;p}$ are unitarily invariant norms, and that when $m = n$, they are matrix norms; see Horn and Johnson [96] (Section 3.4, Corollary 3.4.4 and Problem 3).

22.6 Summary

The main concepts and results of this chapter are listed below:

- For any linear map $f\colon E \to E$ on a Euclidean space E, the maps $f^* \circ f$ and $f \circ f^*$ are self-adjoint and positive semidefinite.
- The *singular values* of a linear map.
- *Positive semidefinite* and *positive definite* self-adjoint maps.
- Relationships between $\mathrm{Im}\, f$, $\mathrm{Ker}\, f$, $\mathrm{Im}\, f^*$, and $\mathrm{Ker}\, f^*$.
- The *singular value decomposition theorem* for square matrices (Theorem 22.5).
- The *SVD* of matrix.
- The *polar decomposition* of a matrix.
- The *Weyl inequalities*.
- The *singular value decomposition theorem* for $m \times n$ matrices (Theorem 22.7).
- Ky Fan k-norms, Ky Fan p-k-norms, Schatten p-norms.

22.7 Problems

Problem 22.1. (1) Let A be a real $n\times n$ matrix and consider the $(2n)\times(2n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}.$$

Suppose that A has rank r. If $A = V\Sigma U^\top$ is an SVD for A, with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$, denoting the columns of U by u_k and the columns of V by v_k, prove that σ_k is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ u_k \end{pmatrix}$ for $k = 1, \ldots, n$, and that $-\sigma_k$ is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ -u_k \end{pmatrix}$ for $k = 1, \ldots, n$.

Hint. We have $Au_k = \sigma_k v_k$ for $k = 1, \ldots, n$. Show that $A^\top v_k = \sigma_k u_k$ for $k = 1, \ldots, n$.

(2) Prove that the $2n$ eigenvectors of S in (1) are pairwise orthogonal. Check that if A has rank r, then S has rank $2r$.

(3) Now assume that A is a real $m \times n$ matrix and consider the $(m + n) \times (m + n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}.$$

Suppose that A has rank r. If $A = V\Sigma U^\top$ is an SVD for A, prove that σ_k is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ u_k \end{pmatrix}$ for $k = 1, \ldots, r$, and that $-\sigma_k$ is an eigenvalue of S with corresponding eigenvector $\begin{pmatrix} v_k \\ -u_k \end{pmatrix}$ for $k = 1, \ldots, r$.

Find the remaining $m + n - 2r$ eigenvectors of S associated with the eigenvalue 0.

(4) Prove that these $m + n$ eigenvectors of S are pairwise orthogonal.

Problem 22.2. Let A be a real $m \times n$ matrix of rank r.

(1) Consider the $(m + n) \times (m + n)$ real symmetric matrix

$$S = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$$

and prove that

$$\begin{pmatrix} I_m & z^{-1}A \\ 0 & I_n \end{pmatrix} \begin{pmatrix} zI_m & -A \\ -A^\top & zI_n \end{pmatrix} = \begin{pmatrix} zI_m - z^{-1}AA^\top & 0 \\ -A^\top & zI_n \end{pmatrix}. \tag{$*$}$$

Use the Equation $(*)$ to prove that if if $n \geq m$, then

$$\det(zI_{m+n} - S) = z^{n-m} \det(z^2 I_m - AA^\top).$$

Permute the two matrices on the lefthand side of Equation $(*)$ to obtain another equation and use this equation to prove that if $m \geq n$, then

$$\det(zI_{m+n} - S) = z^{m-n} \det(z^2 I_n - A^\top A).$$

(2) Prove that the eigenvalues of S are $\pm\sigma_1, \ldots, \pm\sigma_r$, with $m + n - 2r$ additional zeros.

Problem 22.3. Let B be a real bidiagonal matrix of the form

$$B = \begin{pmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n-1} & b_{n-1} \\ 0 & 0 & \cdots & 0 & a_n \end{pmatrix}.$$

Let A be the $(2n) \times (2n)$ symmetric matrix

$$A = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix},$$

and let P be the permutation matrix given by $P = [e_1, e_{n+1}, e_2, e_{n+2}, \cdots, e_n, e_{2n}]$.

(1) Prove that $T = P^\top AP$ is a symmetric tridiagonal $(2n) \times (2n)$ matrix with zero main diagonal of the form

$$T = \begin{pmatrix} 0 & a_1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_1 & 0 & b_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & b_1 & 0 & a_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & a_2 & 0 & b_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1} & 0 & b_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & b_{n-1} & 0 & a_n \\ 0 & 0 & 0 & \cdots & 0 & 0 & a_n & 0 \end{pmatrix}.$$

(2) Prove that if x_i is a unit eigenvector for an eigenvalue λ_i of T, then $\lambda_i = \pm\sigma_i$ where σ_i is a singular value of B, and that

$$Px_i = \frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ \pm v_i \end{pmatrix},$$

where the u_i are unit eigenvectors of $B^\top B$ and the v_i are unit eigenvectors of $BB^\top$.

Problem 22.4. Find the SVD of the matrix

$$A = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}.$$

Problem 22.5. Let $u, v \in \mathbb{R}^n$ be two nonzero vectors, and let $A = uv^\top$ be the corresponding rank 1 matrix. Prove that the nonzero singular value of A is $\|u\|_2 \|v\|_2$.

Problem 22.6. Let A be a $n \times n$ real matrix. Prove that if $\sigma_1, \ldots, \sigma_n$ are the singular values of A, then $\sigma_1^3, \ldots, \sigma_n^3$ are the singular values of $AA^\top A$.

Problem 22.7. Let A be a real $n \times n$ matrix.

(1) Prove that the largest singular value σ_1 of A is given by

$$\sigma_1 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

and that this supremum is achieved at $x = u_1$, the first column in U in an SVD $A = V\Sigma U^\top$.

(2) Extend the above result to real $m \times n$ matrices.

Problem 22.8. Let A be a real $m \times n$ matrix. Prove that if B is any submatrix of A (by keeping $M \leq m$ rows and $N \leq n$ columns of A), then $(\sigma_1)_B \leq (\sigma_1)_A$ (where $(\sigma_1)_A$ is the largest singular value of A and similarly for $(\sigma_1)_B$).

Problem 22.9. Let A be a real $n \times n$ matrix.

(1) Assume A is invertible. Prove that if $A = Q_1 S_1 = Q_2 S_2$ are two polar decompositions of A, then $Q_1 = Q_2$ and $S_1 = S_2$.

Hint. $A^\top A = S_1^2 = S_2^2$, with S_1 and S_2 symmetric positive definite. Then use Problem 17.7.

(2) Now assume that A is singular. Prove that if $A = Q_1 S_1 = Q_2 S_2$ are two polar decompositions of A, then $S_1 = S_2$, but Q_1 may not be equal to Q_2.

Problem 22.10. (1) Let A be any invertible (real) $n \times n$ matrix. Prove that for every SVD, $A = VDU^\top$ of A, the product $VU^\top$ is the same (i.e., if $V_1 D U_1^\top = V_2 D U_2^\top$, then $V_1 U_1^\top = V_2 U_2^\top$). What does $VU^\top$ have to do with the polar form of A?

(2) Given any invertible (real) $n \times n$ matrix, A, prove that there is a unique orthogonal matrix, $Q \in \mathbf{O}(n)$, such that $\|A - Q\|_F$ is minimal (under the Frobenius norm). In fact, prove that $Q = VU^\top$, where $A = VDU^\top$ is an SVD of A. Moreover, if $\det(A) > 0$, show that $Q \in \mathbf{SO}(n)$.

What can you say if A is singular (i.e., non-invertible)?

Problem 22.11. (1) Prove that for any $n \times n$ matrix A and any orthogonal matrix Q, we have

$$\max\{\mathrm{tr}(QA) \mid Q \in \mathbf{O}(n)\} = \sigma_1 + \cdots + \sigma_n,$$

where $\sigma_1 \geq \cdots \geq \sigma_n$ are the singular values of A. Furthermore, this maximum is achieved by $Q = UV^\top$, where $A = V\Sigma U^\top$ is any SVD for A.

(2) By applying the above result with $A = X^\top Z$ and $Q = R$, deduce the following result: for any two fixed $n \times k$ matrices X and Z, the minimum of the set

$$\{\|X - ZR\|_F \mid R \in \mathbf{O}(k)\}$$

is achieved by $R = UV^\top$ for any SVD decomposition $V\Sigma U^\top = X^\top Z$ of $X^\top Z$.

Remark: The problem of finding an orthogonal matrix R such that ZR comes as close as possible to X is called the *orthogonal Procrustes problem*; see Strang [171] (Section IV.9) for the history of this problem.

Chapter 23

Applications of SVD and Pseudo-Inverses

> De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile, que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre *minimum* la somme des carrés des erreurs. Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui, empêchant les extrêmes de prévaloir, est très propre às faire connaitre l'état du système le plus proche de la vérité.
> —**Legendre, 1805,** *Nouvelles Méthodes pour la détermination des Orbites des Comètes*

23.1 Least Squares Problems and the Pseudo-Inverse

This chapter presents several applications of SVD. The first one is the pseudo-inverse, which plays a crucial role in solving linear systems by the method of least squares. The second application is data compression. The third application is principal component analysis (PCA), whose purpose is to identify patterns in data and understand the variance–covariance structure of the data. The fourth application is the best affine approximation of a set of data, a problem closely related to PCA.

The method of least squares is a way of "solving" an overdetermined system of linear equations

$$Ax = b,$$

i.e., a system in which A is a rectangular $m \times n$ matrix with more equations than unknowns (when $m > n$). Historically, the method of least squares was used by Gauss and Legendre to solve problems in astronomy and geodesy. The method was first published by Legendre in 1805 in a paper on methods for determining the orbits of comets. However, Gauss had already used the method of least squares as early as 1801 to determine the orbit of the asteroid

Ceres, and he published a paper about it in 1810 after the discovery of the asteroid Pallas. Incidentally, it is in that same paper that Gaussian elimination using pivots is introduced.

The reason why more equations than unknowns arise in such problems is that repeated measurements are taken to minimize errors. This produces an overdetermined and often inconsistent system of linear equations. For example, Gauss solved a system of eleven equations in six unknowns to determine the orbit of the asteroid Pallas.

Example 23.1. As a concrete illustration, suppose that we observe the motion of a small object, assimilated to a point, in the plane. From our observations, we suspect that this point moves along a straight line, say of equation $y = cx + d$. Suppose that we observed the moving point at three different locations (x_1, y_1), (x_2, y_2), and (x_3, y_3). Then we should have

$$\begin{aligned} d + cx_1 &= y_1, \\ d + cx_2 &= y_2, \\ d + cx_3 &= y_3. \end{aligned}$$

If there were no errors in our measurements, these equations would be compatible, and c and d would be determined by only two of the equations. However, in the presence of errors, the system may be inconsistent. Yet we would like to find c and d!

The idea of the method of least squares is to determine (c, d) such that it minimizes the sum of the squares of the errors, namely,

$$(d + cx_1 - y_1)^2 + (d + cx_2 - y_2)^2 + (d + cx_3 - y_3)^2.$$

See Figure 23.1.

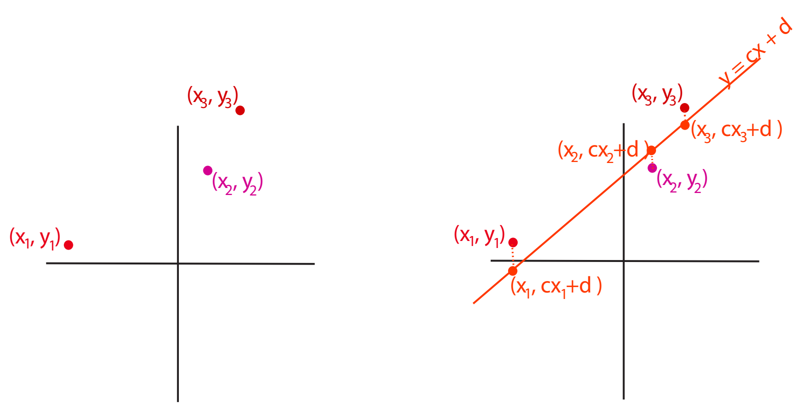


Figure 23.1: Given three points (x_1, y_1), (x_2, y_2), (x_3, y_3), we want to determine the line $y = cx + d$ which minimizes the lengths of the dashed vertical lines.

In general, for an overdetermined $m \times n$ system $Ax = b$, what Gauss and Legendre discovered is that there are solutions x minimizing

$$\|Ax - b\|_2^2$$

(where $\|u\|_2^2 = u_1^2 + \cdots + u_n^2$, the square of the Euclidean norm of the vector $u = (u_1, \ldots, u_n)$), and that these solutions are given by the square $n \times n$ system

$$A^\top Ax = A^\top b,$$

called the *normal equations*. Furthermore, when the columns of A are linearly independent, it turns out that $A^\top A$ is invertible, and so x is unique and given by

$$x = (A^\top A)^{-1} A^\top b.$$

Note that $A^\top A$ is a symmetric matrix, one of the nice features of the normal equations of a least squares problem. For instance, since the above problem in matrix form is represented as

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

the normal equations are

$$\begin{pmatrix} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + y_3 \\ x_1 y_1 + x_2 y_2 + x_3 y_3 \end{pmatrix}.$$

In fact, given any real $m \times n$ matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|_2^2$, even when the columns of A are linearly dependent. How do we prove this, and how do we find x^+?

Theorem 23.1. *Every linear system $Ax = b$, where A is an $m \times n$ matrix, has a unique least squares solution x^+ of smallest norm.*

Proof. Geometry offers a nice proof of the existence and uniqueness of x^+. Indeed, we can interpret b as a point in the Euclidean (affine) space $\mathbb{R}^m$, and the image subspace of A (also called the column space of A) as a subspace U of $\mathbb{R}^m$ (passing through the origin). Then it is clear that

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \inf_{y \in U} \|y - b\|_2^2,$$

with $U = \operatorname{Im} A$, and we claim that x minimizes $\|Ax - b\|_2^2$ iff $Ax = p$, where p the orthogonal projection of b onto the subspace U.

Recall from Section 13.1 that the orthogonal projection $p_U \colon U \oplus U^\perp \to U$ is the linear map given by

$$p_U(u + v) = u,$$

with $u \in U$ and $v \in U^{\perp}$. If we let $p = p_U(b) \in U$, then for any point $y \in U$, the vectors $\overrightarrow{py} = y - p \in U$ and $\overrightarrow{bp} = p - b \in U^{\perp}$ are orthogonal, which implies that

$$\|\overrightarrow{by}\|_2^2 = \|\overrightarrow{bp}\|_2^2 + \|\overrightarrow{py}\|_2^2,$$

where $\overrightarrow{by} = y - b$. Thus, p is indeed the unique point in U that minimizes the distance from b to any point in U. See Figure 23.2.

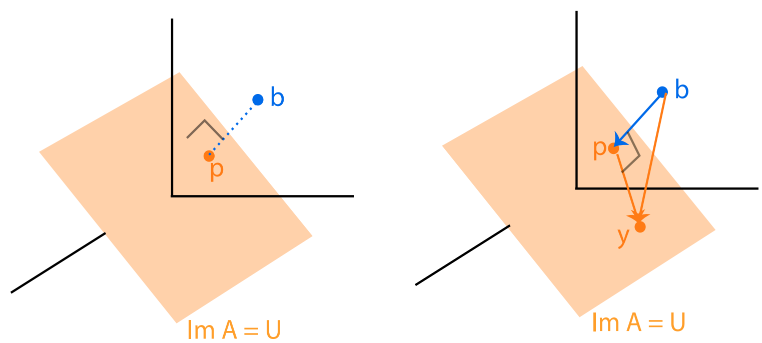


Figure 23.2: Given a 3×2 matrix A, $U = \operatorname{Im} A$ is the peach plane in $\mathbb{R}^3$ and p is the orthogonal projection of b onto U. Furthermore, given $y \in U$, the points b, y, and p are the vertices of a right triangle.

Thus the problem has been reduced to proving that there is a unique x^+ of minimum norm such that $Ax^+ = p$, with $p = p_U(b) \in U$, the orthogonal projection of b onto U. We use the fact that

$$\mathbb{R}^n = \operatorname{Ker} A \oplus (\operatorname{Ker} A)^{\perp}.$$

Consequently, every $x \in \mathbb{R}^n$ can be written uniquely as $x = u + v$, where $u \in \operatorname{Ker} A$ and $v \in (\operatorname{Ker} A)^{\perp}$, and since u and v are orthogonal,

$$\|x\|_2^2 = \|u\|_2^2 + \|v\|_2^2.$$

Furthermore, since $u \in \operatorname{Ker} A$, we have $Au = 0$, and thus $Ax = p$ iff $Av = p$, which shows that the solutions of $Ax = p$ for which x has minimum norm must belong to $(\operatorname{Ker} A)^{\perp}$. However, the restriction of A to $(\operatorname{Ker} A)^{\perp}$ is injective. This is because if $Av_1 = Av_2$, where $v_1, v_2 \in (\operatorname{Ker} A)^{\perp}$, then $A(v_2 - v_1) = 0$, which implies $v_2 - v_1 \in \operatorname{Ker} A$, and since $v_1, v_2 \in (\operatorname{Ker} A)^{\perp}$, we also have $v_2 - v_1 \in (\operatorname{Ker} A)^{\perp}$, and consequently, $v_2 - v_1 = 0$. This shows that there is a unique x^+ of minimum norm such that $Ax^+ = p$, and that x^+ must belong to $(\operatorname{Ker} A)^{\perp}$. By our previous reasoning, x^+ is the unique vector of minimum norm minimizing $\|Ax - b\|_2^2$. $\square$

The proof also shows that x minimizes $\|Ax - b\|_2^2$ iff $\overrightarrow{pb} = b - Ax$ is orthogonal to U, which can be expressed by saying that $b - Ax$ is orthogonal to every column of A. However, this is equivalent to

$$A^\top(b - Ax) = 0, \quad \text{i.e.,} \quad A^\top Ax = A^\top b.$$

Finally, it turns out that the minimum norm least squares solution x^+ can be found in terms of the pseudo-inverse A^+ of A, which is itself obtained from any SVD of A.

Definition 23.1. Given any nonzero $m \times n$ matrix A of rank r, if $A = VDU^\top$ is an SVD of A such that

$$D = \begin{pmatrix} \Lambda & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

with

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$$

an $r \times r$ diagonal matrix consisting of the nonzero singular values of A, then if we let D^+ be the $n \times m$ matrix

$$D^+ = \begin{pmatrix} \Lambda^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{pmatrix},$$

with

$$\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_r),$$

the *pseudo-inverse* of A is defined by

$$A^+ = UD^+V^\top.$$

If $A = 0_{m,n}$ is the zero matrix, we set $A^+ = 0_{n,m}$. Observe that D^+ is obtained from D by inverting the nonzero diagonal entries of D, leaving all zeros in place, and then transposing the matrix. For example, given the matrix

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

its pseudo-inverse is

$$D^+ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pseudo-inverse of a matrix is also known as the *Moore–Penrose pseudo-inverse*.

Actually, it seems that A^+ depends on the specific choice of U and V in an SVD (U, D, V) for A, but the next theorem shows that this is not so.

Theorem 23.2. *The least squares solution of smallest norm of the linear system $Ax = b$, where A is an $m \times n$ matrix, is given by*

$$x^+ = A^+ b = UD^+ V^\top b.$$

Proof. First assume that A is a (rectangular) diagonal matrix D, as above. Then since x minimizes $\|Dx - b\|_2^2$ iff Dx is the projection of b onto the image subspace F of D, it is fairly obvious that $x^+ = D^+ b$. Otherwise, we can write

$$A = VDU^\top,$$

where U and V are orthogonal. However, since V is an isometry,

$$\|Ax - b\|_2 = \|VDU^\top x - b\|_2 = \|DU^\top x - V^\top b\|_2.$$

Letting $y = U^\top x$, we have $\|x\|_2 = \|y\|_2$, since U is an isometry, and since U is surjective, $\|Ax - b\|_2$ is minimized iff $\|Dy - V^\top b\|_2$ is minimized, and we have shown that the least solution is

$$y^+ = D^+ V^\top b.$$

Since $y = U^\top x$, with $\|x\|_2 = \|y\|_2$, we get

$$x^+ = UD^+ V^\top b = A^+ b.$$

Thus, the pseudo-inverse provides the optimal solution to the least squares problem. $\square$

By Theorem 23.2 and Theorem 23.1, $A^+ b$ is uniquely defined by every b, and thus A^+ depends only on A.

The `Matlab` command for computing the pseudo-inverse B of the matrix A is
`B = pinv(A)`.

Example 23.2. If A is the rank 2 matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$$

whose eigenvalues are $-1.1652, 0, 0, 17.1652$, using `Matlab` we obtain the SVD $A = VDU^\top$ with

$$U = \begin{pmatrix} -0.3147 & 0.7752 & 0.2630 & -0.4805 \\ -0.4275 & 0.3424 & 0.0075 & 0.8366 \\ -0.5402 & -0.0903 & -0.8039 & -0.2319 \\ -0.6530 & -0.5231 & 0.5334 & -0.1243 \end{pmatrix},$$

$$V = \begin{pmatrix} -0.3147 & -0.7752 & 0.5452 & 0.0520 \\ -0.4275 & -0.3424 & -0.7658 & 0.3371 \\ -0.5402 & 0.0903 & -0.1042 & -0.8301 \\ -0.6530 & 0.5231 & 0.3247 & 0.4411 \end{pmatrix}, \quad D = \begin{pmatrix} 17.1652 & 0 & 0 & 0 \\ 0 & 1.1652 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$D^+ = \begin{pmatrix} 0.0583 & 0 & 0 & 0 \\ 0 & 0.8583 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$A^+ = UD^+V^\top = \begin{pmatrix} -0.5100 & -0.2200 & 0.0700 & 0.3600 \\ -0.2200 & -0.0900 & 0.0400 & 0.1700 \\ 0.0700 & 0.0400 & 0.0100 & -0.0200 \\ 0.3600 & 0.1700 & -0.0200 & -0.2100 \end{pmatrix},$$

which is also the result obtained by calling `pinv(A)`.

If A is an $m \times n$ matrix of rank n (and so $m \geq n$), it is immediately shown that the QR-decomposition in terms of Householder transformations applies as follows:

There are n $m \times m$ matrices $H_1, \ldots, H_n$, Householder matrices or the identity, and an upper triangular $m \times n$ matrix R of rank n such that

$$A = H_1 \cdots H_n R.$$

Then because each H_i is an isometry,

$$\|Ax - b\|_2 = \|Rx - H_n \cdots H_1 b\|_2,$$

and the least squares problem $Ax = b$ is equivalent to the system

$$Rx = H_n \cdots H_1 b.$$

Now the system

$$Rx = H_n \cdots H_1 b$$

is of the form

$$\begin{pmatrix} R_1 \\ 0_{m-n} \end{pmatrix} x = \begin{pmatrix} c \\ d \end{pmatrix},$$

where R_1 is an invertible $n \times n$ matrix (since A has rank n), $c \in \mathbb{R}^n$, and $d \in \mathbb{R}^{m-n}$, and the least squares solution of smallest norm is

$$x^+ = R_1^{-1} c.$$

Since R_1 is a triangular matrix, it is very easy to invert R_1.

The method of least squares is one of the most effective tools of the mathematical sciences. There are entire books devoted to it. Readers are advised to consult Strang [170], Golub and Van Loan [80], Demmel [48], and Trefethen and Bau [176], where extensions and applications of least squares (such as weighted least squares and recursive least squares) are described. Golub and Van Loan [80] also contains a very extensive bibliography, including a list of books on least squares.

23.2 Properties of the Pseudo-Inverse

We begin this section with a proposition which provides a way to calculate the pseudo-inverse of an $m \times n$ matrix A without first determining an SVD factorization.

Proposition 23.3. *When A has full rank, the pseudo-inverse A^+ can be expressed as $A^+ = (A^\top A)^{-1} A^\top$ when $m \geq n$, and as $A^+ = A^\top (AA^\top)^{-1}$ when $n \geq m$. In the first case ($m \geq n$), observe that $A^+ A = I$, so A^+ is a left inverse of A; in the second case ($n \geq m$), we have $AA^+ = I$, so A^+ is a right inverse of A.*

Proof. If $m \geq n$ and A has full rank n, we have

$$A = V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top$$

with Λ an $n \times n$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top.$$

We find that

$$A^\top A = U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top V \begin{pmatrix} \Lambda \\ 0_{m-n,n} \end{pmatrix} U^\top = U \Lambda^2 U^\top,$$

which yields

$$(A^\top A)^{-1} A^\top = U \Lambda^{-2} U^\top U \begin{pmatrix} \Lambda & 0_{n,m-n} \end{pmatrix} V^\top = U \begin{pmatrix} \Lambda^{-1} & 0_{n,m-n} \end{pmatrix} V^\top = A^+.$$

Therefore, if $m \geq n$ and A has full rank n, then

$$A^+ = (A^\top A)^{-1} A^\top.$$

If $n \geq m$ and A has full rank m, then

$$A = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top$$

with Λ an $m \times m$ diagonal invertible matrix (with positive entries), so

$$A^+ = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top.$$

We find that

$$AA^\top = V \begin{pmatrix} \Lambda & 0_{m,n-m} \end{pmatrix} U^\top U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top = V \Lambda^2 V^\top,$$

which yields

$$A^\top (AA^\top)^{-1} = U \begin{pmatrix} \Lambda \\ 0_{n-m,m} \end{pmatrix} V^\top V \Lambda^{-2} V^\top = U \begin{pmatrix} \Lambda^{-1} \\ 0_{n-m,m} \end{pmatrix} V^\top = A^+.$$

Therefore, if $n \geq m$ and A has full rank m, then $A^+ = A^\top (AA^\top)^{-1}$. $\square$

For example, if $A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 0 & 1 \end{pmatrix}$, then A has rank 2 and since $m \geq n$, $A^+ = (A^\top A)^{-1} A^\top$ where

$$A^+ = \begin{pmatrix} 5 & 8 \\ 8 & 14 \end{pmatrix}^{-1} A^\top = \begin{pmatrix} 7/3 & -4/3 \\ 4/3 & 5/6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} -1/3 & 2/3 & -4/3 \\ 1/3 & -1/6 & 5/6 \end{pmatrix}.$$

If $A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix}$, since A has rank 2 and $n \geq m$, then $A^+ = A^\top (AA^\top)^{-1}$ where

$$A^+ = A^\top \begin{pmatrix} 14 & 5 \\ 5 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 3/17 & -5/17 \\ -5/17 & 14/17 \end{pmatrix} = \begin{pmatrix} 3/17 & -5/17 \\ 1/17 & 4/17 \\ 4/17 & -1/17 \\ 5/17 & -14/17 \end{pmatrix}.$$

Let $A = V\Sigma U^\top$ be an SVD for any $m \times n$ matrix A. It is easy to check that both AA^+ and A^+A are symmetric matrices. In fact,

$$AA^+ = V\Sigma U^\top U \Sigma^+ V^\top = V\Sigma\Sigma^+ V^\top = V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top$$

and

$$A^+A = U\Sigma^+ V^\top V \Sigma U^\top = U\Sigma^+\Sigma U^\top = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top.$$

From the above expressions we immediately deduce that

$$\begin{aligned} AA^+A &= A, \\ A^+AA^+ &= A^+, \end{aligned}$$

and that

$$\begin{aligned} (AA^+)^2 &= AA^+, \\ (A^+A)^2 &= A^+A, \end{aligned}$$

so both AA^+ and A^+A are orthogonal projections (since they are both symmetric).

Proposition 23.4. *The matrix AA^+ is the orthogonal projection onto the range of A and A^+A is the orthogonal projection onto* $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^\top)$*, the range of $A^\top$.*

Proof. Obviously, we have $\mathrm{range}(AA^+) \subseteq \mathrm{range}(A)$, and for any $y = Ax \in \mathrm{range}(A)$, since $AA^+A = A$, we have

$$AA^+y = AA^+Ax = Ax = y,$$

so the image of AA^+ is indeed the range of A. It is also clear that $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(A^+A)$, and since $AA^+A = A$, we also have $\mathrm{Ker}(A^+A) \subseteq \mathrm{Ker}(A)$, and so

$$\mathrm{Ker}(A^+A) = \mathrm{Ker}(A).$$

Since A^+A is symmetric, $\mathrm{range}(A^+A) = \mathrm{range}((A^+A)^\top) = \mathrm{Ker}(A^+A)^\perp = \mathrm{Ker}(A)^\perp$, as claimed. $\square$

Proposition 23.5. *The set* $\mathrm{range}(A) = \mathrm{range}(AA^+)$ *consists of all vectors* $y \in \mathbb{R}^m$ *such that*

$$V^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. Indeed, if $y = Ax$, then

$$V^\top y = V^\top Ax = V^\top V\Sigma U^\top x = \Sigma U^\top x = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} U^\top x = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

where Σ_r is the $r \times r$ diagonal matrix $\mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Conversely, if $V^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = V\left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and

$$\begin{aligned}
AA^+y &= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top y \\
&= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} V^\top V \begin{pmatrix} z \\ 0 \end{pmatrix} \\
&= V \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\
&= V \begin{pmatrix} z \\ 0 \end{pmatrix} = y,
\end{aligned}$$

which shows that y belongs to the range of A. $\square$

Similarly, we have the following result.

Proposition 23.6. *The set* $\mathrm{range}(A^+A) = \mathrm{Ker}(A)^\perp$ *consists of all vectors* $y \in \mathbb{R}^n$ *such that*

$$U^\top y = \begin{pmatrix} z \\ 0 \end{pmatrix},$$

with $z \in \mathbb{R}^r$.

Proof. If $y = A^+Au$, then

$$y = A^+Au = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top u = U \begin{pmatrix} z \\ 0 \end{pmatrix},$$

for some $z \in \mathbb{R}^r$. Conversely, if $U^\top y = \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, then $y = U \left(\begin{smallmatrix} z \\ 0 \end{smallmatrix}\right)$, and so

$$\begin{aligned} A^+AU \begin{pmatrix} z \\ 0 \end{pmatrix} &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} U^\top U \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix} \begin{pmatrix} z \\ 0 \end{pmatrix} \\ &= U \begin{pmatrix} z \\ 0 \end{pmatrix} = y, \end{aligned}$$

which shows that $y \in \mathrm{range}(A^+A)$. ∎

Analogous results hold for complex matrices, but in this case, V and U are unitary matrices and AA^+ and A^+A are Hermitian orthogonal projections.

If A is a normal matrix, which means that $AA^\top = A^\top A$, then there is an intimate relationship between SVD's of A and block diagonalizations of A. As a consequence, the pseudo-inverse of a normal matrix A can be obtained directly from a block diagonalization of A.

If A is a (real) normal matrix, then we know from Theorem 17.18 that A can be block diagonalized with respect to an orthogonal matrix U as

$$A = U\Lambda U^\top,$$

where Λ is the (real) block diagonal matrix

$$\Lambda = \mathrm{diag}(B_1, \ldots, B_n),$$

consisting either of 2×2 blocks of the form

$$B_j = \begin{pmatrix} \lambda_j & -\mu_j \\ \mu_j & \lambda_j \end{pmatrix}$$

with $\mu_j \neq 0$, or of one-dimensional blocks $B_k = (\lambda_k)$. Then we have the following proposition:

Proposition 23.7. *For any (real) normal matrix A and any block diagonalization $A = U\Lambda U^\top$ of A as above, the pseudo-inverse of A is given by*

$$A^+ = U\Lambda^+U^\top,$$

where Λ^+ is the pseudo-inverse of Λ. Furthermore, if

$$\Lambda = \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_r has rank r, then

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. Assume that $B_1, \ldots, B_p$ are 2×2 blocks and that $\lambda_{2p+1}, \ldots, \lambda_n$ are the scalar entries. We know that the numbers $\lambda_j \pm i\mu_j$, and the λ_{2p+k} are the eigenvalues of A. Let $\rho_{2j-1} = \rho_{2j} = \sqrt{\lambda_j^2 + \mu_j^2} = \sqrt{\det(B_i)}$ for $j = 1, \ldots, p$, $\rho_j = |\lambda_j|$ for $j = 2p+1, \ldots, r$. Multiplying U by a suitable permutation matrix, we may assume that the blocks of Λ are ordered so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$. Then it is easy to see that

$$AA^\top = A^\top A = U\Lambda U^\top U\Lambda^\top U^\top = U\Lambda\Lambda^\top U^\top,$$

with

$$\Lambda\Lambda^\top = \mathrm{diag}(\rho_1^2, \ldots, \rho_r^2, 0, \ldots, 0),$$

so $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0$ are the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ of A. Define the diagonal matrix

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0),$$

where $r = \mathrm{rank}(A)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and the block diagonal matrix Θ defined such that the block B_i in Λ is replaced by the block $\sigma^{-1}B_i$ where $\sigma = \sqrt{\det(B_i)}$, the nonzero scalar λ_j is replaced $\lambda_j/|\lambda_j|$, and a diagonal zero is replaced by 1. Observe that Θ is an orthogonal matrix and

$$\Lambda = \Theta\Sigma.$$

But then we can write

$$A = U\Lambda U^\top = U\Theta\Sigma U^\top,$$

and we if let $V = U\Theta$, since U is orthogonal and Θ is also orthogonal, V is also orthogonal and $A = V\Sigma U^\top$ *is an SVD for A*. Now we get

$$A^+ = U\Sigma^+ V^\top = U\Sigma^+\Theta^\top U^\top.$$

However, since Θ is an orthogonal matrix, $\Theta^\top = \Theta^{-1}$, and a simple calculation shows that

$$\Sigma^+\Theta^\top = \Sigma^+\Theta^{-1} = \Lambda^+,$$

which yields the formula

$$A^+ = U\Lambda^+ U^\top.$$

Also observe that Λ_r is invertible and

$$\Lambda^+ = \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, the pseudo-inverse of a normal matrix can be computed directly from any block diagonalization of A, as claimed. $\square$

Example 23.3. Consider the following real diagonal form of the normal matrix

$$A = \begin{pmatrix} -2.7500 & 2.1651 & -0.8660 & 0.5000 \\ 2.1651 & -0.2500 & -1.5000 & 0.8660 \\ 0.8660 & 1.5000 & 0.7500 & -0.4330 \\ -0.5000 & -0.8660 & -0.4330 & 0.2500 \end{pmatrix} = U\Lambda U^\top,$$

with

$$U = \begin{pmatrix} \cos(\pi/3) & 0 & \sin(\pi/3) & 0 \\ \sin(\pi/3) & 0 & -\cos(\pi/3) & 0 \\ 0 & \cos(\pi/6) & 0 & \sin(\pi/6) \\ 0 & -\cos(\pi/6) & 0 & \sin(\pi/6) \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We obtain

$$\Lambda^+ = \begin{pmatrix} 1/5 & 2/5 & 0 & 0 \\ -2/5 & 1/5 & 0 & 0 \\ 0 & 0 & -1/4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the pseudo-inverse of A is

$$A^+ = U\Lambda^+ U^\top = \begin{pmatrix} -0.1375 & 0.1949 & 0.1732 & -0.1000 \\ 0.1949 & 0.0875 & 0.3000 & -0.1732 \\ -0.1732 & -0.3000 & 0.1500 & -0.0866 \\ 0.1000 & 0.1732 & -0.0866 & 0.0500 \end{pmatrix},$$

which agrees with `pinv(A)`.

The following properties, due to Penrose, characterize the pseudo-inverse of a matrix. We have already proved that the pseudo-inverse satisfies these equations. For a proof of the converse, see Kincaid and Cheney [102].

Proposition 23.8. *Given any $m \times n$ matrix A (real or complex), the pseudo-inverse A^+ of A is the unique $n \times m$ matrix satisfying the following properties:*

$$\begin{aligned} AA^+A &= A, \\ A^+AA^+ &= A^+, \\ (AA^+)^\top &= AA^+, \\ (A^+A)^\top &= A^+A. \end{aligned}$$

23.3 Data Compression and SVD

Among the many applications of SVD, a very useful one is *data compression*, notably for images. In order to make precise the notion of closeness of matrices, we use the notion of *matrix norm*. This concept is defined in Chapter 9, and the reader may want to review it before reading any further.

Given an $m \times n$ matrix of rank r, we would like to find a best approximation of A by a matrix B of rank $k \leq r$ (actually, $k < r$) such that $\|A - B\|_2$ (or $\|A - B\|_F$) is minimized. The following proposition is known as the *Eckart–Young theorem*.

Proposition 23.9. *Let A be an $m \times n$ matrix of rank r and let $VDU^\top = A$ be an SVD for A. Write u_i for the columns of U, v_i for the columns of V, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ for the singular values of A ($p = \min(m, n)$). Then a matrix of rank $k < r$ closest to A (in the $\| \, \|_2$ norm) is given by*

$$A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^\top = V \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) U^\top$$

and $\|A - A_k\|_2 = \sigma_{k+1}$.

Proof. By construction, A_k has rank k, and we have

$$\|A - A_k\|_2 = \left\| \sum_{i=k+1}^{p} \sigma_i v_i u_i^\top \right\|_2 = \left\| V \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p) U^\top \right\|_2 = \sigma_{k+1}.$$

It remains to show that $\|A - B\|_2 \geq \sigma_{k+1}$ for all rank k matrices B. Let B be any rank k matrix, so its kernel has dimension $n - k$. The subspace U_{k+1} spanned by $(u_1, \ldots, u_{k+1})$ has dimension $k + 1$, and because the sum of the dimensions of the kernel of B and of U_{k+1} is $(n - k) + k + 1 = n + 1$, these two subspaces must intersect in a subspace of dimension at least 1. Pick any unit vector h in $\mathrm{Ker}(B) \cap U_{k+1}$. Then since $Bh = 0$, and since U and V are isometries, we have

$$\|A - B\|_2^2 \geq \|(A - B)h\|_2^2 = \|Ah\|_2^2 = \left\| VDU^\top h \right\|_2^2 = \left\| DU^\top h \right\|_2^2 \geq \sigma_{k+1}^2 \left\| U^\top h \right\|_2^2 = \sigma_{k+1}^2,$$

which proves our claim. ∎

Note that A_k can be stored using $(m + n)k$ entries, as opposed to mn entries. When $k \ll m$, this is a substantial gain.

Example 23.4. Consider the badly conditioned symmetric matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

from Section 9.5. Since A is SPD, we have the SVD

$$A = UDU^\top,$$

with

$$U = \begin{pmatrix} -0.5286 & -0.6149 & 0.3017 & -0.5016 \\ -0.3803 & -0.3963 & -0.0933 & 0.8304 \\ -0.5520 & 0.2716 & -0.7603 & -0.2086 \\ -0.5209 & 0.6254 & 0.5676 & 0.1237 \end{pmatrix}, D = \begin{pmatrix} 30.2887 & 0 & 0 & 0 \\ 0 & 3.8581 & 0 & 0 \\ 0 & 0 & 0.8431 & 0 \\ 0 & 0 & 0 & 0.0102 \end{pmatrix}.$$

If we set $\sigma_3 = \sigma_4 = 0$, we obtain the best rank 2 approximation

$$A_2 = U(:, 1:2) * D(:, 1:2) * U(:, 1:2)' = \begin{pmatrix} 9.9207 & 7.0280 & 8.1923 & 6.8563 \\ 7.0280 & 4.9857 & 5.9419 & 5.0436 \\ 8.1923 & 5.9419 & 9.5122 & 9.3641 \\ 6.8563 & 5.0436 & 9.3641 & 9.7282 \end{pmatrix}.$$

A nice example of the use of Proposition 23.9 in image compression is given in Demmel [48], Chapter 3, Section 3.2.3, pages 113–115; see the Matlab demo.

Proposition 23.9 also holds for the Frobenius norm; see Problem 23.4.

An interesting topic that we have not addressed is the actual computation of an SVD. This is a very interesting but tricky subject. Most methods reduce the computation of an SVD to the diagonalization of a well-chosen symmetric matrix which is not $A^\top A$; see Problem 22.1 and Problem 22.3. Interested readers should read Section 5.4 of Demmel's excellent book [48], which contains an overview of most known methods and an extensive list of references.

23.4 Principal Components Analysis (PCA)

Suppose we have a set of data consisting of n points $X_1, \ldots, X_n$, with each $X_i \in \mathbb{R}^d$ *viewed as a row vector*. Think of the X_i's as persons, and if $X_i = (x_{i\,1}, \ldots, x_{i\,d})$, each $x_{i\,j}$ is the value of some *feature* (or *attribute*) of that person.

Example 23.5. For example, the X_i's could be mathematicians, $d = 2$, and the first component, $x_{i\,1}$, of X_i could be the year that X_i was born, and the second component, $x_{i\,2}$, the length of the beard of X_i in centimeters. Here is a small data set.

Name	year	length
Carl Friedrich Gauss	1777	0
Camille Jordan	1838	12
Adrien-Marie Legendre	1752	0
Bernhard Riemann	1826	15
David Hilbert	1862	2
Henri Poincaré	1854	5
Emmy Noether	1882	0
Karl Weierstrass	1815	0
Eugenio Beltrami	1835	2
Hermann Schwarz	1843	20

We usually form the $n \times d$ matrix X whose ith row is X_i, with $1 \leq i \leq n$. Then the jth column is denoted by C_j ($1 \leq j \leq d$). It is sometimes called a *feature vector*, but this terminology is far from being universally accepted. In fact, many people in computer vision call the data points X_i feature vectors!

The purpose of *principal components analysis*, for short *PCA*, is to identify patterns in data and understand the *variance–covariance* structure of the data. This is useful for the following tasks:

1. Data reduction: Often much of the variability of the data can be accounted for by a smaller number of *principal components*.
2. Interpretation: PCA can show relationships that were not previously suspected.

Given a vector (a *sample* of measurements) $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, recall that the *mean* (or *average*) $\overline{x}$ of x is given by

$$\overline{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

We let $x - \overline{x}$ denote the *centered data point*

$$x - \overline{x} = (x_1 - \overline{x}, \ldots, x_n - \overline{x}).$$

In order to *measure the spread* of the x_i's around the mean, we define the *sample variance* (for short, *variance*) $\mathrm{var}(x)$ (or s^2) of the sample x by

$$\mathrm{var}(x) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}.$$

Example 23.6. If $x = (1, 3, -1)$, $\overline{x} = \frac{1+3-1}{3} = 1$, $x - \overline{x} = (0, 2, -2)$, and $\mathrm{var}(x) = \frac{0^2+2^2+(-2)^2}{2} = 4$. If $y = (1, 2, 3)$, $\overline{y} = \frac{1+2+3}{3} = 2$, $y - \overline{y} = (-1, 0, 1)$, and $\mathrm{var}(y) = \frac{(-1)^2+0^2+1^2}{2} = 2$.

There is a reason for using $n-1$ instead of n. The above definition makes $\mathrm{var}(x)$ an unbiased estimator of the variance of the random variable being sampled. However, we don't need to worry about this. Curious readers will find an explanation of these peculiar definitions in Epstein [57] (Chapter 14, Section 14.5) or in any decent statistics book.

Given two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the *sample covariance* (for short, *covariance*) of x and y is given by

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Example 23.7. If we take $x = (1, 3, -1)$ and $y = (0, 2, -2)$, we know from Example 23.6 that $x - \overline{x} = (0, 2, -2)$ and $y - \overline{y} = (-1, 0, 1)$. Thus, $\mathrm{cov}(x, y) = \frac{0(-1)+2(0)+(-2)(1)}{2} = -1$.

The covariance of x and y measures how x and y vary from the mean with respect to each other. Obviously, $\mathrm{cov}(x, y) = \mathrm{cov}(y, x)$ and $\mathrm{cov}(x, x) = \mathrm{var}(x)$.

Note that

$$\mathrm{cov}(x, y) = \frac{(x - \overline{x})^\top (y - \overline{y})}{n-1}.$$

We say that x and y are *uncorrelated* iff $\mathrm{cov}(x, y) = 0$.

Finally, given an $n \times d$ matrix X of n points X_i, for PCA to be meaningful, it will be necessary to translate the origin to the *centroid* (or *center of gravity*) μ of the X_i's, defined by

$$\mu = \frac{1}{n}(X_1 + \cdots + X_n).$$

Observe that if $\mu = (\mu_1, \ldots, \mu_d)$, then μ_j is the mean of the vector C_j (the jth column of X).

We let $X - \mu$ denote the *matrix* whose ith row is the centered data point $X_i - \mu$ ($1 \leq i \leq n$). Then the *sample covariance matrix* (for short, *covariance matrix*) of X is the $d \times d$ symmetric matrix

$$\Sigma = \frac{1}{n-1}(X - \mu)^\top (X - \mu) = (\mathrm{cov}(C_i, C_j)).$$

Example 23.8. Let $X = \begin{pmatrix} 1 & 1 \\ 3 & 2 \\ -1 & 3 \end{pmatrix}$, the 3×2 matrix whose columns are the vector x and y of Example 23.6. Then

$$\mu = \frac{1}{3}\left[(1, 1) + (3, 2) + (-1, 3)\right] = (1, 2),$$

$$X - \mu = \begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix},$$

and

$$\Sigma = \frac{1}{2} \begin{pmatrix} 0 & 2 & -2 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 2 & 0 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}.$$

Remark: The factor $\frac{1}{n-1}$ is irrelevant for our purposes and can be ignored.

Example 23.9. Here is the matrix $X - \mu$ in the case of our bearded mathematicians: since

$$\mu_1 = 1828.4, \quad \mu_2 = 5.6,$$

we get the following centered data set.

Name	year	length
Carl Friedrich Gauss	−51.4	−5.6
Camille Jordan	9.6	6.4
Adrien-Marie Legendre	−76.4	−5.6
Bernhard Riemann	−2.4	9.4
David Hilbert	33.6	−3.6
Henri Poincaré	25.6	−0.6
Emmy Noether	53.6	−5.6
Karl Weierstrass	13.4	−5.6
Eugenio Beltrami	6.6	−3.6
Hermann Schwarz	14.6	14.4

See Figure 23.3.

We can think of the vector C_j as representing the features of X in the direction e_j (the jth canonical basis vector in $\mathbb{R}^d$, namely $e_j = (0, \ldots, 1, \ldots 0)$, with a 1 in the jth position).

If $v \in \mathbb{R}^d$ is a unit vector, we wish to consider the projection of the data points $X_1, \ldots, X_n$ onto the line spanned by v. Recall from Euclidean geometry that if $x \in \mathbb{R}^d$ is any vector and $v \in \mathbb{R}^d$ is a unit vector, the projection of x onto the line spanned by v is

$$\langle x, v \rangle v.$$

Thus, with respect to the basis v, the projection of x has coordinate $\langle x, v \rangle$. If x is represented by a row vector and v by a column vector, then

$$\langle x, v \rangle = xv.$$

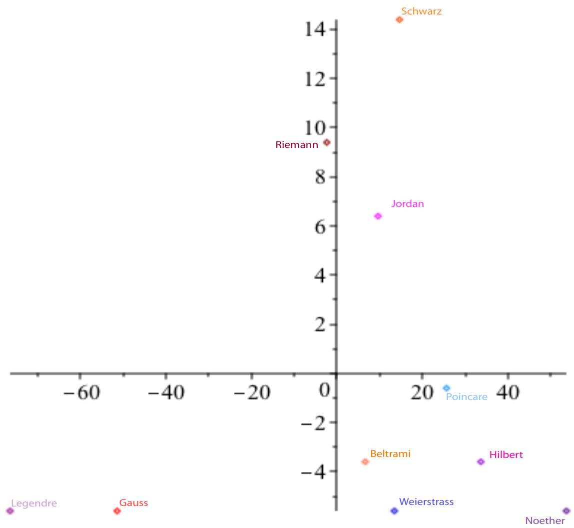


Figure 23.3: The centered data points of Example 23.9.

Therefore, the vector $Y \in \mathbb{R}^n$ consisting of the coordinates of the projections of $X_1, \ldots, X_n$ onto the line spanned by v is given by $Y = Xv$, and this is the linear combination

$$Xv = v_1 C_1 + \cdots + v_d C_d$$

of the columns of X (with $v = (v_1, \ldots, v_d)$).

Observe that because μ_j is the mean of the vector C_j (the jth column of X), we get

$$\overline{Y} = \overline{Xv} = v_1 \mu_1 + \cdots + v_d \mu_d,$$

and so the centered point $Y - \overline{Y}$ is given by

$$Y - \overline{Y} = v_1(C_1 - \mu_1) + \cdots + v_d(C_d - \mu_d) = (X - \mu)v.$$

Furthermore, if $Y = Xv$ and $Z = Xw$, then

$$\begin{aligned} \mathrm{cov}(Y, Z) &= \frac{((X-\mu)v)^\top (X-\mu)w}{n-1} \\ &= v^\top \frac{1}{n-1}(X-\mu)^\top (X-\mu)w \\ &= v^\top \Sigma w, \end{aligned}$$

where Σ is the covariance matrix of X. Since $Y - \overline{Y}$ has zero mean, we have

$$\mathrm{var}(Y) = \mathrm{var}(Y - \overline{Y}) = v^\top \frac{1}{n-1}(X-\mu)^\top (X-\mu)v.$$

The above suggests that we should move the origin to the centroid μ of the X_i's and consider the matrix $X - \mu$ of the centered data points $X_i - \mu$.

From now on beware that we denote the columns of $X - \mu$ by $C_1, \ldots, C_d$ and that Y denotes the *centered* point $Y = (X - \mu)v = \sum_{j=1}^{d} v_j C_j$, where v is a unit vector.

Basic idea of PCA: The principal components of X are *uncorrelated* projections Y of the data points X_1, …, X_n onto some directions v (where the v's are unit vectors) such that $\mathrm{var}(Y)$ is maximal.

This suggests the following definition:

Definition 23.2. Given an $n \times d$ matrix X of data points $X_1, \ldots, X_n$, if μ is the centroid of the X_i's, then a *first principal component of X (first PC)* is a centered point $Y_1 = (X - \mu)v_1$, the projection of $X_1, \ldots, X_n$ onto a direction v_1 such that $\mathrm{var}(Y_1)$ is maximized, where v_1 is a unit vector (recall that $Y_1 = (X - \mu)v_1$ is a linear combination of the C_j's, the columns of $X - \mu$).

More generally, if $Y_1, \ldots, Y_k$ are k principal components of X along some unit vectors $v_1, \ldots, v_k$, where $1 \leq k < d$, a *$(k+1)$th principal component of X ($(k+1)$th PC)* is a centered point $Y_{k+1} = (X - \mu)v_{k+1}$, the projection of $X_1, \ldots, X_n$ onto some direction v_{k+1} such that $\mathrm{var}(Y_{k+1})$ is maximized, subject to $\mathrm{cov}(Y_h, Y_{k+1}) = 0$ for all h with $1 \leq h \leq k$, and where v_{k+1} is a unit vector (recall that $Y_h = (X - \mu)v_h$ is a linear combination of the C_j's). The v_h are called *principal directions*.

The following proposition is the key to the main result about PCA. This result was already proven in Proposition 17.23 except that the eigenvalues were listed in increasing order. For the reader's convenience we prove it again.

Proposition 23.10. *If A is a symmetric $d \times d$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and if $(u_1, \ldots, u_d)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then*

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \lambda_1$$

(with the maximum attained for $x = u_1$) and

$$\max_{x \neq 0, x \in \{u_1, \ldots, u_k\}^\perp} \frac{x^\top A x}{x^\top x} = \lambda_{k+1}$$

(with the maximum attained for $x = u_{k+1}$), where $1 \leq k \leq d - 1$.

Proof. First observe that

$$\max_{x \neq 0} \frac{x^\top A x}{x^\top x} = \max_{x} \{x^\top A x \mid x^\top x = 1\},$$

and similarly,

$$\max_{x\neq 0, x\in\{u_1,\ldots,u_k\}^{\perp}} \frac{x^{\top}Ax}{x^{\top}x} = \max_x \left\{x^{\top}Ax \mid (x\in\{u_1,\ldots,u_k\}^{\perp}) \wedge (x^{\top}x = 1)\right\}.$$

Since A is a symmetric matrix, its eigenvalues are real and it can be diagonalized with respect to an orthonormal basis of eigenvectors, so let $(u_1,\ldots,u_d)$ be such a basis. If we write

$$x = \sum_{i=1}^{d} x_i u_i,$$

a simple computation shows that

$$x^{\top}Ax = \sum_{i=1}^{d} \lambda_i x_i^2.$$

If $x^{\top}x = 1$, then $\sum_{i=1}^{d} x_i^2 = 1$, and since we assumed that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, we get

$$x^{\top}Ax = \sum_{i=1}^{d} \lambda_i x_i^2 \leq \lambda_1 \left(\sum_{i=1}^{d} x_i^2\right) = \lambda_1.$$

Thus,

$$\max_x \left\{x^{\top}Ax \mid x^{\top}x = 1\right\} \leq \lambda_1,$$

and since this maximum is achieved for $e_1 = (1,0,\ldots,0)$, we conclude that

$$\max_x \left\{x^{\top}Ax \mid x^{\top}x = 1\right\} = \lambda_1.$$

Next observe that $x\in\{u_1,\ldots,u_k\}^{\perp}$ and $x^{\top}x = 1$ iff $x_1 = \cdots = x_k = 0$ and $\sum_{i=1}^{d} x_i = 1$. Consequently, for such an x, we have

$$x^{\top}Ax = \sum_{i=k+1}^{d} \lambda_i x_i^2 \leq \lambda_{k+1} \left(\sum_{i=k+1}^{d} x_i^2\right) = \lambda_{k+1}.$$

Thus,

$$\max_x \left\{x^{\top}Ax \mid (x\in\{u_1,\ldots,u_k\}^{\perp}) \wedge (x^{\top}x = 1)\right\} \leq \lambda_{k+1},$$

and since this maximum is achieved for $e_{k+1} = (0,\ldots,0,1,0,\ldots,0)$ with a 1 in position $k+1$, we conclude that

$$\max_x \left\{x^{\top}Ax \mid (x\in\{u_1,\ldots,u_k\}^{\perp}) \wedge (x^{\top}x = 1)\right\} = \lambda_{k+1},$$

as claimed. $\square$

The quantity

$$\frac{x^\top Ax}{x^\top x}$$

is known as the *Rayleigh ratio* or *Rayleigh–Ritz ratio* (see Section 17.6) and Proposition 23.10 is often known as part of the *Rayleigh–Ritz theorem*.

Proposition 23.10 also holds if A is a Hermitian matrix and if we replace $x^\top Ax$ by x^*Ax and $x^\top x$ by x^*x. The proof is unchanged, since a Hermitian matrix has real eigenvalues and is diagonalized with respect to an orthonormal basis of eigenvectors (with respect to the Hermitian inner product).

We then have the following fundamental result showing how *the SVD of X yields the PCs*:

Theorem 23.11. *(SVD yields PCA) Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then the centered points $Y_1, \ldots, Y_d$, where*

$$Y_k = (X - \mu)u_k = k\textit{th column of } VD$$

and u_k is the kth column of U, are d principal components of X. Furthermore,

$$\mathrm{var}(Y_k) = \frac{\sigma_k^2}{n-1}$$

and $\mathrm{cov}(Y_h, Y_k) = 0$*, whenever $h \neq k$ and $1 \leq k, h \leq d$.*

Proof. Recall that for any unit vector v, the centered projection of the points $X_1, \ldots, X_n$ onto the line of direction v is $Y = (X - \mu)v$ and that the variance of Y is given by

$$\mathrm{var}(Y) = v^\top \frac{1}{n-1}(X - \mu)^\top (X - \mu)v.$$

Since $X - \mu = VDU^\top$, we get

$$\begin{aligned}
\mathrm{var}(Y) &= v^\top \frac{1}{(n-1)}(X - \mu)^\top (X - \mu)v \\
&= v^\top \frac{1}{(n-1)} UDV^\top VDU^\top v \\
&= v^\top U \frac{1}{(n-1)} D^2 U^\top v.
\end{aligned}$$

Similarly, if $Y = (X - \mu)v$ and $Z = (X - \mu)w$, then the covariance of Y and Z is given by

$$\mathrm{cov}(Y, Z) = v^\top U \frac{1}{(n-1)} D^2 U^\top w.$$

Obviously, $U\frac{1}{(n-1)}D^2U^\top$ is a symmetric matrix whose eigenvalues are $\frac{\sigma_1^2}{n-1} \geq \cdots \geq \frac{\sigma_d^2}{n-1}$, and the columns of U form an orthonormal basis of unit eigenvectors.

We proceed by induction on k. For the base case, $k = 1$, maximizing $\mathrm{var}(Y)$ is equivalent to maximizing

$$v^\top U\frac{1}{(n-1)}D^2U^\top v,$$

where v is a unit vector. By Proposition 23.10, the maximum of the above quantity is the largest eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_1^2}{n-1}$, and it is achieved for u_1, the first columnn of U. Now we get

$$Y_1 = (X - \mu)u_1 = VDU^\top u_1,$$

and since the columns of U form an orthonormal basis, $U^\top u_1 = e_1 = (1, 0, \ldots, 0)$, and so Y_1 is indeed the first column of VD.

By the induction hypothesis, the centered points $Y_1, \ldots, Y_k$, where $Y_h = (X - \mu)u_h$ and $u_1, \ldots, u_k$ are the first k columns of U, are k principal components of X. Because

$$\mathrm{cov}(Y, Z) = v^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where $Y = (X - \mu)v$ and $Z = (X - \mu)w$, the condition $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to the fact that w belongs to the orthogonal complement of the subspace spanned by $\{u_1, \ldots, u_k\}$, and maximizing $\mathrm{var}(Z)$ subject to $\mathrm{cov}(Y_h, Z) = 0$ for $h = 1, \ldots, k$ is equivalent to maximizing

$$w^\top U\frac{1}{(n-1)}D^2U^\top w,$$

where w is a unit vector orthogonal to the subspace spanned by $\{u_1, \ldots, u_k\}$. By Proposition 23.10, the maximum of the above quantity is the $(k+1)$th eigenvalue of $U\frac{1}{(n-1)}D^2U^\top$, namely $\frac{\sigma_{k+1}^2}{n-1}$, and it is achieved for u_{k+1}, the $(k+1)$th columnn of U. Now we get

$$Y_{k+1} = (X - \mu)u_{k+1} = VDU^\top u_{k+1},$$

and since the columns of U form an orthonormal basis, $U^\top u_{k+1} = e_{k+1}$, and Y_{k+1} is indeed the $(k+1)$th column of VD, which completes the proof of the induction step. $\square$

The d columns $u_1, \ldots, u_d$ of U are usually called the *principal directions* of $X - \mu$ (and X). We note that not only do we have $\mathrm{cov}(Y_h, Y_k) = 0$ whenever $h \neq k$, but the directions $u_1, \ldots, u_d$ along which the data are projected are mutually orthogonal.

Example 23.10. For the centered data set of our bearded mathematicians (Example 23.9) we have $X - \mu = V\Sigma U^\top$, where Σ has two nonzero singular values, $\sigma_1 = 116.9803, \sigma_2 = 21.7812$, and with

$$U = \begin{pmatrix} 0.9995 & 0.0325 \\ 0.0325 & -0.9995 \end{pmatrix},$$

so the principal directions are $u_1 = (0.9995, 0.0325)$ and $u_2 = (0.0325, -0.9995)$. Observe that u_1 is almost the direction of the x-axis, and u_2 is almost the opposite direction of the y-axis. We also find that the projections Y_1 and Y_2 along the principal directions are

$$VD = \begin{pmatrix} -51.5550 & 3.9249 \\ 9.8031 & -6.0843 \\ -76.5417 & 3.1116 \\ -2.0929 & -9.4731 \\ 33.4651 & 4.6912 \\ 25.5669 & 1.4325 \\ 53.3894 & 7.3408 \\ 13.2107 & 6.0330 \\ 6.4794 & 3.8128 \\ 15.0607 & -13.9174 \end{pmatrix}, \quad \text{with} \quad X - \mu = \begin{pmatrix} -51.4000 & -5.6000 \\ 9.6000 & 6.4000 \\ -76.4000 & -5.6000 \\ -2.4000 & 9.4000 \\ 33.6000 & -3.6000 \\ 25.6000 & -0.6000 \\ 53.6000 & -5.6000 \\ 13.4000 & -5.6000 \\ 6.6000 & -3.6000 \\ 14.6000 & 14.4000 \end{pmatrix}.$$

See Figures 23.4, 23.5, and 23.6.

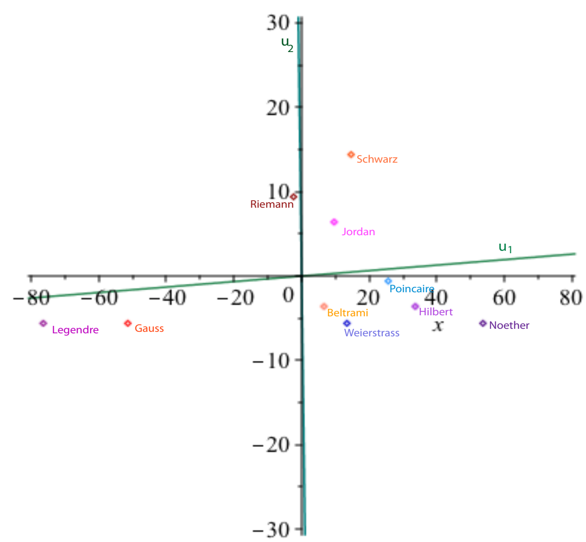


Figure 23.4: The centered data points of Example 23.9 and the two principal directions of Example 23.10.

We know from our study of SVD that $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of the symmetric positive semidefinite matrix $(X - \mu)^\top (X - \mu)$ and that $u_1, \ldots, u_d$ are corresponding eigenvectors. Numerically, it is preferable to use SVD on $X - \mu$ rather than to compute explicitly $(X - \mu)^\top (X - \mu)$ and then diagonalize it. Indeed, the explicit computation of $A^\top A$ from a matrix A can be numerically quite unstable, and good SVD algorithms avoid computing $A^\top A$ explicitly.

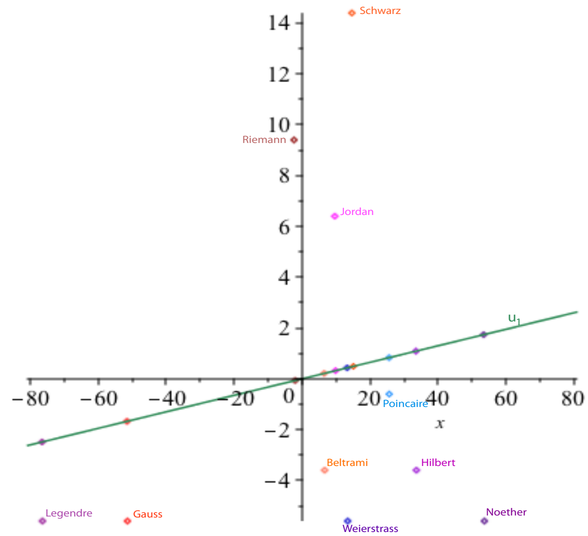


Figure 23.5: The first principal components of Example 23.10, i.e. the projection of the centered data points onto the u_1 line.

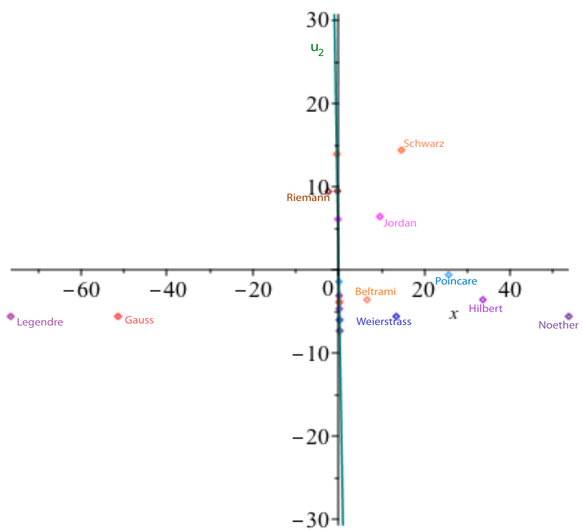


Figure 23.6: The second principal components of Example 23.10, i.e. the projection of the centered data points onto the u_2 line.

In general, since an SVD of X is not unique, *the principal directions* $u_1, \ldots, u_d$ *are not unique*. This can happen when a data set has some *rotational symmetries*, and in such a case, PCA is not a very good method for analyzing the data set.

23.5 Best Affine Approximation

A problem very close to PCA (and based on least squares) is to *best approximate a data set of* n *points* $X_1, \ldots, X_n$, *with* $X_i \in \mathbb{R}^d$, *by a* p*-dimensional affine subspace* A *of* $\mathbb{R}^d$, *with* $1 \leq p \leq d-1$ (the terminology rank $d-p$ is also used).

First consider $p = d-1$. Then $A = A_1$ is an affine hyperplane (in $\mathbb{R}^d$), and it is given by an equation of the form

$$a_1x_1 + \cdots + a_dx_d + c = 0.$$

By *best approximation*, we mean that $(a_1, \ldots, a_d, c)$ solves the homogeneous linear system

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_d \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

in the *least squares sense, subject to the condition that* $a = (a_1, \ldots, a_d)$ *is a unit vector*, that is, $a^\top a = 1$, where $X_i = (x_{i\,1}, \cdots, x_{i\,d})$.

If we form the symmetric matrix

$$\begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}^\top \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n\,1} & \cdots & x_{n\,d} & 1 \end{pmatrix}$$

involved in the normal equations, we see that the bottom row (and last column) of that matrix is

$$n\mu_1 \quad \cdots \quad n\mu_d \quad n,$$

where $n\mu_j = \sum_{i=1}^n x_{i\,j}$ is n times the mean of the column C_j of X.

Therefore, if $(a_1, \ldots, a_d, c)$ is a least squares solution, that is, a solution of the normal equations, we must have

$$n\mu_1a_1 + \cdots + n\mu_da_d + nc = 0,$$

that is,

$$a_1\mu_1 + \cdots + a_d\mu_d + c = 0,$$

which means that the *hyperplane* A_1 *must pass through the centroid* μ *of the data points* $X_1, \ldots, X_n$. Then we can rewrite the original system with respect to the centered data $X_i - \mu$, find that the variable c drops out, get the system

$$(X - \mu)a = 0,$$

where $a = (a_1, \ldots, a_d)$.

Thus, we are looking for a unit vector a solving $(X - \mu)a = 0$ in the least squares sense, that is, some a such that $a^\top a = 1$ minimizing

$$a^\top (X - \mu)^\top (X - \mu)a.$$

Compute some SVD $VDU^\top$ of $X - \mu$, where the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. Then

$$a^\top (X - \mu)^\top (X - \mu)a = a^\top U D^2 U^\top a,$$

where $D^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix, so pick a to be *the last column in* U (corresponding to the smallest eigenvalue σ_d^2 of $(X - \mu)^\top (X - \mu)$). This is a solution to our best fit problem.

Therefore, if U_{d-1} is the linear hyperplane defined by a, that is,

$$U_{d-1} = \{u \in \mathbb{R}^d \mid \langle u, a \rangle = 0\},$$

where a is the last column in U for some SVD $VDU^\top$ of $X - \mu$, we have shown that the affine hyperplane $A_1 = \mu + U_{d-1}$ is a best approximation of the data set $X_1, \ldots, X_n$ in the least squares sense.

It is easy to show that this hyperplane $A_1 = \mu + U_{d-1}$ minimizes the sum of the square distances of each X_i to its orthogonal projection onto A_1. Also, since U_{d-1} is the orthogonal complement of a, the last column of U, we see that U_{d-1} is spanned by the first $d-1$ columns of U, that is, the first $d - 1$ principal directions of $X - \mu$.

All this can be generalized to a *best* $(d-k)$*-dimensional affine subspace* A_k *approximating* $X_1, \ldots, X_n$ *in the least squares sense* $(1 \leq k \leq d-1)$. Such an affine subspace A_k is cut out by k independent hyperplanes H_i (with $1 \leq i \leq k$), each given by some equation

$$a_{i\,1} x_1 + \cdots + a_{i\,d} x_d + c_i = 0.$$

If we write $a_i = (a_{i\,1}, \cdots, a_{i\,d})$, to say that the H_i are independent means that $a_1, \ldots, a_k$ are linearly independent. In fact, we may assume that $a_1, \ldots, a_k$ form an *orthonormal system*.

Then finding a best $(d - k)$-dimensional affine subspace A_k amounts to solving the homogeneous linear system

$$\begin{pmatrix} X & \mathbf{1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X & \mathbf{1} \end{pmatrix} \begin{pmatrix} a_1 \\ c_1 \\ \vdots \\ a_k \\ c_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

in the least squares sense, subject to the conditions $a_i^\top a_j = \delta_{i\,j}$, for all i, j with $1 \leq i, j \leq k$, where the matrix of the system is a block diagonal matrix consisting of k diagonal blocks $(X, \mathbf{1})$, where $\mathbf{1}$ denotes the column vector $(1, \ldots, 1) \in \mathbb{R}^n$.

Again it is easy to see that each hyperplane H_i *must pass through the centroid* μ of $X_1, \ldots, X_n$, and by switching to the centered data $X_i - \mu$ we get the system

$$\begin{pmatrix} X-\mu & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X-\mu \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

with $a_i^\top a_j = \delta_{i\,j}$ for all i, j with $1 \leq i, j \leq k$.

If $VDU^\top = X - \mu$ is an SVD decomposition, it is easy to see that a least squares solution of this system is given by the last k columns of U, assuming that the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ of $X - \mu$ arranged in descending order. But now the $(d-k)$-dimensional subspace U_{d-k} cut out by the hyperplanes defined by $a_1, \ldots, a_k$ is simply the orthogonal complement of $(a_1, \ldots, a_k)$, which is the subspace spanned by the first $d - k$ columns of U.

So the best $(d-k)$-dimensional affine subpsace A_k approximating $X_1, \ldots, X_n$ in the least squares sense is

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ principal directions of $X-\mu$, that is, the first $d-k$ columns of U. Consequently, we get the following interesting interpretation of PCA (actually, principal directions):

Theorem 23.12. *Let X be an $n \times d$ matrix of data points $X_1, \ldots, X_n$, and let μ be the centroid of the X_i's. If $X - \mu = VDU^\top$ is an SVD decomposition of $X - \mu$ and if the main diagonal of D consists of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then a best $(d-k)$-dimensional affine approximation A_k of $X_1, \ldots, X_n$ in the least squares sense is given by*

$$A_k = \mu + U_{d-k},$$

where U_{d-k} is the linear subspace spanned by the first $d-k$ columns of U, the first $d-k$ principal directions of $X - \mu$ ($1 \leq k \leq d-1$).

Example 23.11. Going back to Example 23.10, a best 1-dimensional affine approximation A_1 is the affine line passing through $(\mu_1, \mu_2) = (1824.4, 5.6)$ of direction $u_1 = (0.9995, 0.0325)$.

Example 23.12. Suppose in the data set of Example 23.5 that we add the month of birth of every mathematician as a feature. We obtain the following data set.

Name	month	year	length
Carl Friedrich Gauss	4	1777	0
Camille Jordan	1	1838	12
Adrien-Marie Legendre	9	1752	0
Bernhard Riemann	9	1826	15
David Hilbert	1	1862	2
Henri Poincaré	4	1854	5
Emmy Noether	3	1882	0
Karl Weierstrass	10	1815	0
Eugenio Beltrami	10	1835	2
Hermann Schwarz	1	1843	20

The mean of the first column is 5.2, and the centered data set is given below.

Name	month	year	length
Carl Friedrich Gauss	−1.2	−51.4	−5.6
Camille Jordan	−4.2	9.6	6.4
Adrien-Marie Legendre	3.8	−76.4	−5.6
Bernhard Riemann	3.8	−2.4	9.4
David Hilbert	−4.2	33.6	−3.6
Henri Poincaré	−1.2	25.6	−0.6
Emmy Noether	−2.2	53.6	−5.6
Karl Weierstrass	4.8	13.4	−5.6
Eugenio Beltrami	4.8	6.6	−3.6
Hermann Schwarz	−4.2	14.6	14.4

Running SVD on this data set we get

$$U = \begin{pmatrix} 0.0394 & 0.1717 & 0.9844 \\ -0.9987 & 0.0390 & 0.0332 \\ -0.0327 & -0.9844 & 0.1730 \end{pmatrix}, \qquad D = \begin{pmatrix} 117.0706 & 0 & 0 \\ 0 & 22.0390 & 0 \\ 0 & 0 & 10.1571 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$VD = \begin{pmatrix} 51.4683 & 3.3013 & -3.8569 \\ -9.9623 & -6.6467 & -2.7082 \\ 76.6327 & 3.1845 & 0.2348 \\ 2.2393 & -8.6943 & 5.2872 \\ -33.6038 & 4.1334 & -3.6415 \\ -25.5941 & 1.3833 & -0.4350 \\ -53.4333 & 7.2258 & -1.3547 \\ -13.0100 & 6.8594 & 4.2010 \\ -6.2843 & 4.6254 & 4.3212 \\ -15.2173 & -14.3266 & -1.1581 \end{pmatrix}, \quad X - \mu = \begin{pmatrix} -1.2000 & -51.4000 & -5.6000 \\ -4.2000 & 9.6000 & 6.4000 \\ 3.8000 & -76.4000 & -5.6000 \\ 3.8000 & -2.4000 & 9.4000 \\ -4.2000 & 33.6000 & -3.6000 \\ -1.2000 & 25.6000 & -0.6000 \\ -2.2000 & 53.6000 & -5.6000 \\ 4.8000 & 13.4000 & -5.6000 \\ 4.8000 & 6.6000 & -3.6000 \\ -4.2000 & 14.6000 & 14.4000 \end{pmatrix}.$$

The first principal direction $u_1 = (0.0394, -0.9987, -0.0327)$ is basically the opposite of the y-axis, and the most significant feature is the year of birth. The second principal direction $u_2 = (0.1717, 0.0390, -0.9844)$ is close to the opposite of the z-axis, and the second most significant feature is the lenght of beards. A best affine plane is spanned by the vectors u_1 and u_2.

There are many applications of PCA to data compression, dimension reduction, and pattern analysis. The basic idea is that in many cases, given a data set $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, only a "small" subset of $m < d$ of the features is needed to describe the data set accurately.

If $u_1, \ldots, u_d$ are the principal directions of $X - \mu$, then the first m projections of the data (the first m principal components, i.e., the first m columns of VD) onto the first m principal directions represent the data without much loss of information. Thus, instead of using the original data points $X_1, \ldots, X_n$, with $X_i \in \mathbb{R}^d$, we can use their projections onto the first m principal directions $Y_1, \ldots, Y_m$, where $Y_i \in \mathbb{R}^m$ and $m < d$, obtaining a compressed version of the original data set.

For example, PCA is used in computer vision for *face recognition*. Sirovitch and Kirby (1987) seem to be the first to have had the idea of using PCA to compress facial images. They introduced the term *eigenpicture* to refer to the principal directions, u_i. However, an explicit face recognition algorithm was given only later by Turk and Pentland (1991). They renamed eigenpictures as *eigenfaces*.

For details on the topic of eigenfaces, see Forsyth and Ponce [64] (Chapter 22, Section 22.3.2), where you will also find exact references to Turk and Pentland's papers.

Another interesting application of PCA is to the *recognition of handwritten digits*. Such an application is described in Hastie, Tibshirani, and Friedman, [88] (Chapter 14, Section 14.5.1).

23.6 Summary

The main concepts and results of this chapter are listed below:

- *Least squares problems.*
- Existence of a least squares solution of smallest norm (Theorem 23.1).
- The *pseudo-inverse* A^+ of a matrix A.
- The least squares solution of smallest norm is given by the pseudo-inverse (Theorem 23.2)
- Projection properties of the pseudo-inverse.
- The pseudo-inverse of a normal matrix.
- The *Penrose characterization* of the pseudo-inverse.
- Data compression and SVD.
- Best approximation of rank $< r$ of a matrix.
- *Principal component analysis.*
- Review of basic statistical concepts: *mean, variance, covariance*, *covariance matrix*.
- Centered data, *centroid*.
- The *principal components* (*PCA*).
- The *Rayleigh–Ritz theorem* (Theorem 23.10).
- The main theorem: *SVD yields PCA* (Theorem 23.11).
- Best affine approximation.
- SVD yields a best affine approximation (Theorem 23.12).
- Face recognition, eigenfaces.

23.7 Problems

Problem 23.1. Consider the overdetermined system in the single variable x:

$$a_1 x = b_1, \ldots, a_m x = b_m,$$

with $a_1^2 + \cdots + a_m^2 \neq 0$. Prove that the least squares solution of smallest norm is given by

$$x^+ = \frac{a_1 b_1 + \cdots + a_m b_m}{a_1^2 + \cdots + a_m^2}.$$

Problem 23.2. Let X be an $m \times n$ real matrix. For any strictly positive constant $K > 0$, the matrix $X^\top X + KI_n$ is invertible. Prove that the limit of the matrix $(X^\top X + KI_n)^{-1}X^\top$ when K goes to zero is equal to the pseudo-inverse X^+ of X.

Problem 23.3. Use `Matlab` to find the pseudo-inverse of the 8×6 matrix

$$A = \begin{pmatrix} 64 & 2 & 3 & 61 & 60 & 6 \\ 9 & 55 & 54 & 12 & 13 & 51 \\ 17 & 47 & 46 & 20 & 21 & 43 \\ 40 & 26 & 27 & 37 & 36 & 30 \\ 32 & 34 & 35 & 29 & 28 & 38 \\ 41 & 23 & 22 & 44 & 45 & 19 \\ 49 & 15 & 14 & 52 & 53 & 11 \\ 8 & 58 & 59 & 5 & 4 & 62 \end{pmatrix}.$$

Observe that the sums of the columns are all equal to 260. Let b be the vector of dimension 8 whose coordinates are all equal to 256. Find the solution x^+ of the system $Ax = b$.

Problem 23.4. The purpose of this problem is to show that Proposition 23.9 (the Eckart–Young theorem) also holds for the Frobenius norm. This problem is adapted from Strang [171], Section I.9.

Suppose the $m \times n$ matrix B of rank at most k minimizes $\|A - B\|_F$. Start with an SVD of B,

$$B = V \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} U^\top,$$

where D is a diagonal $k \times k$ matrix. We can write

$$A = V \begin{pmatrix} L + E + R & F \\ G & H \end{pmatrix} U^\top,$$

where L is strictly lower triangular in the first k rows, E is diagonal, and R is strictly upper triangular, and let

$$C = V \begin{pmatrix} L + D + R & F \\ 0 & 0 \end{pmatrix} U^\top,$$

which clearly has rank $\leq k$.

(1) Prove that

$$\|A - B\|_F^2 = \|A - C\|_F^2 + \|L\|_F^2 + \|R\|_F^2 + \|F\|_F^2.$$

Since $\|A - B\|_F$ is minimal, show that $L = R = F = 0$.

Similarly, show that $G = 0$.

(2) We have

$$V^\top AU = \begin{pmatrix} E & 0 \\ 0 & H \end{pmatrix}, \quad V^\top BU = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},$$

where E is diagonal, so deduce that

1. $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$.
2. The singular values of H must be the smallest $n-k$ singular values of A.
3. The minimum of $\|A - B\|_F$ must be $\|H\|_F = (\sigma_{k+1}^2 + \cdots + \sigma_r^2)^{1/2}$.

Problem 23.5. Prove that the closest rank 1 approximation (in $\| \|_2$) of the matrix

$$A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$$

is

$$A_1 = \frac{3}{2}\begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix}.$$

Show that the Eckart–Young theorem fails for the operator norm $\| \|_\infty$ by finding a rank 1 matrix B such that $\|A - B\|_\infty < \|A - A_1\|_\infty$.

Problem 23.6. Find a closest rank 1 approximation (in $\| \|_2$) for the matrices

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 3 \\ 2 & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Problem 23.7. Find a closest rank 1 approximation (in $\| \|_2$) for the matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Problem 23.8. Let S be a real symmetric positive definite matrix and let $S = U\Sigma U^\top$ be a diagonalization of S. Prove that the closest rank 1 matrix (in the L^2-norm) to S is $u_1\sigma_1 u_1^\top$, where u_1 is the first column of U.

Part II

Affine and Projective Geometry

Chapter 24

Basics of Affine Geometry

L'algèbre n'est qu'une géométrie écrite; la géométrie n'est qu'une algèbre figurée.
—**Sophie Germain**

24.1 Affine Spaces

Geometrically, curves and surfaces are usually considered to be sets of points with some special properties, living in a space consisting of "points." Typically, one is also interested in geometric properties invariant under certain transformations, for example, translations, rotations, projections, etc. One could model the space of points as a vector space, but this is not very satisfactory for a number of reasons. One reason is that the point corresponding to the zero vector (0), called the origin, plays a special role, when there is really no reason to have a privileged origin. Another reason is that certain notions, such as parallelism, are handled in an awkward manner. But the deeper reason is that vector spaces and affine spaces really have different geometries. The geometric properties of a vector space are invariant under the group of bijective linear maps, whereas the geometric properties of an affine space are invariant under the group of bijective affine maps, and these two groups are not isomorphic. Roughly speaking, there are more affine maps than linear maps.

Affine spaces provide a better framework for doing geometry. In particular, it is possible to deal with points, curves, surfaces, etc., in an **intrinsic manner**, that is, independently of any specific choice of a coordinate system. As in physics, this is highly desirable to really understand what is going on. Of course, coordinate systems have to be chosen to finally carry out computations, but one should learn to resist the temptation to resort to coordinate systems until it is really necessary.

Affine spaces are the right framework for dealing with motions, trajectories, and physical forces, among other things. Thus, affine geometry is crucial to a clean presentation of kinematics, dynamics, and other parts of physics (for example, elasticity). After all, a rigid motion is an affine map, but not a linear map in general. Also, given an $m \times n$ matrix A

and a vector $b \in \mathbb{R}^m$, the set $U = \{x \in \mathbb{R}^n \mid Ax = b\}$ of solutions of the system $Ax = b$ is an affine space, but not a vector space (linear space) in general.

Use coordinate systems only when needed!

This chapter proceeds as follows. We take advantage of the fact that almost every affine concept is the counterpart of some concept in linear algebra. We begin by defining affine spaces, stressing the physical interpretation of the definition in terms of points (particles) and vectors (forces). Corresponding to linear combinations of vectors, we define affine combinations of points (barycenters), realizing that we are forced to restrict our attention to families of scalars adding up to 1. Corresponding to linear subspaces, we introduce affine subspaces as subsets closed under affine combinations. Then, we characterize affine subspaces in terms of certain vector spaces called their directions. This allows us to define a clean notion of parallelism. Next, corresponding to linear independence and bases, we define affine independence and affine frames. We also define convexity. Corresponding to linear maps, we define affine maps as maps preserving affine combinations. We show that every affine map is completely defined by the image of one point and a linear map. Then, we investigate briefly some simple affine maps, the translations and the central dilatations. At this point, we give a glimpse of affine geometry. We prove the theorems of Thales, Pappus, and Desargues. After this, the definition of affine hyperplanes in terms of affine forms is reviewed. The section ends with a closer look at the intersection of affine subspaces.

Our presentation of affine geometry is far from being comprehensive, and it is biased toward the algorithmic geometry of curves and surfaces. For more details, the reader is referred to Pedoe [136], Snapper and Troyer [162], Berger [11, 12], Coxeter [44], Samuel [142], Tisseron [175], Fresnel [65], Vienne [185], and Hilbert and Cohn-Vossen [92].

Suppose we have a particle moving in 3D space and that we want to describe the trajectory of this particle. If one looks up a good textbook on dynamics, such as Greenwood [82], one finds out that the particle is modeled as a point, and that the position of this point x is determined with respect to a "frame" in $\mathbb{R}^3$ by a vector. Curiously, the notion of a frame is rarely defined precisely, but it is easy to infer that a frame is a pair $(O, (e_1, e_2, e_3))$ consisting of an origin O (which is a point) together with a basis of three vectors (e_1, e_2, e_3). For example, the standard frame in $\mathbb{R}^3$ has origin $O = (0, 0, 0)$ and the basis of three vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. The position of a point x is then defined by the "unique vector" from O to x.

But wait a minute, this definition seems to be defining frames and the position of a point without defining what a point is! Well, let us identify points with elements of $\mathbb{R}^3$. If so, given any two points $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$, there is a unique *free vector*, denoted by $\overrightarrow{ab}$, from a to b, the vector $\overrightarrow{ab} = (b_1 - a_1, b_2 - a_2, b_3 - a_3)$. Note that

$$b = a + \overrightarrow{ab},$$

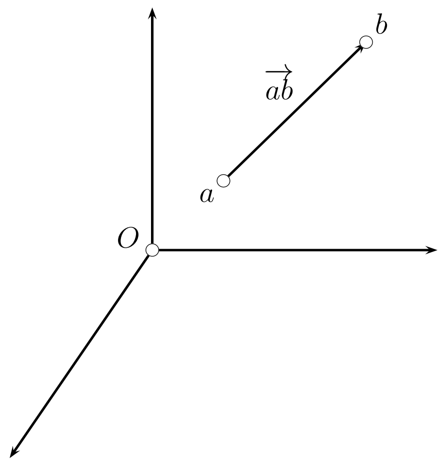


Figure 24.1: Points and free vectors.

addition being understood as addition in $\mathbb{R}^3$. Then, in the standard frame, given a point $x = (x_1, x_2, x_3)$, the position of x is the vector $\overrightarrow{Ox} = (x_1, x_2, x_3)$, which coincides with the point itself. In the standard frame, points and vectors are identified. Points and free vectors are illustrated in Figure 24.1.

What if we pick a frame with a different origin, say $\Omega = (\omega_1, \omega_2, \omega_3)$, but the same basis vectors (e_1, e_2, e_3)? This time, the point $x = (x_1, x_2, x_3)$ is defined by two position vectors:

$$\overrightarrow{Ox} = (x_1, x_2, x_3)$$

in the frame $(O, (e_1, e_2, e_3))$ and

$$\overrightarrow{\Omega x} = (x_1 - \omega_1, x_2 - \omega_2, x_3 - \omega_3)$$

in the frame $(\Omega, (e_1, e_2, e_3))$. See Figure 24.2.

This is because

$$\overrightarrow{Ox} = \overrightarrow{O\Omega} + \overrightarrow{\Omega x} \quad \text{and} \quad \overrightarrow{O\Omega} = (\omega_1, \omega_2, \omega_3).$$

We note that in the second frame $(\Omega, (e_1, e_2, e_3))$, points and position vectors are no longer identified. This gives us evidence that points are not vectors. It may be computationally convenient to deal with points using position vectors, but such a treatment is not frame invariant, which has undesirable effects.

Inspired by physics, we deem it important to define points and properties of points that are frame invariant. An undesirable side effect of the present approach shows up if we attempt to define linear combinations of points. First, let us review the notion of linear combination of vectors. Given two vectors u and v of coordinates (u_1, u_2, u_3) and (v_1, v_2, v_3) with respect

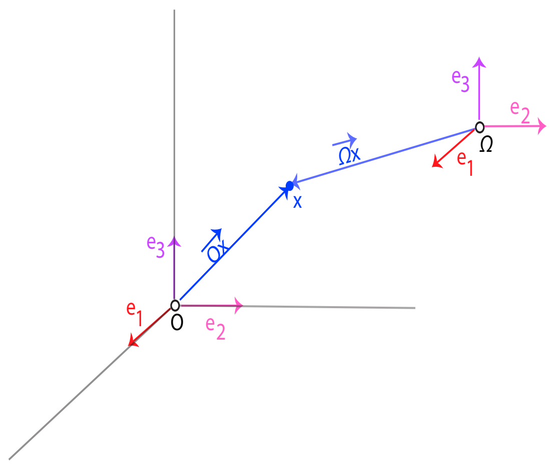


Figure 24.2: The two position vectors for the point x.

to the basis (e_1, e_2, e_3), for any two scalars λ, μ, we can define the linear combination $\lambda u + \mu v$ as the vector of coordinates

$$(\lambda u_1 + \mu v_1, \lambda u_2 + \mu v_2, \lambda u_3 + \mu v_3).$$

If we choose a different basis (e'_1, e'_2, e'_3) and if the matrix P expressing the vectors (e'_1, e'_2, e'_3) over the basis (e_1, e_2, e_3) is

$$P = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix},$$

which means that the columns of P are the coordinates of the e'_j over the basis (e_1, e_2, e_3), since

$$u_1 e_1 + u_2 e_2 + u_3 e_3 = u'_1 e'_1 + u'_2 e'_2 + u'_3 e'_3$$

and

$$v_1 e_1 + v_2 e_2 + v_3 e_3 = v'_1 e'_1 + v'_2 e'_2 + v'_3 e'_3,$$

it is easy to see that the coordinates (u_1, u_2, u_3) and (v_1, v_2, v_3) of u and v with respect to the basis (e_1, e_2, e_3) are given in terms of the coordinates (u'_1, u'_2, u'_3) and (v'_1, v'_2, v'_3) of u and v with respect to the basis (e'_1, e'_2, e'_3) by the matrix equations

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = P \begin{pmatrix} u'_1 \\ u'_2 \\ u'_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = P \begin{pmatrix} v'_1 \\ v'_2 \\ v'_3 \end{pmatrix}.$$

From the above, we get

$$\begin{pmatrix} u_1' \\ u_2' \\ u_3' \end{pmatrix} = P^{-1} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v_1' \\ v_2' \\ v_3' \end{pmatrix} = P^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

and by linearity, the coordinates

$$(\lambda u_1' + \mu v_1', \lambda u_2' + \mu v_2', \lambda u_3' + \mu v_3')$$

of $\lambda u + \mu v$ with respect to the basis (e_1', e_2', e_3') are given by

$$\begin{pmatrix} \lambda u_1' + \mu v_1' \\ \lambda u_2' + \mu v_2' \\ \lambda u_3' + \mu v_3' \end{pmatrix} = \lambda P^{-1} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \mu P^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = P^{-1} \begin{pmatrix} \lambda u_1 + \mu v_1 \\ \lambda u_2 + \mu v_2 \\ \lambda u_3 + \mu v_3 \end{pmatrix}.$$

Everything worked out because the change of basis does not involve a change of origin. On the other hand, if we consider the change of frame from the frame $(O, (e_1, e_2, e_3))$ to the frame $(\Omega, (e_1, e_2, e_3))$, where $\overrightarrow{O\Omega} = (\omega_1, \omega_2, \omega_3)$, given two points a, b of coordinates (a_1, a_2, a_3) and (b_1, b_2, b_3) with respect to the frame $(O, (e_1, e_2, e_3))$ and of coordinates (a_1', a_2', a_3') and (b_1', b_2', b_3') with respect to the frame $(\Omega, (e_1, e_2, e_3))$, since

$$(a_1', a_2', a_3') = (a_1 - \omega_1, a_2 - \omega_2, a_3 - \omega_3)$$

and

$$(b_1', b_2', b_3') = (b_1 - \omega_1, b_2 - \omega_2, b_3 - \omega_3),$$

the coordinates of $\lambda a + \mu b$ with respect to the frame $(O, (e_1, e_2, e_3))$ are

$$(\lambda a_1 + \mu b_1, \lambda a_2 + \mu b_2, \lambda a_3 + \mu b_3),$$

but the coordinates

$$(\lambda a_1' + \mu b_1', \lambda a_2' + \mu b_2', \lambda a_3' + \mu b_3')$$

of $\lambda a + \mu b$ with respect to the frame $(\Omega, (e_1, e_2, e_3))$ are

$$(\lambda a_1 + \mu b_1 - (\lambda + \mu)\omega_1, \lambda a_2 + \mu b_2 - (\lambda + \mu)\omega_2, \lambda a_3 + \mu b_3 - (\lambda + \mu)\omega_3),$$

which are different from

$$(\lambda a_1 + \mu b_1 - \omega_1, \lambda a_2 + \mu b_2 - \omega_2, \lambda a_3 + \mu b_3 - \omega_3),$$

unless $\lambda + \mu = 1$. See Figure 24.3.

Thus, we have discovered a major difference between vectors and points: The notion of linear combination of vectors is basis independent, but the notion of linear combination of points is frame dependent. In order to salvage the notion of linear combination of points, some restriction is needed: The scalar coefficients must add up to 1.

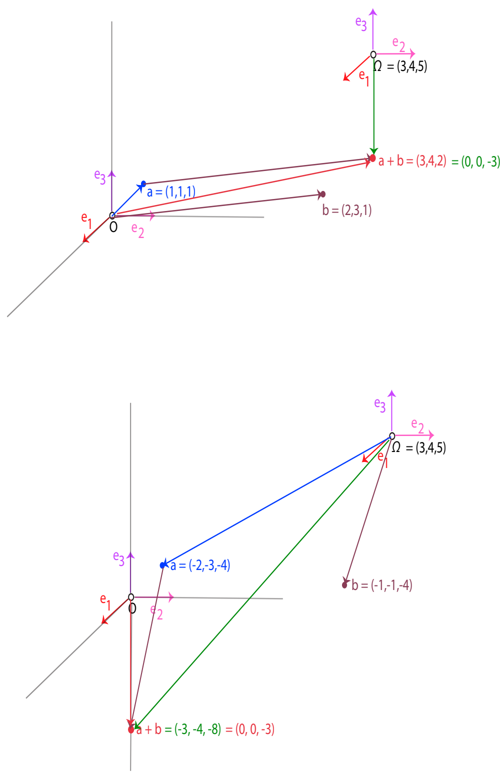


Figure 24.3: The top figure shows the location of the "point" sum $a+b$ with respect to the frame $(O,(e_1,e_2,e_3))$, while the bottom figure shows the location of the "point" sum $a+b$ with respect to the frame $(\Omega,(e_1,e_2,e_3))$.

A clean way to handle the problem of frame invariance and to deal with points in a more intrinsic manner is to make a clearer distinction between points and vectors. We duplicate $\mathbb{R}^3$ into two copies, the first copy corresponding to points, where we forget the vector space structure, and the second copy corresponding to free vectors, where the vector space structure is important. Furthermore, we make explicit the important fact that the vector space $\mathbb{R}^3$ acts on the set of points $\mathbb{R}^3$: Given any **point** $a=(a_1,a_2,a_3)$ and any **vector** $v=(v_1,v_2,v_3)$, we obtain the **point**

$$a+v=(a_1+v_1,a_2+v_2,a_3+v_3),$$

which can be thought of as the result of translating a to b using the vector v. We can imagine that v is placed such that its origin coincides with a and that its tip coincides with b. This action $+\colon \mathbb{R}^3\times\mathbb{R}^3\to\mathbb{R}^3$ satisfies some crucial properties. For example,

$$\begin{aligned} a+0 &= a,\\ (a+u)+v &= a+(u+v), \end{aligned}$$

and for any two points a, b, there is a unique free vector $\overrightarrow{ab}$ such that

$$b = a + \overrightarrow{ab}.$$

It turns out that the above properties, although trivial in the case of $\mathbb{R}^3$, are all that is needed to define the abstract notion of affine space (or affine structure). The basic idea is to consider two (distinct) sets E and $\overrightarrow{E}$, where E is a set of points (with no structure) and $\overrightarrow{E}$ is a vector space (of free vectors) acting on the set E.

Did you say "A fine space"?

Intuitively, we can think of the elements of $\overrightarrow{E}$ as forces moving the points in E, considered as physical particles. The effect of applying a force (free vector) $u \in \overrightarrow{E}$ to a point $a \in E$ is a translation. By this, we mean that for every force $u \in \overrightarrow{E}$, the action of the force u is to "move" every point $a \in E$ to the point $a + u \in E$ obtained by the translation corresponding to u viewed as a vector. Since translations can be composed, it is natural that $\overrightarrow{E}$ is a vector space.

For simplicity, it is assumed that all vector spaces under consideration are defined over the field $\mathbb{R}$ of real numbers. Most of the definitions and results also hold for an arbitrary field K, although some care is needed when dealing with fields of characteristic different from zero. It is also assumed that all families $(\lambda_i)_{i\in I}$ of scalars have finite support. Recall that a family $(\lambda_i)_{i\in I}$ of scalars has *finite support* if $\lambda_i = 0$ for all $i \in I - J$, where J is a finite subset of I. Obviously, finite families of scalars have finite support, and for simplicity, the reader may assume that all families of scalars are finite. The formal definition of an affine space is as follows.

Definition 24.1. An *affine space* is either the degenerate space reduced to the empty set, or a triple $\langle E, \overrightarrow{E}, +\rangle$ consisting of a nonempty set E (of *points*), a vector space $\overrightarrow{E}$ (of *translations*, or *free vectors*), and an action $+\colon E \times \overrightarrow{E} \to E$, satisfying the following conditions.

(A1) $a + 0 = a$, for every $a \in E$.

(A2) $(a + u) + v = a + (u + v)$, for every $a \in E$, and every $u, v \in \overrightarrow{E}$.

(A3) For any two points $a, b \in E$, there is a unique $u \in \overrightarrow{E}$ such that $a + u = b$.

The unique vector $u \in \overrightarrow{E}$ such that $a + u = b$ is denoted by $\overrightarrow{ab}$, or sometimes by $\mathbf{ab}$, or even by $b - a$. Thus, we also write

$$b = a + \overrightarrow{ab}$$

(or $b = a + \mathbf{ab}$, or even $b = a + (b - a)$).

The *dimension of the affine space* $\langle E, \overrightarrow{E}, +\rangle$ is the dimension $\dim(\overrightarrow{E})$ of the vector space $\overrightarrow{E}$. For simplicity, it is denoted by $\dim(E)$.

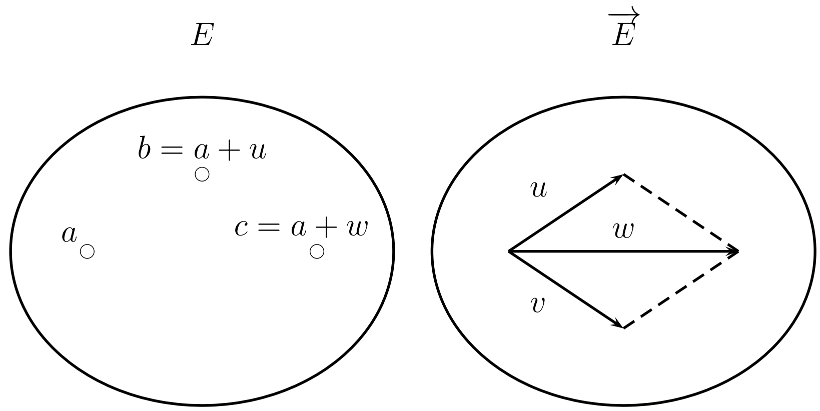


Figure 24.4: Intuitive picture of an affine space.

Conditions (A1) and (A2) say that the (abelian) group $\overrightarrow{E}$ acts on E, and Condition (A3) says that $\overrightarrow{E}$ acts transitively and faithfully on E. Note that

$$\overrightarrow{a(a+v)} = v$$

for all $a \in E$ and all $v \in \overrightarrow{E}$, since $\overrightarrow{a(a+v)}$ is the unique vector such that $a+v = a+\overrightarrow{a(a+v)}$. Thus, $b = a + v$ is equivalent to $\overrightarrow{ab} = v$. Figure 24.4 gives an intuitive picture of an affine space. It is natural to think of all vectors as having the same origin, the null vector.

The axioms defining an affine space $\langle E, \overrightarrow{E}, +\rangle$ can be interpreted intuitively as saying that E and $\overrightarrow{E}$ are two different ways of looking at the same object, but wearing different sets of glasses, the second set of glasses depending on the choice of an "origin" in E. Indeed, we can choose to look at the points in E, forgetting that every pair (a, b) of points defines a unique vector $\overrightarrow{ab}$ in $\overrightarrow{E}$, or we can choose to look at the vectors u in $\overrightarrow{E}$, forgetting the points in E. Furthermore, if we also pick any point a in E, a point that can be viewed as an *origin* in E, then we can recover all the points in E as the translated points $a + u$ for all $u \in \overrightarrow{E}$. This can be formalized by defining two maps between E and $\overrightarrow{E}$.

For every $a \in E$, consider the mapping from $\overrightarrow{E}$ to E given by

$$u \mapsto a + u,$$

where $u \in \overrightarrow{E}$, and consider the mapping from E to $\overrightarrow{E}$ given by

$$b \mapsto \overrightarrow{ab},$$

where $b \in E$. The composition of the first mapping with the second is

$$u \mapsto a + u \mapsto \overrightarrow{a(a+u)},$$

which, in view of (A3), yields u. The composition of the second with the first mapping is

$$b \mapsto \overrightarrow{ab} \mapsto a + \overrightarrow{ab},$$

which, in view of (A3), yields b. Thus, these compositions are the identity from $\overrightarrow{E}$ to $\overrightarrow{E}$ and the identity from E to E, and the mappings are both bijections.

When we identify E with $\overrightarrow{E}$ via the mapping $b \mapsto \overrightarrow{ab}$, we say that we consider E as the vector space obtained *by taking a as the origin in E*, and we denote it by E_a. Because E_a is a vector space, to be consistent with our notational conventions we should use the notation $\overrightarrow{E_a}$ (using an arrow), instead of E_a. However, for simplicity, we stick to the notation E_a.

Thus, an affine space $\langle E, \overrightarrow{E}, +\rangle$ is a way of defining a vector space structure on a set of points E, without making a commitment to a **fixed** origin in E. Nevertheless, as soon as we commit to an origin a in E, we can view E as the vector space E_a. However, we urge the reader to think of E as a physical set of points and of $\overrightarrow{E}$ as a set of forces acting on E, rather than reducing E to some isomorphic copy of $\mathbb{R}^n$. After all, points are points, and not vectors! For notational simplicity, we will often denote an affine space $\langle E, \overrightarrow{E}, +\rangle$ by $(E, \overrightarrow{E})$, or even by E. The vector space $\overrightarrow{E}$ is called the *vector space associated with E*.

One should be careful about the overloading of the addition symbol $+$. Addition is well-defined on vectors, as in $u + v$; the translate $a + u$ of a point $a \in E$ by a vector $u \in \overrightarrow{E}$ is also well-defined, but addition of points $a + b$ **does not make sense**. In this respect, the notation $b - a$ for the unique vector u such that $b = a + u$ is somewhat confusing, since it suggests that points can be subtracted (but not added!).

Any vector space $\overrightarrow{E}$ has an affine space structure specified by choosing $E = \overrightarrow{E}$, and letting $+$ be addition in the vector space $\overrightarrow{E}$. We will refer to the affine structure $\langle \overrightarrow{E}, \overrightarrow{E}, +\rangle$ on a vector space $\overrightarrow{E}$ as the *canonical (or natural) affine structure on $\overrightarrow{E}$*. In particular, the vector space $\mathbb{R}^n$ can be viewed as the affine space $\langle \mathbb{R}^n, \mathbb{R}^n, +\rangle$, denoted by $\mathbb{A}^n$. In general, if K is any field, the affine space $\langle K^n, K^n, +\rangle$ is denoted by $\mathbb{A}^n_K$. In order to distinguish between the double role played by members of $\mathbb{R}^n$, points and vectors, we will denote points by row vectors, and vectors by column vectors. Thus, the action of the vector space $\mathbb{R}^n$ over the set $\mathbb{R}^n$ simply viewed as a set of points is given by

$$(a_1, \ldots, a_n) + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = (a_1 + u_1, \ldots, a_n + u_n).$$

We will also use the convention that if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, then the column vector associated with x is denoted by $\mathbf{x}$ (in boldface notation). Abusing the notation slightly, if $a \in \mathbb{R}^n$ is a point, we also write $a \in \mathbb{A}^n$. The affine space $\mathbb{A}^n$ is called the *real affine space of dimension n*. In most cases, we will consider $n = 1, 2, 3$.

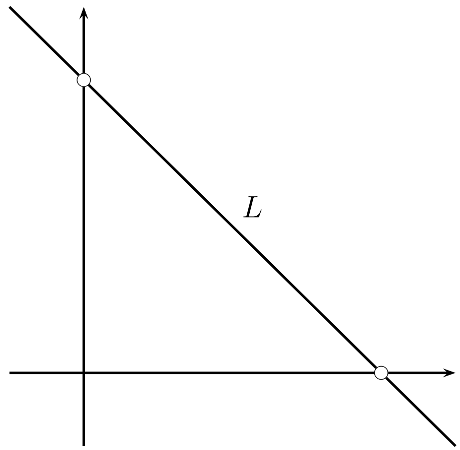


Figure 24.5: An affine space: the line of equation $x + y - 1 = 0$.

24.2 Examples of Affine Spaces

Let us now give an example of an affine space that is not given as a vector space (at least, not in an obvious fashion). Consider the subset L of $\mathbb{A}^2$ consisting of all points (x, y) satisfying the equation

$$x + y - 1 = 0.$$

The set L is the line of slope -1 passing through the points $(1, 0)$ and $(0, 1)$ shown in Figure 24.5.

The line L can be made into an official affine space by defining the action $+\colon L \times \mathbb{R} \to L$ of $\mathbb{R}$ on L defined such that for every point $(x, 1 - x)$ on L and any $u \in \mathbb{R}$,

$$(x, 1 - x) + u = (x + u, 1 - x - u).$$

It is immediately verified that this action makes L into an affine space. For example, for any two points $a = (a_1, 1 - a_1)$ and $b = (b_1, 1 - b_1)$ on L, the unique (vector) $u \in \mathbb{R}$ such that $b = a + u$ is $u = b_1 - a_1$. Note that the vector space $\mathbb{R}$ is isomorphic to the line of equation $x + y = 0$ passing through the origin.

Similarly, consider the subset H of $\mathbb{A}^3$ consisting of all points (x, y, z) satisfying the equation

$$x + y + z - 1 = 0.$$

The set H is the plane passing through the points $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. The plane H can be made into an official affine space by defining the action $+\colon H \times \mathbb{R}^2 \to H$ of $\mathbb{R}^2$ on

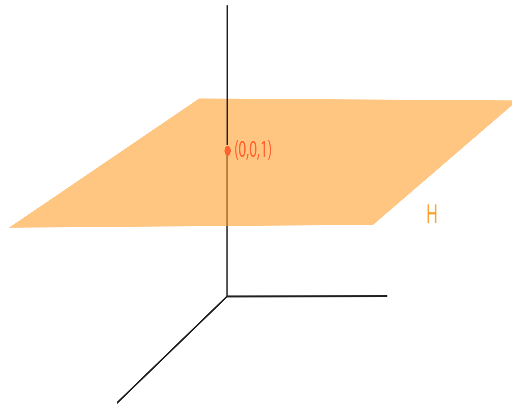


Figure 24.6: An affine space: the plane $x + y + z - 1 = 0$.

H defined such that for every point $(x, y, 1 - x - y)$ on H and any $\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2$,

$$(x, y, 1 - x - y) + \begin{pmatrix} u \\ v \end{pmatrix} = (x + u, y + v, 1 - x - u - y - v).$$

For a slightly wilder example, consider the subset P of $\mathbb{A}^3$ consisting of all points (x, y, z) satisfying the equation

$$x^2 + y^2 - z = 0.$$

The set P is a paraboloid of revolution, with axis Oz. The surface P can be made into an official affine space by defining the action $+\colon P \times \mathbb{R}^2 \to P$ of $\mathbb{R}^2$ on P defined such that for every point $(x, y, x^2 + y^2)$ on P and any $\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2$,

$$(x, y, x^2 + y^2) + \begin{pmatrix} u \\ v \end{pmatrix} = (x + u, y + v, (x + u)^2 + (y + v)^2).$$

See Figure 24.7.

This should dispel any idea that affine spaces are dull. Affine spaces not already equipped with an obvious vector space structure arise in projective geometry.

24.3 Chasles's Identity

Given any three points $a, b, c \in E$, since $c = a + \overrightarrow{ac}$, $b = a + \overrightarrow{ab}$, and $c = b + \overrightarrow{bc}$, we get

$$c = b + \overrightarrow{bc} = (a + \overrightarrow{ab}) + \overrightarrow{bc} = a + (\overrightarrow{ab} + \overrightarrow{bc})$$

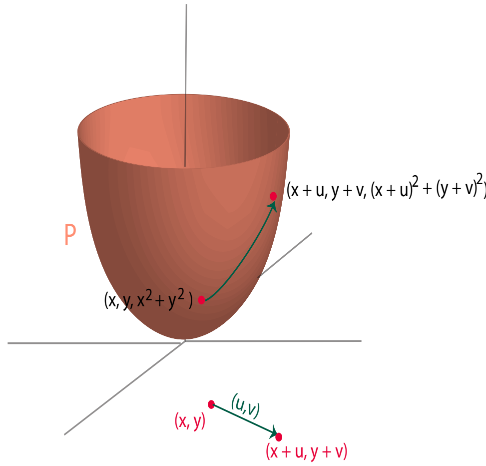


Figure 24.7: The paraboloid of revolution P viewed as a two-dimensional affine space.

by (A2), and thus, by (A3),

$$\overrightarrow{ab} + \overrightarrow{bc} = \overrightarrow{ac},$$

which is known as *Chasles's identity*, and illustrated in Figure 24.8.

Since $a = a + \overrightarrow{aa}$ and by (A1) $a = a + 0$, by (A3) we get

$$\overrightarrow{aa} = 0.$$

Thus, letting $a = c$ in Chasles's identity, we get

$$\overrightarrow{ba} = -\overrightarrow{ab}.$$

Given any four points $a, b, c, d \in E$, since by Chasles's identity

$$\overrightarrow{ab} + \overrightarrow{bc} = \overrightarrow{ad} + \overrightarrow{dc} = \overrightarrow{ac},$$

we have the *parallelogram law*

$$\overrightarrow{ab} = \overrightarrow{dc} \quad \text{iff} \quad \overrightarrow{bc} = \overrightarrow{ad}.$$

24.4 Affine Combinations, Barycenters

A fundamental concept in linear algebra is that of a linear combination. The corresponding concept in affine geometry is that of an *affine combination*, also called a *barycenter*. However, there is a problem with the naive approach involving a coordinate system, as we saw in Section 24.1. Since this problem is the reason for introducing affine combinations, at the

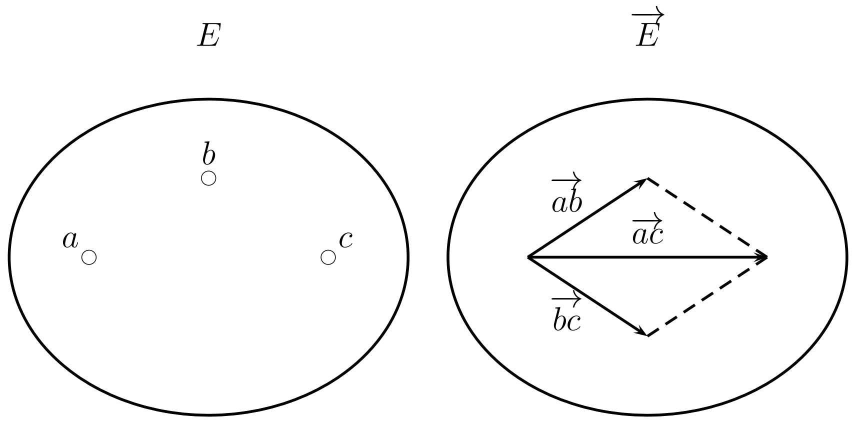


Figure 24.8: Points and corresponding vectors in affine geometry.

risk of boring certain readers, we give another example showing what goes wrong if we are not careful in defining linear combinations of points.

Consider $\mathbb{R}^2$ as an affine space, under its natural coordinate system with origin $O = (0, 0)$ and basis vectors $\begin{pmatrix}1\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$. Given any two points $a = (a_1, a_2)$ and $b = (b_1, b_2)$, it is natural to define the affine combination $\lambda a + \mu b$ as the point of coordinates

$$(\lambda a_1 + \mu b_1, \lambda a_2 + \mu b_2).$$

Thus, when $a = (-1, -1)$ and $b = (2, 2)$, the point $a + b$ is the point $c = (1, 1)$.

Let us now consider the new coordinate system with respect to the origin $c = (1, 1)$ (and the same basis vectors). This time, the coordinates of a are $(-2, -2)$, the coordinates of b are $(1, 1)$, and the point $a + b$ is the point d of coordinates $(-1, -1)$. However, it is clear that the point d is identical to the origin $O = (0, 0)$ of the first coordinate system. This situation is illustrated in Figure 24.9.

Thus, $a + b$ corresponds to two different points depending on which coordinate system is used for its computation!

This shows that some extra condition is needed in order for affine combinations to make sense. It turns out that if the scalars sum up to 1, the definition is intrinsic, as the following proposition shows.

Proposition 24.1. *Given an affine space E, let $(a_i)_{i\in I}$ be a family of points in E, and let $(\lambda_i)_{i\in I}$ be a family of scalars. For any two points $a, b \in E$, the following properties hold:*

(1) If $\sum_{i\in I} \lambda_i = 1$, then

$$a + \sum_{i\in I} \lambda_i \overrightarrow{aa_i} = b + \sum_{i\in I} \lambda_i \overrightarrow{ba_i}.$$

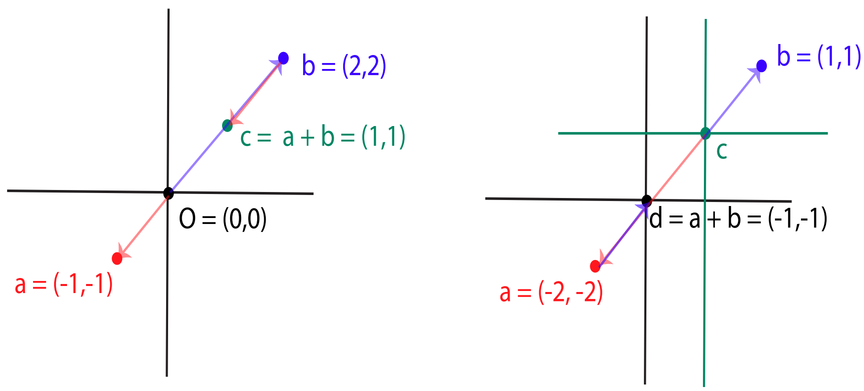


Figure 24.9: The example from the beginning of Section 24.4.

(2) If $\sum_{i\in I}\lambda_i = 0$, *then*

$$\sum_{i\in I}\lambda_i\overrightarrow{aa_i} = \sum_{i\in I}\lambda_i\overrightarrow{ba_i}.$$

Proof. (1) By Chasles's identity (see Section 24.3), we have

$$\begin{aligned}
a + \sum_{i\in I}\lambda_i\overrightarrow{aa_i} &= a + \sum_{i\in I}\lambda_i(\overrightarrow{ab} + \overrightarrow{ba_i}) \\
&= a + \left(\sum_{i\in I}\lambda_i\right)\overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} \\
&= a + \overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} && \text{since } \textstyle\sum_{i\in I}\lambda_i = 1 \\
&= b + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} && \text{since } b = a + \overrightarrow{ab}.
\end{aligned}$$

An illustration of this calculation in $\mathbb{A}^2$ is provided by Figure 24.10.

(2) We also have

$$\begin{aligned}
\sum_{i\in I}\lambda_i\overrightarrow{aa_i} &= \sum_{i\in I}\lambda_i(\overrightarrow{ab} + \overrightarrow{ba_i}) \\
&= \left(\sum_{i\in I}\lambda_i\right)\overrightarrow{ab} + \sum_{i\in I}\lambda_i\overrightarrow{ba_i} \\
&= \sum_{i\in I}\lambda_i\overrightarrow{ba_i},
\end{aligned}$$

since $\sum_{i\in I}\lambda_i = 0$. $\square$

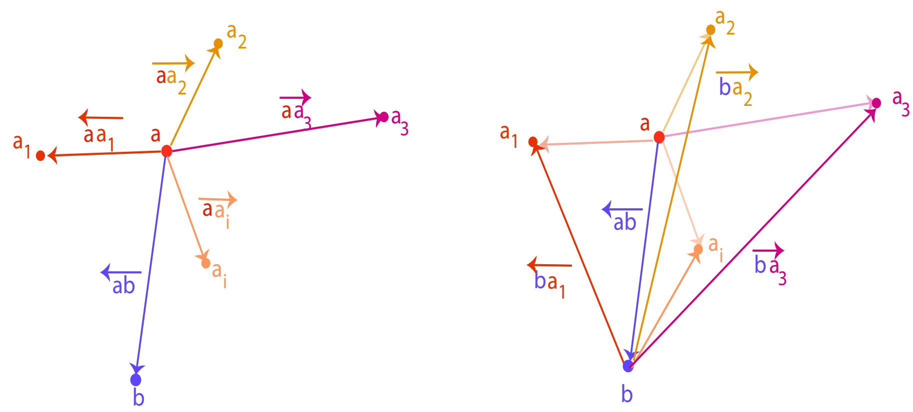


Figure 24.10: Part (1) of Proposition 24.1.

Thus, by Proposition 24.1, for any family of points $(a_i)_{i\in I}$ in E, for any family $(\lambda_i)_{i\in I}$ of scalars such that $\sum_{i\in I}\lambda_i = 1$, the point

$$x = a + \sum_{i\in I}\lambda_i\overrightarrow{aa_i}$$

is independent of the choice of the origin $a \in E$. This property motivates the following definition.

Definition 24.2. For any family of points $(a_i)_{i\in I}$ in E, for any family $(\lambda_i)_{i\in I}$ of scalars such that $\sum_{i\in I}\lambda_i = 1$, and for any $a \in E$, the point

$$a + \sum_{i\in I}\lambda_i\overrightarrow{aa_i}$$

(which is independent of $a \in E$, by Proposition 24.1) is called the *barycenter (or barycentric combination, or affine combination) of the points a_i assigned the weights λ_i*, and it is denoted by

$$\sum_{i\in I}\lambda_i a_i.$$

In dealing with barycenters, it is convenient to introduce the notion of a *weighted point*, which is just a pair (a, λ), where $a \in E$ is a point, and $\lambda \in \mathbb{R}$ is a scalar. Then, given a family of weighted points $((a_i, \lambda_i))_{i\in I}$, where $\sum_{i\in I}\lambda_i = 1$, we also say that the point $\sum_{i\in I}\lambda_i a_i$ is the *barycenter of the family of weighted points* $((a_i, \lambda_i))_{i\in I}$.

Note that the barycenter x of the family of weighted points $((a_i, \lambda_i))_{i\in I}$ is the unique point such that

$$\overrightarrow{ax} = \sum_{i\in I}\lambda_i\overrightarrow{aa_i} \quad \text{for every } a \in E,$$

and setting $a = x$, the point x is the unique point such that

$$\sum_{i\in I} \lambda_i \overrightarrow{xa_i} = 0.$$

In physical terms, the barycenter is the *center of mass* of the family of weighted points $((a_i, \lambda_i))_{i\in I}$ (where the masses have been normalized, so that $\sum_{i\in I} \lambda_i = 1$, and negative masses are allowed).

Remarks:

(1) Since the barycenter of a family $((a_i, \lambda_i))_{i\in I}$ of weighted points is defined for families $(\lambda_i)_{i\in I}$ of scalars with finite support (and such that $\sum_{i\in I} \lambda_i = 1$), we might as well assume that I is finite. Then, for all $m \geq 2$, it is easy to prove that the barycenter of m weighted points can be obtained by repeated computations of barycenters of two weighted points.

(2) This result still holds, provided that the field K has at least three distinct elements, but the proof is trickier!

(3) When $\sum_{i\in I} \lambda_i = 0$, the vector $\sum_{i\in I} \lambda_i \overrightarrow{aa_i}$ does not depend on the point a, and we may denote it by $\sum_{i\in I} \lambda_i a_i$. This observation will be used to define a vector space in which linear combinations of both points and vectors make sense, regardless of the value of $\sum_{i\in I} \lambda_i$.

Figure 24.11 illustrates the geometric construction of the barycenters g_1 and g_2 of the weighted points $\left(a, \frac{1}{4}\right)$, $\left(b, \frac{1}{4}\right)$, and $\left(c, \frac{1}{2}\right)$, and $(a, -1)$, $(b, 1)$, and $(c, 1)$.

The point g_1 can be constructed geometrically as the middle of the segment joining c to the middle $\frac{1}{2}a + \frac{1}{2}b$ of the segment (a, b), since

$$g_1 = \frac{1}{2}\left(\frac{1}{2}a + \frac{1}{2}b\right) + \frac{1}{2}c.$$

The point g_2 can be constructed geometrically as the point such that the middle $\frac{1}{2}b + \frac{1}{2}c$ of the segment (b, c) is the middle of the segment (a, g_2), since

$$g_2 = -a + 2\left(\frac{1}{2}b + \frac{1}{2}c\right).$$

Later on, we will see that a polynomial curve can be defined as a set of barycenters of a fixed number of points. For example, let (a, b, c, d) be a sequence of points in $\mathbb{A}^2$. Observe that

$$(1-t)^3 + 3t(1-t)^2 + 3t^2(1-t) + t^3 = 1,$$

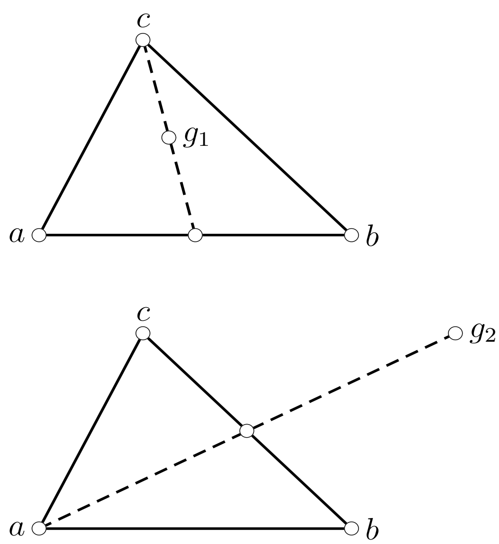


Figure 24.11: Barycenters, $g_1 = \frac{1}{4}a + \frac{1}{4}b + \frac{1}{2}c$, $\quad g_2 = -a + b + c$

since the sum on the left-hand side is obtained by expanding $(t + (1 - t))^3 = 1$ using the binomial formula. Thus,

$$(1 - t)^3\, a + 3t(1 - t)^2\, b + 3t^2(1 - t)\, c + t^3\, d$$

is a well-defined affine combination. Then, we can define the curve $F\colon \mathbb{A} \to \mathbb{A}^2$ such that

$$F(t) = (1 - t)^3\, a + 3t(1 - t)^2\, b + 3t^2(1 - t)\, c + t^3\, d.$$

Such a curve is called a *Bézier curve*, and (a, b, c, d) are called its *control points*. Note that the curve passes through a and d, but generally not through b and c. It can be sbown that any point $F(t)$ on the curve can be constructed using an algorithm performing affine interpolation steps (the *de Casteljau algorithm*).

24.5 Affine Subspaces

In linear algebra, a (linear) subspace can be characterized as a nonempty subset of a vector space closed under linear combinations. In affine spaces, the notion corresponding to the notion of (linear) subspace is the notion of affine subspace. It is natural to define an affine subspace as a subset of an affine space closed under affine combinations.

Definition 24.3. Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, a subset V of E is an *affine subspace* (*of* $\langle E, \overrightarrow{E}, +\rangle$) if for every family of weighted points $((a_i, \lambda_i))_{i\in I}$ in V such that $\sum_{i\in I} \lambda_i = 1$, the barycenter $\sum_{i\in I} \lambda_i a_i$ belongs to V.

An affine subspace is also called a *flat* by some authors. According to Definition 24.3, the empty set is trivially an affine subspace, and every intersection of affine subspaces is an affine subspace.

As an example, consider the subset U of $\mathbb{R}^2$ defined by

$$U = \left\{(x, y) \in \mathbb{R}^2 \mid ax + by = c\right\},$$

i.e., the set of solutions of the equation

$$ax + by = c,$$

where it is assumed that $a \neq 0$ or $b \neq 0$. Given any m points $(x_i, y_i) \in U$ and any m scalars λ_i such that $\lambda_1 + \cdots + \lambda_m = 1$, we claim that

$$\sum_{i=1}^{m} \lambda_i (x_i, y_i) \in U.$$

Indeed, $(x_i, y_i) \in U$ means that

$$ax_i + by_i = c,$$

and if we multiply both sides of this equation by λ_i and add up the resulting m equations, we get

$$\sum_{i=1}^{m} (\lambda_i a x_i + \lambda_i b y_i) = \sum_{i=1}^{m} \lambda_i c,$$

and since $\lambda_1 + \cdots + \lambda_m = 1$, we get

$$a\left(\sum_{i=1}^{m} \lambda_i x_i\right) + b\left(\sum_{i=1}^{m} \lambda_i y_i\right) = \left(\sum_{i=1}^{m} \lambda_i\right) c = c,$$

which shows that

$$\left(\sum_{i=1}^{m} \lambda_i x_i, \sum_{i=1}^{m} \lambda_i y_i\right) = \sum_{i=1}^{m} \lambda_i (x_i, y_i) \in U.$$

Thus, U is an affine subspace of $\mathbb{A}^2$. In fact, it is just a usual line in $\mathbb{A}^2$.

It turns out that U is closely related to the subset of $\mathbb{R}^2$ defined by

$$\overrightarrow{U} = \left\{(x, y) \in \mathbb{R}^2 \mid ax + by = 0\right\},$$

i.e., the set of solutions of the homogeneous equation

$$ax + by = 0$$

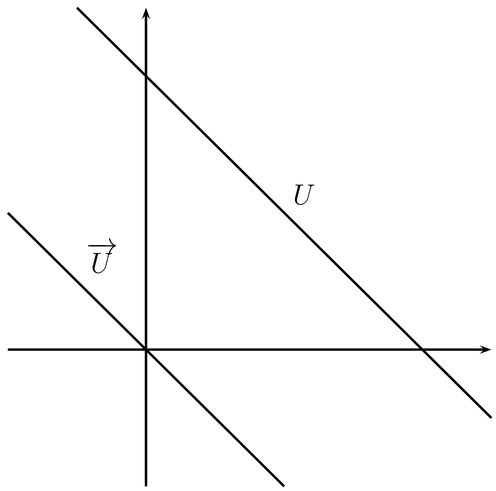


Figure 24.12: An affine line U and its direction.

obtained by setting the right-hand side of $ax + by = c$ to zero. Indeed, for any m scalars λ_i, the same calculation as above yields that

$$\sum_{i=1}^{m} \lambda_i(x_i, y_i) \in \overrightarrow{U},$$

this time **without any restriction on the** λ_i, since the right-hand side of the equation is null. Thus, $\overrightarrow{U}$ is a subspace of $\mathbb{R}^2$. In fact, $\overrightarrow{U}$ is one-dimensional, and it is just a usual line in $\mathbb{R}^2$. This line can be identified with a line passing through the origin of $\mathbb{A}^2$, a line that is parallel to the line U of equation $ax + by = c$, as illustrated in Figure 24.12.

Now, if (x_0, y_0) is any point in U, we claim that

$$U = (x_0, y_0) + \overrightarrow{U},$$

where

$$(x_0, y_0) + \overrightarrow{U} = \left\{ (x_0 + u_1, y_0 + u_2) \mid (u_1, u_2) \in \overrightarrow{U} \right\}.$$

First, $(x_0, y_0) + \overrightarrow{U} \subseteq U$, since $ax_0 + by_0 = c$ and $au_1 + bu_2 = 0$ for all $(u_1, u_2) \in \overrightarrow{U}$. Second, if $(x, y) \in U$, then $ax + by = c$, and since we also have $ax_0 + by_0 = c$, by subtraction, we get

$$a(x - x_0) + b(y - y_0) = 0,$$

which shows that $(x - x_0, y - y_0) \in \overrightarrow{U}$, and thus $(x, y) \in (x_0, y_0) + \overrightarrow{U}$. Hence, we also have $U \subseteq (x_0, y_0) + \overrightarrow{U}$, and $U = (x_0, y_0) + \overrightarrow{U}$.

The above example shows that the affine line U defined by the equation

$$ax + by = c$$

is obtained by "translating" the parallel line $\overrightarrow{U}$ of equation

$$ax + by = 0$$

passing through the origin. In fact, given any point $(x_0, y_0) \in U$,

$$U = (x_0, y_0) + \overrightarrow{U}.$$

More generally, it is easy to prove the following fact. Given any $m \times n$ matrix A and any vector $b \in \mathbb{R}^m$, the subset U of $\mathbb{R}^n$ defined by

$$U = \{x \in \mathbb{R}^n \mid Ax = b\}$$

is an affine subspace of $\mathbb{A}^n$.

Actually, observe that $Ax = b$ should really be written as $Ax^\top = b$, to be consistent with our convention that points are represented by row vectors. We can also use the boldface notation for column vectors, in which case the equation is written as $A\mathbf{x} = b$. For the sake of minimizing the amount of notation, we stick to the simpler (yet incorrect) notation $Ax = b$. If we consider the corresponding homogeneous equation $Ax = 0$, the set

$$\overrightarrow{U} = \{x \in \mathbb{R}^n \mid Ax = 0\}$$

is a subspace of $\mathbb{R}^n$, and for any $x_0 \in U$, we have

$$U = x_0 + \overrightarrow{U}.$$

This is a general situation. Affine subspaces can be characterized in terms of subspaces of $\overrightarrow{E}$. Let V be a nonempty subset of E. For every family $(a_1, \ldots, a_n)$ in V, for any family $(\lambda_1, \ldots, \lambda_n)$ of scalars, and for every point $a \in V$, observe that $x \in E$ given by

$$x = a + \sum_{i=1}^{n} \lambda_i \overrightarrow{aa_i}$$

is the barycenter of the family of weighted points

$$\left((a_1, \lambda_1), \ldots, (a_n, \lambda_n), \left(a, 1 - \sum_{i=1}^{n} \lambda_i\right)\right),$$

since

$$\sum_{i=1}^{n} \lambda_i + \left(1 - \sum_{i=1}^{n} \lambda_i\right) = 1.$$

Given any point $a \in E$ and any subset $\overrightarrow{V}$ of $\overrightarrow{E}$, let $a + \overrightarrow{V}$ denote the following subset of E:

$$a + \overrightarrow{V} = \left\{a + v \mid v \in \overrightarrow{V}\right\}.$$

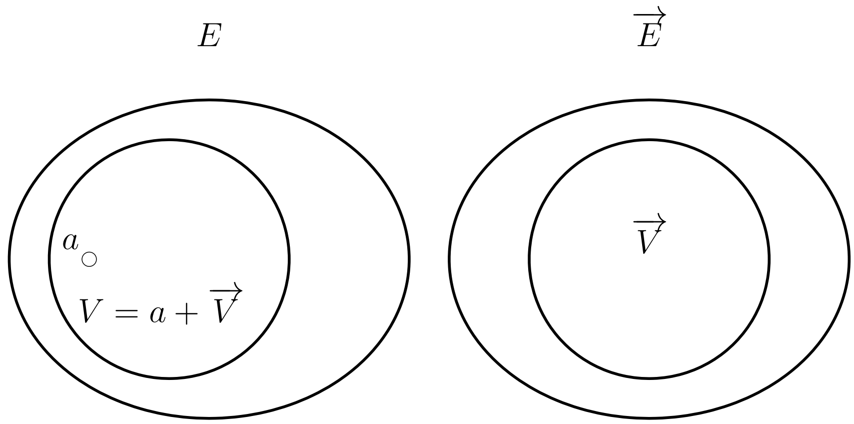


Figure 24.13: An affine subspace V and its direction $\overrightarrow{V}$.

Proposition 24.2. *Let $\langle E, \overrightarrow{E}, +\rangle$ be an affine space.*

(1) A nonempty subset V of E is an affine subspace iff for every point $a \in V$, the set

$$\overrightarrow{V_a} = \{\overrightarrow{ax} \mid x \in V\}$$

is a subspace of $\overrightarrow{E}$. Consequently, $V = a + \overrightarrow{V_a}$. Furthermore,

$$\overrightarrow{V} = \{\overrightarrow{xy} \mid x, y \in V\}$$

is a subspace of $\overrightarrow{E}$ and $\overrightarrow{V_a} = \overrightarrow{V}$ for all $a \in E$. Thus, $V = a + \overrightarrow{V}$.

(2) For any subspace $\overrightarrow{V}$ of $\overrightarrow{E}$ and for any $a \in E$, the set $V = a + \overrightarrow{V}$ is an affine subspace.

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [70]. $\square$

In particular, when E is the natural affine space associated with a vector space $\overrightarrow{E}$, Proposition 24.2 shows that every affine subspace of E is of the form $u + \overrightarrow{U}$, for a subspace $\overrightarrow{U}$ of $\overrightarrow{E}$. The subspaces of $\overrightarrow{E}$ are the affine subspaces of E that contain 0.

The subspace $\overrightarrow{V}$ associated with an affine subspace V is called the *direction of* V. It is also clear that the map $+: V \times \overrightarrow{V} \to V$ induced by $+: E \times \overrightarrow{E} \to E$ confers to $\langle V, \overrightarrow{V}, +\rangle$ an affine structure. Figure 24.13 illustrates the notion of affine subspace.

By the dimension of the subspace V, we mean the dimension of $\overrightarrow{V}$.

An affine subspace of dimension 1 is called a *line*, and an affine subspace of dimension 2 is called a *plane*.

An affine subspace of codimension 1 is called a *hyperplane* (recall that a subspace F of a vector space E has codimension 1 iff there is some subspace G of dimension 1 such that $E = F \oplus G$, the direct sum of F and G, see Strang [170] or Lang [109]).

We say that two affine subspaces U and V are *parallel* if their directions are identical. Equivalently, since $\overrightarrow{U} = \overrightarrow{V}$, we have $U = a + \overrightarrow{U}$ and $V = b + \overrightarrow{U}$ for any $a \in U$ and any $b \in V$, and thus V is obtained from U by the translation $\overrightarrow{ab}$.

In general, when we talk about n points $a_1, \ldots, a_n$, we mean the sequence $(a_1, \ldots, a_n)$, and not the set $\{a_1, \ldots, a_n\}$ (the a_i's need not be distinct).

By Proposition 24.2, a line is specified by a point $a \in E$ and a nonzero vector $v \in \overrightarrow{E}$, i.e., a line is the set of all points of the form $a + \lambda v$, for $\lambda \in \mathbb{R}$.

We say that three points a, b, c are *collinear* if the vectors $\overrightarrow{ab}$ and $\overrightarrow{ac}$ are linearly dependent. If two of the points a, b, c are distinct, say $a \neq b$, then there is a unique $\lambda \in \mathbb{R}$ such that $\overrightarrow{ac} = \lambda \overrightarrow{ab}$, and we define the ratio $\frac{\overrightarrow{ac}}{\overrightarrow{ab}} = \lambda$.

A plane is specified by a point $a \in E$ and two linearly independent vectors $u, v \in \overrightarrow{E}$, i.e., a plane is the set of all points of the form $a + \lambda u + \mu v$, for $\lambda, \mu \in \mathbb{R}$.

We say that four points a, b, c, d are *coplanar* if the vectors $\overrightarrow{ab}, \overrightarrow{ac}$, and $\overrightarrow{ad}$ are linearly dependent. Hyperplanes will be characterized a little later.

Proposition 24.3. *Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, for any family $(a_i)_{i \in I}$ of points in E, the set V of barycenters $\sum_{i \in I} \lambda_i a_i$ (where $\sum_{i \in I} \lambda_i = 1$) is the smallest affine subspace containing $(a_i)_{i \in I}$.*

Proof. If $(a_i)_{i \in I}$ is empty, then $V = \emptyset$, because of the condition $\sum_{i \in I} \lambda_i = 1$. If $(a_i)_{i \in I}$ is nonempty, then the smallest affine subspace containing $(a_i)_{i \in I}$ must contain the set V of barycenters $\sum_{i \in I} \lambda_i a_i$, and thus, it is enough to show that V is closed under affine combinations, which is immediately verified. $\square$

Given a nonempty subset S of E, the smallest affine subspace of E generated by S is often denoted by $\langle S \rangle$. For example, a line specified by two distinct points a and b is denoted by $\langle a, b\rangle$, or even (a, b), and similarly for planes, etc.

Remarks:

(1) Since it can be shown that the barycenter of n weighted points can be obtained by repeated computations of barycenters of two weighted points, a nonempty subset V of E is an affine subspace iff for every two points $a, b \in V$, the set V contains all barycentric combinations of a and b. If V contains at least two points, then V is an affine subspace iff for any two distinct points $a, b \in V$, the set V contains the line determined by a and b, that is, the set of all points $(1 - \lambda)a + \lambda b$, $\lambda \in \mathbb{R}$.

(2) This result still holds if the field K has at least three distinct elements, but the proof is trickier!

24.6 Affine Independence and Affine Frames

Corresponding to the notion of linear independence in vector spaces, we have the notion of affine independence. Given a family $(a_i)_{i\in I}$ of points in an affine space E, we will reduce the notion of (affine) independence of these points to the (linear) independence of the families $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ of vectors obtained by choosing any a_i as an origin. First, the following proposition shows that it is sufficient to consider only one of these families.

Proposition 24.4. *Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, let $(a_i)_{i\in I}$ be a family of points in E. If the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly independent for some $i \in I$, then $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly independent for every $i \in I$.*

Proof. Assume that the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly independent for some specific $i \in I$. Let $k \in I$ with $k \neq i$, and assume that there are some scalars $(\lambda_j)_{j\in(I-\{k\})}$ such that

$$\sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_ka_j} = 0.$$

Since

$$\overrightarrow{a_ka_j} = \overrightarrow{a_ka_i} + \overrightarrow{a_ia_j},$$

we have

$$\begin{aligned}
\sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_ka_j} &= \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_ka_i} + \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_ia_j}, \\
&= \sum_{j\in(I-\{k\})} \lambda_j \overrightarrow{a_ka_i} + \sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_ia_j}, \\
&= \sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_ia_j} - \Bigg(\sum_{j\in(I-\{k\})} \lambda_j\Bigg) \overrightarrow{a_ia_k},
\end{aligned}$$

and thus

$$\sum_{j\in(I-\{i,k\})} \lambda_j \overrightarrow{a_ia_j} - \Bigg(\sum_{j\in(I-\{k\})} \lambda_j\Bigg) \overrightarrow{a_ia_k} = 0.$$

Since the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly independent, we must have $\lambda_j = 0$ for all $j \in (I - \{i, k\})$ and $\sum_{j\in(I-\{k\})} \lambda_j = 0$, which implies that $\lambda_j = 0$ for all $j \in (I - \{k\})$. ☐

We define affine independence as follows.

Definition 24.4. Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, a family $(a_i)_{i\in I}$ of points in E is *affinely independent* if the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly independent for some $i \in I$.

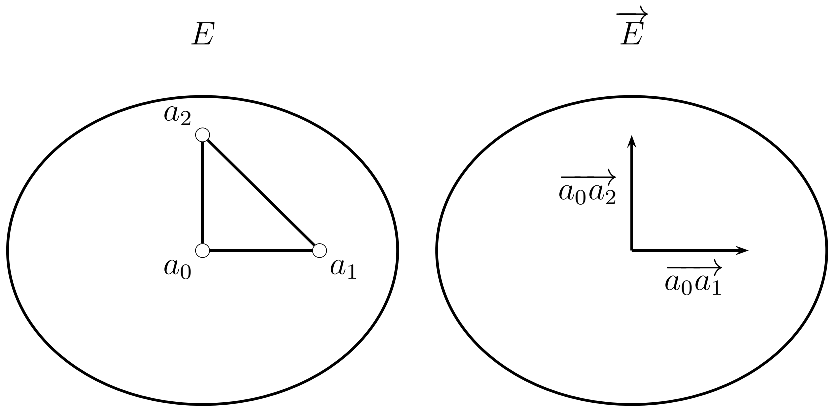


Figure 24.14: Affine independence and linear independence

Definition 24.4 is reasonable, because by Proposition 24.4, the independence of the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ does not depend on the choice of a_i. A crucial property of linearly independent vectors $(u_1, \ldots, u_m)$ is that if a vector v is a linear combination

$$v = \sum_{i=1}^{m} \lambda_i u_i$$

of the u_i, then the λ_i are unique. A similar result holds for affinely independent points.

Proposition 24.5. *Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, let $(a_0, \ldots, a_m)$ be a family of $m+1$ points in E. Let $x \in E$, and assume that $x = \sum_{i=0}^{m} \lambda_i a_i$, where $\sum_{i=0}^{m} \lambda_i = 1$. Then, the family $(\lambda_0, \ldots, \lambda_m)$ such that $x = \sum_{i=0}^{m} \lambda_i a_i$ is unique iff the family $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is linearly independent.*

Proof. The proof is straightforward and is omitted. It is also given in Gallier [70]. $\square$

Proposition 24.5 suggests the notion of affine frame. Affine frames are the affine analogues of bases in vector spaces. Let $\langle E, \overrightarrow{E}, +\rangle$ be a nonempty affine space, and let $(a_0, \ldots, a_m)$ be a family of $m+1$ points in E. The family $(a_0, \ldots, a_m)$ determines the family of m vectors $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ in $\overrightarrow{E}$. Conversely, given a point a_0 in E and a family of m vectors $(u_1, \ldots, u_m)$ in $\overrightarrow{E}$, we obtain the family of $m+1$ points $(a_0, \ldots, a_m)$ in E, where $a_i = a_0 + u_i$, $1 \leq i \leq m$.

Thus, for any $m \geq 1$, it is equivalent to consider a family of $m+1$ points $(a_0, \ldots, a_m)$ in E, and a pair $(a_0, (u_1, \ldots, u_m))$, where the u_i are vectors in $\overrightarrow{E}$. Figure 24.14 illustrates the notion of affine independence.

Remark: The above observation also applies to infinite families $(a_i)_{i\in I}$ of points in E and families $(u_i)_{i\in I-\{0\}}$ of vectors in $\overrightarrow{E}$, provided that the index set I contains 0.

When $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is a basis of $\overrightarrow{E}$ then, for every $x \in E$, since $x = a_0 + \overrightarrow{a_0x}$, there is a unique family $(x_1, \ldots, x_m)$ of scalars such that

$$x = a_0 + x_1\overrightarrow{a_0a_1} + \cdots + x_m\overrightarrow{a_0a_m}.$$

The scalars $(x_1, \ldots, x_m)$ may be considered as coordinates with respect to $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$. Since

$$x = a_0 + \sum_{i=1}^{m} x_i\overrightarrow{a_0a_i} \quad \text{iff} \quad x = \left(1 - \sum_{i=1}^{m} x_i\right) a_0 + \sum_{i=1}^{m} x_i a_i,$$

$x \in E$ can also be expressed uniquely as

$$x = \sum_{i=0}^{m} \lambda_i a_i$$

with $\sum_{i=0}^{m} \lambda_i = 1$, and where $\lambda_0 = 1 - \sum_{i=1}^{m} x_i$, and $\lambda_i = x_i$ for $1 \leq i \leq m$. The scalars $(\lambda_0, \ldots, \lambda_m)$ are also certain kinds of coordinates with respect to $(a_0, \ldots, a_m)$. All this is summarized in the following definition.

Definition 24.5. Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, an *affine frame with origin* a_0 is a family $(a_0, \ldots, a_m)$ of $m + 1$ points in E such that the list of vectors $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is a basis of $\overrightarrow{E}$. The pair $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ is also called an *affine frame with origin* a_0. Then, every $x \in E$ can be expressed as

$$x = a_0 + x_1\overrightarrow{a_0a_1} + \cdots + x_m\overrightarrow{a_0a_m}$$

for a unique family $(x_1, \ldots, x_m)$ of scalars, called the *coordinates of x w.r.t. the affine frame* $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$. Furthermore, every $x \in E$ can be written as

$$x = \lambda_0 a_0 + \cdots + \lambda_m a_m$$

for some unique family $(\lambda_0, \ldots, \lambda_m)$ of scalars such that $\lambda_0 + \cdots + \lambda_m = 1$ called the *barycentric coordinates of x with respect to the affine frame* $(a_0, \ldots, a_m)$. See Figure 24.15.

The coordinates $(x_1, \ldots, x_m)$ and the barycentric coordinates $(\lambda_0, \ldots, \lambda_m)$ are related by the equations $\lambda_0 = 1 - \sum_{i=1}^{m} x_i$ and $\lambda_i = x_i$, for $1 \leq i \leq m$. An affine frame is called an *affine basis* by some authors. A family $(a_i)_{i\in I}$ of points in E is *affinely dependent* if it is not affinely independent. We can also characterize affinely dependent families as follows.

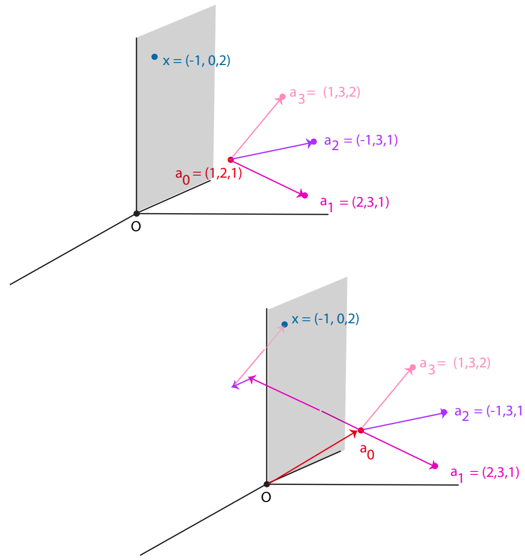


Figure 24.15: The affine frame (a_0, a_1, a_2, a_3) for $\mathbb{A}^3$. The coordinates for $x = (-1, 0, 2)$ are $x_1 = -8/3$, $x_2 = -1/3$, $x_3 = 1$, while the barycentric coordinates for x are $\lambda_0 = 3$, $\lambda_1 = -8/3$, $\lambda_2 = -1/3$, $\lambda_3 = 1$.

Proposition 24.6. *Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, let $(a_i)_{i\in I}$ be a family of points in E. The family $(a_i)_{i\in I}$ is affinely dependent iff there is a family $(\lambda_i)_{i\in I}$ such that $\lambda_j \neq 0$ for some $j \in I$, $\sum_{i\in I} \lambda_i = 0$, and $\sum_{i\in I} \lambda_i \overrightarrow{xa_i} = 0$ for every $x \in E$.*

Proof. By Proposition 24.5, the family $(a_i)_{i\in I}$ is affinely dependent iff the family of vectors $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly dependent for some $i \in I$. For any $i \in I$, the family $(\overrightarrow{a_ia_j})_{j\in(I-\{i\})}$ is linearly dependent iff there is a family $(\lambda_j)_{j\in(I-\{i\})}$ such that $\lambda_j \neq 0$ for some j, and such that

$$\sum_{j\in(I-\{i\})} \lambda_j \overrightarrow{a_ia_j} = 0.$$

Then, for any $x \in E$, we have

$$\begin{aligned}\sum_{j\in(I-\{i\})} \lambda_j \overrightarrow{a_ia_j} &= \sum_{j\in(I-\{i\})} \lambda_j (\overrightarrow{xa_j} - \overrightarrow{xa_i}) \\ &= \sum_{j\in(I-\{i\})} \lambda_j \overrightarrow{xa_j} - \Bigg(\sum_{j\in(I-\{i\})} \lambda_j\Bigg) \overrightarrow{xa_i},\end{aligned}$$

and letting $\lambda_i = -\left(\sum_{j\in(I-\{i\})} \lambda_j\right)$, we get $\sum_{i\in I} \lambda_i \overrightarrow{xa_i} = 0$, with $\sum_{i\in I} \lambda_i = 0$ and $\lambda_j \neq 0$ for some $j \in I$. The converse is obvious by setting $x = a_i$ for some i such that $\lambda_i \neq 0$, since $\sum_{i\in I} \lambda_i = 0$ implies that $\lambda_j \neq 0$, for some $j \neq i$. $\square$

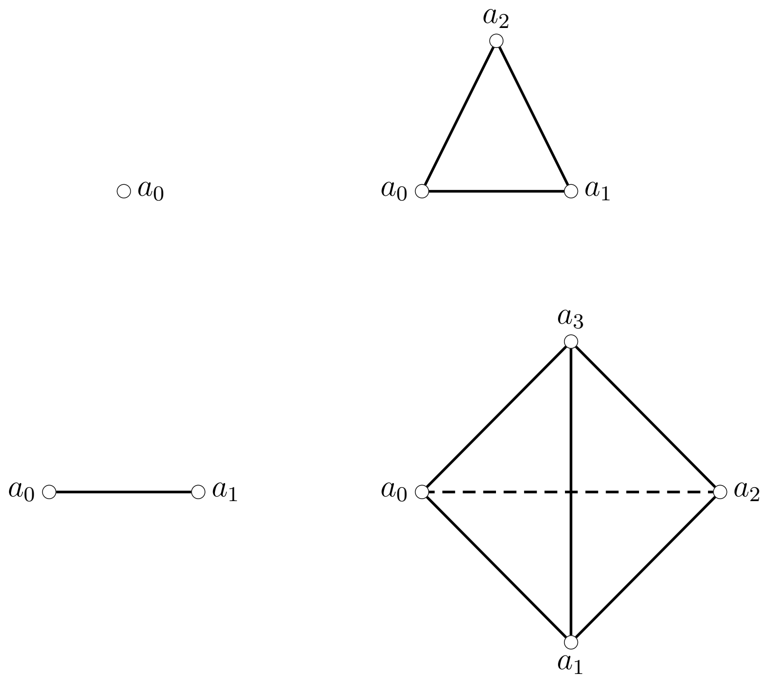


Figure 24.16: Examples of affine frames and their convex hulls.

Even though Proposition 24.6 is rather dull, it is one of the key ingredients in the proof of beautiful and deep theorems about convex sets, such as Carathéodory's theorem, Radon's theorem, and Helly's theorem.

A family of two points (a, b) in E is affinely independent iff $\overrightarrow{ab} \neq 0$, iff $a \neq b$. If $a \neq b$, the affine subspace generated by a and b is the set of all points $(1-\lambda)a + \lambda b$, which is the unique line passing through a and b. A family of three points (a, b, c) in E is affinely independent iff $\overrightarrow{ab}$ and $\overrightarrow{ac}$ are linearly independent, which means that a, b, and c are not on the same line (they are not collinear). In this case, the affine subspace generated by (a, b, c) is the set of all points $(1 - \lambda - \mu)a + \lambda b + \mu c$, which is the unique plane containing a, b, and c. A family of four points (a, b, c, d) in E is affinely independent iff $\overrightarrow{ab}$, $\overrightarrow{ac}$, and $\overrightarrow{ad}$ are linearly independent, which means that a, b, c, and d are not in the same plane (they are not coplanar). In this case, a, b, c, and d are the vertices of a tetrahedron. Figure 24.16 shows affine frames and their convex hulls for $|I| = 0, 1, 2, 3$.

Given $n+1$ affinely independent points $(a_0, \ldots, a_n)$ in E, we can consider the set of points $\lambda_0 a_0 + \cdots + \lambda_n a_n$, where $\lambda_0 + \cdots + \lambda_n = 1$ and $\lambda_i \geq 0$ ($\lambda_i \in \mathbb{R}$). Such affine combinations are called *convex combinations*. This set is called the *convex hull* of $(a_0, \ldots, a_n)$ (or *n-simplex spanned by* $(a_0, \ldots, a_n)$). When $n = 1$, we get the segment between a_0 and a_1, including a_0 and a_1. When $n = 2$, we get the interior of the triangle whose vertices are a_0, a_1, a_2, including boundary points (the edges). When $n = 3$, we get the interior of the tetrahedron whose vertices are a_0, a_1, a_2, a_3, including boundary points (faces and edges). The set

$$\{a_0 + \lambda_1 \overrightarrow{a_0 a_1} + \cdots + \lambda_n \overrightarrow{a_0 a_n} \mid \text{where } 0 \leq \lambda_i \leq 1 \ (\lambda_i \in \mathbb{R})\}$$

is called the *parallelotope spanned by* $(a_0, \ldots, a_n)$. When E has dimension 2, a parallelotope is also called a *parallelogram*, and when E has dimension 3, a *parallelepiped*. Figure 24.17 shows the convex hulls and associated parallelotopes for $|I| = 0, 1, 2, 3$.

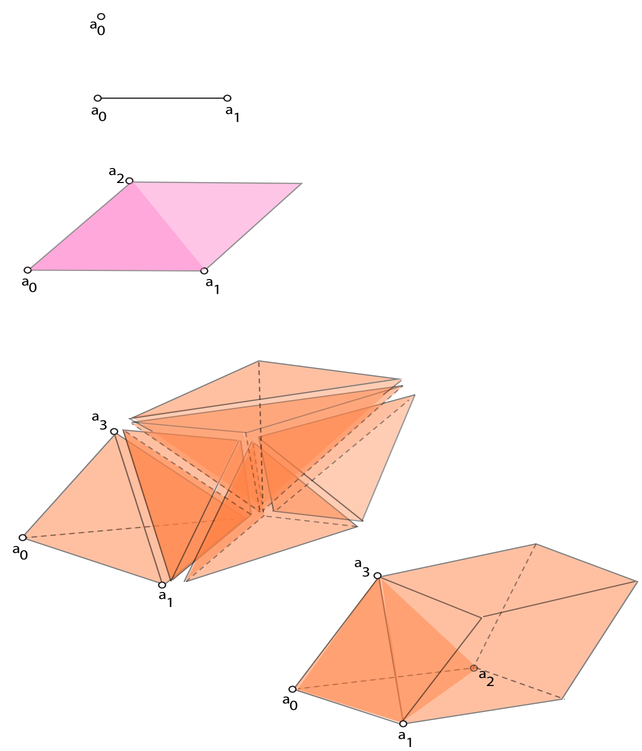


Figure 24.17: Examples of affine frames, convex hulls, and their associated parallelotopes.

More generally, we say that a subset V of E is *convex* if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$).

Points are not vectors! The following example illustrates why treating points as vectors may cause problems. Let a, b, c be three affinely independent points in $\mathbb{A}^3$. Any point x in the plane (a, b, c) can be expressed as

$$x = \lambda_0 a + \lambda_1 b + \lambda_2 c,$$

where $\lambda_0 + \lambda_1 + \lambda_2 = 1$. How can we compute $\lambda_0, \lambda_1, \lambda_2$? Letting $a = (a_1, a_2, a_3)$, $b = (b_1, b_2, b_3)$, $c = (c_1, c_2, c_3)$, and $x = (x_1, x_2, x_3)$ be the coordinates of a, b, c, x in the standard frame of $\mathbb{A}^3$, it is tempting to solve the system of equations

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

However, there is a problem when the origin of the coordinate system belongs to the plane (a, b, c), since in this case, the matrix is not invertible! What we should really be doing is to solve the system

$$\lambda_0\overrightarrow{Oa} + \lambda_1\overrightarrow{Ob} + \lambda_2\overrightarrow{Oc} = \overrightarrow{Ox},$$

where O is any point **not** in the plane (a, b, c). An alternative is to use certain well-chosen cross products.

It can be shown that barycentric coordinates correspond to various ratios of areas and volumes; see the problems.

24.7 Affine Maps

Corresponding to linear maps we have the notion of an affine map. An affine map is defined as a map preserving affine combinations.

Definition 24.6. Given two affine spaces $\langle E, \overrightarrow{E}, +\rangle$ and $\langle E', \overrightarrow{E'}, +'\rangle$, a function $f\colon E \to E'$ is an *affine map* iff for every family $((a_i, \lambda_i))_{i\in I}$ of weighted points in E such that $\sum_{i\in I} \lambda_i = 1$, we have

$$f\left(\sum_{i\in I} \lambda_i a_i\right) = \sum_{i\in I} \lambda_i f(a_i).$$

In other words, f preserves barycenters.

Affine maps can be obtained from linear maps as follows. For simplicity of notation, the same symbol $+$ is used for both affine spaces (instead of using both $+$ and $+'$).

Proposition 24.7. *Given any point $a \in E$, any point $b \in E'$, and any linear map $h\colon \overrightarrow{E} \to \overrightarrow{E'}$, the map $f\colon E \to E'$ defined such that*

$$f(a + v) = b + h(v)$$

is an affine map.

Proof. Indeed, for any family $(\lambda_i)_{i\in I}$ of scalars with $\sum_{i\in I} \lambda_i = 1$ and any family $(v_i)_{i\in I}$, since

$$\sum_{i\in I} \lambda_i(a + v_i) = a + \sum_{i\in I} \lambda_i\overrightarrow{a(a + v_i)} = a + \sum_{i\in I} \lambda_i v_i$$

and

$$\sum_{i \in I} \lambda_i (b + h(v_i)) = b + \sum_{i \in I} \lambda_i \overrightarrow{b(b + h(v_i))} = b + \sum_{i \in I} \lambda_i h(v_i),$$

we have

$$\begin{aligned} f\left(\sum_{i \in I} \lambda_i (a + v_i)\right) &= f\left(a + \sum_{i \in I} \lambda_i v_i\right) \\ &= b + h\left(\sum_{i \in I} \lambda_i v_i\right) \\ &= b + \sum_{i \in I} \lambda_i h(v_i) \\ &= \sum_{i \in I} \lambda_i (b + h(v_i)) \\ &= \sum_{i \in I} \lambda_i f(a + v_i), \end{aligned}$$

as claimed. $\square$

Note that the condition $\sum_{i \in I} \lambda_i = 1$ was implicitly used (in a hidden call to Proposition 24.1) in deriving that

$$\sum_{i \in I} \lambda_i (a + v_i) = a + \sum_{i \in I} \lambda_i v_i$$

and

$$\sum_{i \in I} \lambda_i (b + h(v_i)) = b + \sum_{i \in I} \lambda_i h(v_i).$$

As a more concrete example, the map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

defines an affine map in $\mathbb{A}^2$. It is a "shear" followed by a translation. The effect of this shear on the square (a, b, c, d) is shown in Figure 24.18. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

Let us consider one more example. The map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

is an affine map. Since we can write

$$\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} = \sqrt{2} \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ 2/2 & \sqrt{2}/2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

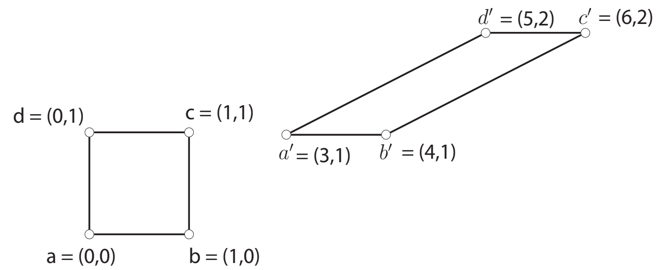


Figure 24.18: The effect of a shear.

this affine map is the composition of a shear, followed by a rotation of angle $\pi/4$, followed by a magnification of ratio $\sqrt{2}$, followed by a translation. The effect of this map on the square (a, b, c, d) is shown in Figure 24.19. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

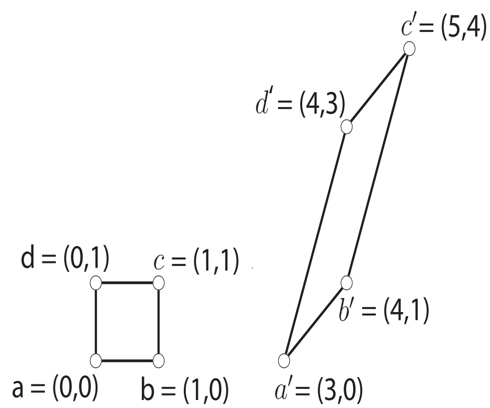


Figure 24.19: The effect of an affine map.

The following proposition shows the converse of what we just showed. Every affine map is determined by the image of any point and a linear map.

Proposition 24.8. *Given an affine map $f\colon E \to E'$, there is a unique linear map $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$ such that*

$$f(a + v) = f(a) + \overrightarrow{f}(v),$$

for every $a \in E$ and every $v \in \overrightarrow{E}$.

Proof. Let $a \in E$ be any point in E. We claim that the map defined such that

$$\overrightarrow{f}(v) = \overrightarrow{f(a)f(a + v)}$$

for every $v \in \overrightarrow{E}$ is a linear map $\overrightarrow{f} \colon \overrightarrow{E} \to \overrightarrow{E'}$. Indeed, we can write

$$a + \lambda v = \lambda(a + v) + (1 - \lambda)a,$$

since $a + \lambda v = a + \lambda\overrightarrow{a(a + v)} + (1 - \lambda)\overrightarrow{aa}$, and also

$$a + u + v = (a + u) + (a + v) - a,$$

since $a + u + v = a + \overrightarrow{a(a + u)} + \overrightarrow{a(a + v)} - \overrightarrow{aa}$. Since f preserves barycenters, we get

$$f(a + \lambda v) = \lambda f(a + v) + (1 - \lambda)f(a).$$

If we recall that $x = \sum_{i\in I} \lambda_i a_i$ is the barycenter of a family $((a_i, \lambda_i))_{i\in I}$ of weighted points (with $\sum_{i\in I} \lambda_i = 1$) iff

$$\overrightarrow{bx} = \sum_{i\in I} \lambda_i \overrightarrow{ba_i} \quad \text{for every } b \in E,$$

we get

$$\overrightarrow{f(a)f(a + \lambda v)} = \lambda\overrightarrow{f(a)f(a + v)} + (1 - \lambda)\overrightarrow{f(a)f(a)} = \lambda\overrightarrow{f(a)f(a + v)},$$

showing that $\overrightarrow{f}(\lambda v) = \lambda\overrightarrow{f}(v)$. We also have

$$f(a + u + v) = f(a + u) + f(a + v) - f(a),$$

from which we get

$$\overrightarrow{f(a)f(a + u + v)} = \overrightarrow{f(a)f(a + u)} + \overrightarrow{f(a)f(a + v)},$$

showing that $\overrightarrow{f}(u + v) = \overrightarrow{f}(u) + \overrightarrow{f}(v)$. Consequently, $\overrightarrow{f}$ is a linear map. For any other point $b \in E$, since

$$b + v = a + \overrightarrow{ab} + v = a + \overrightarrow{a(a + v)} - \overrightarrow{aa} + \overrightarrow{ab},$$

$b + v = (a + v) - a + b$, and since f preserves barycenters, we get

$$f(b + v) = f(a + v) - f(a) + f(b),$$

which implies that

$$\begin{aligned}
\overrightarrow{f(b)f(b + v)} &= \overrightarrow{f(b)f(a + v)} - \overrightarrow{f(b)f(a)} + \overrightarrow{f(b)f(b)}, \\
&= \overrightarrow{f(a)f(b)} + \overrightarrow{f(b)f(a + v)}, \\
&= \overrightarrow{f(a)f(a + v)}.
\end{aligned}$$

Thus, $\overrightarrow{f(b)f(b + v)} = \overrightarrow{f(a)f(a + v)}$, which shows that the definition of $\overrightarrow{f}$ does not depend on the choice of $a \in E$. The fact that $\overrightarrow{f}$ is unique is obvious: We must have $\overrightarrow{f}(v) = \overrightarrow{f(a)f(a + v)}$. $\square$

The unique linear map $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$ given by Proposition 24.8 is called the *linear map associated with the affine map* f.

Note that the condition

$$f(a + v) = f(a) + \overrightarrow{f}(v),$$

for every $a \in E$ and every $v \in \overrightarrow{E}$, can be stated equivalently as

$$f(x) = f(a) + \overrightarrow{f}(\overrightarrow{ax}), \quad \text{or} \quad \overrightarrow{f(a)f(x)} = \overrightarrow{f}(\overrightarrow{ax}),$$

for all $a, x \in E$. Proposition 24.8 shows that for any affine map $f\colon E \to E'$, there are points $a \in E$, $b \in E'$, and a unique linear map $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$, such that

$$f(a + v) = b + \overrightarrow{f}(v),$$

for all $v \in \overrightarrow{E}$ (just let $b = f(a)$, for any $a \in E$). Affine maps for which $\overrightarrow{f}$ is the identity map are called *translations*. Indeed, if $\overrightarrow{f} = \mathrm{id}$,

$$\begin{aligned} f(x) &= f(a) + \overrightarrow{f}(\overrightarrow{ax}) = f(a) + \overrightarrow{ax} = x + \overrightarrow{xa} + \overrightarrow{af(a)} + \overrightarrow{ax} \\ &= x + \overrightarrow{xa} + \overrightarrow{af(a)} - \overrightarrow{xa} = x + \overrightarrow{af(a)}, \end{aligned}$$

and so

$$\overrightarrow{xf(x)} = \overrightarrow{af(a)},$$

which shows that f is the translation induced by the vector $\overrightarrow{af(a)}$ (which does not depend on a).

Since an affine map preserves barycenters, and since an affine subspace V is closed under barycentric combinations, the image $f(V)$ of V is an affine subspace in E'. So, for example, the image of a line is a point or a line, and the image of a plane is either a point, a line, or a plane.

It is easily verified that the composition of two affine maps is an affine map. Also, given affine maps $f\colon E \to E'$ and $g\colon E' \to E''$, we have

$$g(f(a + v)) = g\Big(f(a) + \overrightarrow{f}(v)\Big) = g(f(a)) + \overrightarrow{g}\Big(\overrightarrow{f}(v)\Big),$$

which shows that $\overrightarrow{g \circ f} = \overrightarrow{g} \circ \overrightarrow{f}$. It is easy to show that an affine map $f\colon E \to E'$ is injective iff $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$ is injective, and that $f\colon E \to E'$ is surjective iff $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$ is surjective. An affine map $f\colon E \to E'$ is constant iff $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$ is the null (constant) linear map equal to 0 for all $v \in \overrightarrow{E}$.

If E is an affine space of dimension m and $(a_0, a_1, \ldots, a_m)$ is an affine frame for E, then for any other affine space F and for any sequence $(b_0, b_1, \ldots, b_m)$ of $m + 1$ points in F, there is a unique affine map $f\colon E \to F$ such that $f(a_i) = b_i$, for $0 \leq i \leq m$. Indeed, f must be such that

$$f(\lambda_0 a_0 + \cdots + \lambda_m a_m) = \lambda_0 b_0 + \cdots + \lambda_m b_m,$$

where $\lambda_0 + \cdots + \lambda_m = 1$, and this defines a unique affine map on all of E, since $(a_0, a_1, \ldots, a_m)$ is an affine frame for E.

Using affine frames, affine maps can be represented in terms of matrices. We explain how an affine map $f\colon E \to E$ is represented with respect to a frame $(a_0, \ldots, a_n)$ in E, the more general case where an affine map $f\colon E \to F$ is represented with respect to two affine frames $(a_0, \ldots, a_n)$ in E and $(b_0, \ldots, b_m)$ in F being analogous. Since

$$f(a_0 + x) = f(a_0) + \overrightarrow{f}(x)$$

for all $x \in \overrightarrow{E}$, we have

$$\overrightarrow{a_0 f(a_0 + x)} = \overrightarrow{a_0 f(a_0)} + \overrightarrow{f}(x).$$

Since x, $\overrightarrow{a_0 f(a_0)}$, and $\overrightarrow{a_0 f(a_0 + x)}$, can be expressed as

$$\begin{aligned}
x &= x_1 \overrightarrow{a_0 a_1} + \cdots + x_n \overrightarrow{a_0 a_n}, \\
\overrightarrow{a_0 f(a_0)} &= b_1 \overrightarrow{a_0 a_1} + \cdots + b_n \overrightarrow{a_0 a_n}, \\
\overrightarrow{a_0 f(a_0 + x)} &= y_1 \overrightarrow{a_0 a_1} + \cdots + y_n \overrightarrow{a_0 a_n},
\end{aligned}$$

if $A = (a_{i\,j})$ is the $n \times n$ matrix of the linear map $\overrightarrow{f}$ over the basis $(\overrightarrow{a_0 a_1}, \ldots, \overrightarrow{a_0 a_n})$, letting x, y, and b denote the column vectors of components $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$, and $(b_1, \ldots, b_n)$,

$$\overrightarrow{a_0 f(a_0 + x)} = \overrightarrow{a_0 f(a_0)} + \overrightarrow{f}(x)$$

is equivalent to

$$y = Ax + b.$$

Note that $b \neq 0$ unless $f(a_0) = a_0$. Thus, f is generally not a linear transformation, unless it has a *fixed point*, i.e., there is a point a_0 such that $f(a_0) = a_0$. The vector b is the "translation part" of the affine map. Affine maps do not always have a fixed point. Obviously, nonnull translations have no fixed point. A less trivial example is given by the affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This map is a reflection about the x-axis followed by a translation along the x-axis. The affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3}/4 & 1/4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

can also be written as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which shows that it is the composition of a rotation of angle $\pi/3$, followed by a stretch (by a factor of 2 along the x-axis, and by a factor of $\frac{1}{2}$ along the y-axis), followed by a translation. It is easy to show that this affine map has a unique fixed point. On the other hand, the affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 8/5 & -6/5 \\ 3/10 & 2/5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

has no fixed point, even though

$$\begin{pmatrix} 8/5 & -6/5 \\ 3/10 & 2/5 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 4/5 & -3/5 \\ 3/5 & 4/5 \end{pmatrix},$$

and the second matrix is a rotation of angle θ such that $\cos\theta = \frac{4}{5}$ and $\sin\theta = \frac{3}{5}$.

There is a useful trick to convert the equation $y = Ax + b$ into what looks like a linear equation. The trick is to consider an $(n+1) \times (n+1)$ matrix. We add 1 as the $(n+1)$th component to the vectors x, y, and b, and form the $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix}$$

so that $y = Ax + b$ is equivalent to

$$\begin{pmatrix} y \\ 1 \end{pmatrix} = \begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

This trick is very useful in kinematics and dynamics, where A is a rotation matrix. Such affine maps are called *rigid motions*.

If $f\colon E \to E'$ is a bijective affine map, given any three collinear points a, b, c in E, with $a \neq b$, where, say, $c = (1-\lambda)a + \lambda b$, since f preserves barycenters, we have $f(c) = (1-\lambda)f(a) + \lambda f(b)$, which shows that $f(a), f(b), f(c)$ are collinear in E'. There is a converse to this property, which is simpler to state when the ground field is $K = \mathbb{R}$. The converse states that given any bijective function $f\colon E \to E'$ between two real affine spaces of the same dimension $n \geq 2$, if f maps any three collinear points to collinear points, then f is affine. The proof is rather long (see Berger [11] or Samuel [142]).

Given three collinear points a, b, c, where $a \neq c$, we have $b = (1-\beta)a + \beta c$ for some unique β, and we define the *ratio of the sequence* a, b, c, as

$$\text{ratio}(a, b, c) = \frac{\beta}{(1-\beta)} = \frac{\overrightarrow{ab}}{\overrightarrow{bc}},$$

provided that $\beta \neq 1$, i.e., $b \neq c$. When $b = c$, we agree that $\mathrm{ratio}(a, b, c) = \infty$. We warn our readers that other authors define the ratio of a, b, c as $-\mathrm{ratio}(a, b, c) = \frac{\overrightarrow{ba}}{\overrightarrow{bc}}$. Since affine maps preserve barycenters, it is clear that affine maps preserve the ratio of three points.

24.8 Affine Groups

We now take a quick look at the bijective affine maps. Given an affine space E, the set of affine bijections $f\colon E \to E$ is clearly a group, called the *affine group of* E, and denoted by $\mathbf{GA}(E)$. Recall that the group of bijective linear maps of the vector space $\overrightarrow{E}$ is denoted by $\mathbf{GL}(\overrightarrow{E})$. Then, the map $f \mapsto \overrightarrow{f}$ defines a group homomorphism $L\colon \mathbf{GA}(E) \to \mathbf{GL}(\overrightarrow{E})$. The kernel of this map is the set of translations on E.

The subset of all linear maps of the form $\lambda\,\mathrm{id}_{\overrightarrow{E}}$, where $\lambda \in \mathbb{R} - \{0\}$, is a subgroup of $\mathbf{GL}(\overrightarrow{E})$, and is denoted by $\mathbb{R}^*\mathrm{id}_{\overrightarrow{E}}$ (where $\lambda\,\mathrm{id}_{\overrightarrow{E}}(u) = \lambda u$, and $\mathbb{R}^* = \mathbb{R} - \{0\}$). The subgroup $\mathbf{DIL}(E) = L^{-1}(\mathbb{R}^*\mathrm{id}_{\overrightarrow{E}})$ of $\mathbf{GA}(E)$ is particularly interesting. It turns out that it is the disjoint union of the translations and of the dilatations of ratio $\lambda \neq 1$. The elements of $\mathbf{DIL}(E)$ are called *affine dilatations*.

Given any point $a \in E$, and any scalar $\lambda \in \mathbb{R}$, a *dilatation or central dilatation (or homothety) of center a and ratio* λ is a map $H_{a,\lambda}$ defined such that

$$H_{a,\lambda}(x) = a + \lambda\overrightarrow{ax},$$

for every $x \in E$.

Remark: The terminology does not seem to be universally agreed upon. The terms *affine dilatation* and *central dilatation* are used by Pedoe [136]. Snapper and Troyer use the term *dilation* for an affine dilatation and *magnification* for a central dilatation [162]. Samuel uses *homothety* for a central dilatation, a direct translation of the French "homothétie" [142]. Since dilation is shorter than dilatation and somewhat easier to pronounce, perhaps we should use that!

Observe that $H_{a,\lambda}(a) = a$, and when $\lambda \neq 0$ and $x \neq a$, $H_{a,\lambda}(x)$ is on the line defined by a and x, and is obtained by "scaling" $\overrightarrow{ax}$ by λ.

Figure 24.20 shows the effect of a central dilatation of center d. The triangle (a, b, c) is magnified to the triangle (a', b', c'). Note how every line is mapped to a parallel line.

When $\lambda = 1$, $H_{a,1}$ is the identity. Note that $\overrightarrow{H_{a,\lambda}} = \lambda\,\mathrm{id}_{\overrightarrow{E}}$. When $\lambda \neq 0$, it is clear that $H_{a,\lambda}$ is an affine bijection. It is immediately verified that

$$H_{a,\lambda} \circ H_{a,\mu} = H_{a,\lambda\mu}.$$

We have the following useful result.

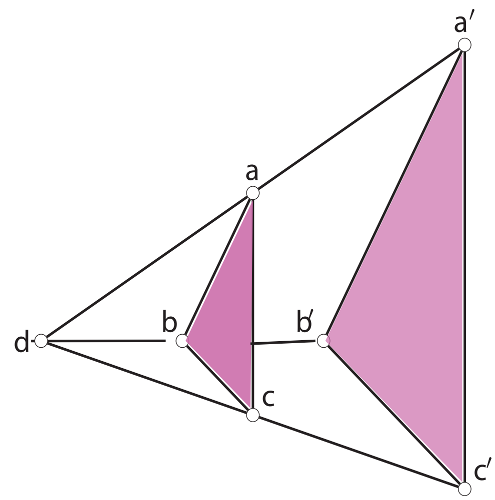


Figure 24.20: The effect of a central dilatation $H_{d,\lambda}(x)$.

Proposition 24.9. *Given any affine space E, for any affine bijection $f \in \mathbf{GA}(E)$, if $\overrightarrow{f} = \lambda\, \mathrm{id}_{\overrightarrow{E}}$, for some $\lambda \in \mathbb{R}^*$ with $\lambda \neq 1$, then there is a unique point $c \in E$ such that $f = H_{c,\lambda}$.*

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [70]. ☐

Clearly, if $\overrightarrow{f} = \mathrm{id}_{\overrightarrow{E}}$, the affine map f is a translation. Thus, the group of affine dilatations $\mathbf{DIL}(E)$ is the disjoint union of the translations and of the dilatations of ratio $\lambda \neq 0, 1$. Affine dilatations can be given a purely geometric characterization.

Another point worth mentioning is that affine bijections preserve the ratio of volumes of parallelotopes. Indeed, given any basis $B = (u_1, \ldots, u_m)$ of the vector space $\overrightarrow{E}$ associated with the affine space E, given any $m+1$ affinely independent points $(a_0, \ldots, a_m)$, we can compute the determinant $\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ w.r.t. the basis B. For any bijective affine map $f\colon E \to E$, since

$$\det_B\Big(\overrightarrow{f}(\overrightarrow{a_0a_1}), \ldots, \overrightarrow{f}(\overrightarrow{a_0a_m})\Big) = \det\big(\overrightarrow{f}\big)\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$$

and the determinant of a linear map is intrinsic (i.e., depends only on $\overrightarrow{f}$, and not on the particular basis B), we conclude that the ratio

$$\frac{\det_B\Big(\overrightarrow{f}(\overrightarrow{a_0a_1}), \ldots, \overrightarrow{f}(\overrightarrow{a_0a_m})\Big)}{\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})} = \det\big(\overrightarrow{f}\big)$$

is independent of the basis B. Since $\det_B(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is the volume of the parallelotope spanned by $(a_0, \ldots, a_m)$, where the parallelotope spanned by any point a and the vectors

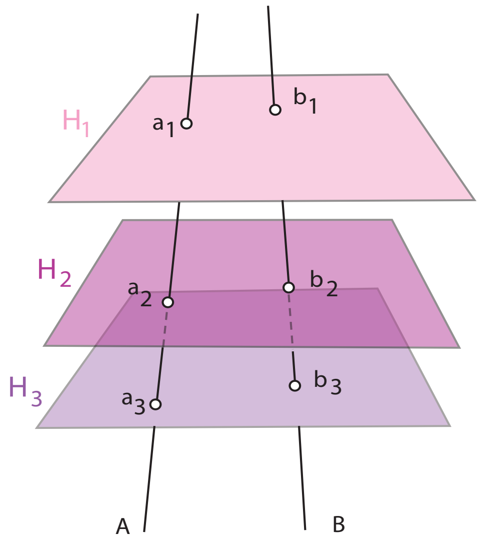


Figure 24.21: The theorem of Thales.

$(u_1, \ldots, u_m)$ has unit volume (see Berger [11], Section 9.12), we see that affine bijections preserve the ratio of volumes of parallelotopes. In fact, this ratio is independent of the choice of the parallelotopes of unit volume. In particular, the affine bijections $f \in \mathbf{GA}(E)$ such that $\det\big(\overrightarrow{f}\big) = 1$ preserve volumes. These affine maps form a subgroup $\mathbf{SA}(E)$ of $\mathbf{GA}(E)$ called the *special affine group of E*. We now take a glimpse at affine geometry.

24.9 Affine Geometry: A Glimpse

In this section we state and prove three fundamental results of affine geometry. Roughly speaking, affine geometry is the study of properties invariant under affine bijections. We now prove one of the oldest and most basic results of affine geometry, the theorem of Thales.

Proposition 24.10. *Given any affine space E, if H_1, H_2, H_3 are any three distinct parallel hyperplanes, and A and B are any two lines not parallel to H_i, letting $a_i = H_i \cap A$ and $b_i = H_i \cap B$, then the following ratios are equal:*

$$\frac{\overrightarrow{a_1a_3}}{\overrightarrow{a_1a_2}} = \frac{\overrightarrow{b_1b_3}}{\overrightarrow{b_1b_2}} = \rho.$$

Conversely, for any point d on the line A, if $\frac{\overrightarrow{a_1d}}{\overrightarrow{a_1a_2}} = \rho$, then $d = a_3$.

Proof. Figure 24.21 illustrates the theorem of Thales. We sketch a proof, leaving the details as an exercise. Since H_1, H_2, H_3 are parallel, they have the same direction $\overrightarrow{H}$, a hyperplane in $\overrightarrow{E}$. Let $u \in \overrightarrow{E} - \overrightarrow{H}$ be any nonnull vector such that $A = a_1 + \mathbb{R}u$. Since A is not parallel to H, we have $\overrightarrow{E} = \overrightarrow{H} \oplus \mathbb{R}u$, and thus we can define the linear map $p\colon \overrightarrow{E} \to \mathbb{R}u$, the projection on $\mathbb{R}u$ parallel to $\overrightarrow{H}$. This linear map induces an affine map $f\colon E \to A$, by defining f such that

$$f(b_1 + w) = a_1 + p(w),$$

for all $w \in \overrightarrow{E}$. Clearly, $f(b_1) = a_1$, and since H_1, H_2, H_3 all have direction $\overrightarrow{H}$, we also have $f(b_2) = a_2$ and $f(b_3) = a_3$. Since f is affine, it preserves ratios, and thus

$$\frac{\overrightarrow{a_1a_3}}{\overrightarrow{a_1a_2}} = \frac{\overrightarrow{b_1b_3}}{\overrightarrow{b_1b_2}}.$$

The converse is immediate. $\square$

We also have the following simple proposition, whose proof is left as an easy exercise.

Proposition 24.11. *Given any affine space E, given any two distinct points $a, b \in E$, and for any affine dilatation f different from the identity, if $a' = f(a)$, $D = \langle a, b \rangle$ is the line passing through a and b, and D' is the line parallel to D and passing through a', the following are equivalent:*

(i) $b' = f(b)$;

(ii) If f is a translation, then b' is the intersection of D' with the line parallel to $\langle a, a' \rangle$ passing through b;

If f is a dilatation of center c, then $b' = D' \cap \langle c, b \rangle$.

The first case is the parallelogram law, and the second case follows easily from Thales' theorem. For an illustration, see Figure 24.22.

We are now ready to prove two classical results of affine geometry, Pappus's theorem and Desargues's theorem. Actually, these results are theorems of projective geometry, and we are stating affine versions of these important results. There are stronger versions that are best proved using projective geometry.

Proposition 24.12. *Given any affine plane E, any two distinct lines D and D', then for any distinct points a, b, c on D and a', b', c' on D', if a, b, c, a', b', c' are distinct from the intersection of D and D' (if D and D' intersect) and if the lines $\langle a, b' \rangle$ and $\langle a', b \rangle$ are parallel, and the lines $\langle b, c' \rangle$ and $\langle b', c \rangle$ are parallel, then the lines $\langle a, c' \rangle$ and $\langle a', c \rangle$ are parallel.*

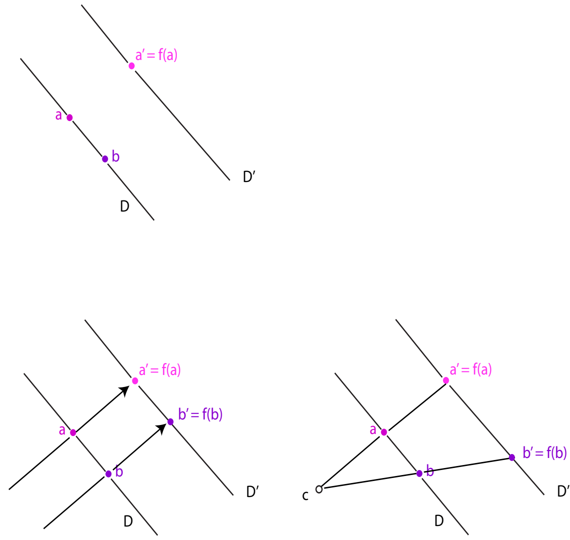


Figure 24.22: An illustration of Proposition 24.11. The bottom left diagram illustrates a translation, while the bottom right illustrates a central dilation through c.

Proof. Pappus's theorem is illustrated in Figure 24.23. If D and D' are not parallel, let d be their intersection. Let f be the dilatation of center d such that $f(a) = b$, and let g be the dilatation of center d such that $g(b) = c$. Since the lines $\langle a, b' \rangle$ and $\langle a', b \rangle$ are parallel, and the lines $\langle b, c' \rangle$ and $\langle b', c \rangle$ are parallel, by Proposition 24.11 we have $a' = f(b')$ and $b' = g(c')$. However, we observed that dilatations with the same center commute, and thus $f \circ g = g \circ f$, and thus, letting $h = g \circ f$, we get $c = h(a)$ and $a' = h(c')$. Again, by Proposition 24.11, the lines $\langle a, c' \rangle$ and $\langle a', c \rangle$ are parallel. If D and D' are parallel, we use translations instead of dilatations. $\square$

There is a converse to Pappus's theorem, which yields a fancier version of Pappus's theorem, but it is easier to prove it using projective geometry. It should be noted that in axiomatic presentations of projective geometry, Pappus's theorem is equivalent to the commutativity of the ground field K (in the present case, $K = \mathbb{R}$). We now prove an affine version of Desargues's theorem.

Proposition 24.13. *Given any affine space E, and given any two triangles (a, b, c) and (a', b', c'), where a, b, c, a', b', c' are all distinct, if $\langle a, b \rangle$ and $\langle a', b' \rangle$ are parallel and $\langle b, c \rangle$ and $\langle b', c' \rangle$ are parallel, then $\langle a, c \rangle$ and $\langle a', c' \rangle$ are parallel iff the lines $\langle a, a' \rangle$, $\langle b, b' \rangle$, and $\langle c, c' \rangle$ are either parallel or concurrent (i.e., intersect in a common point).*

Proof. We prove half of the proposition, the direction in which it is assumed that $\langle a, c \rangle$ and $\langle a', c' \rangle$ are parallel, leaving the converse as an exercise. Since the lines $\langle a, b \rangle$ and $\langle a', b' \rangle$ are

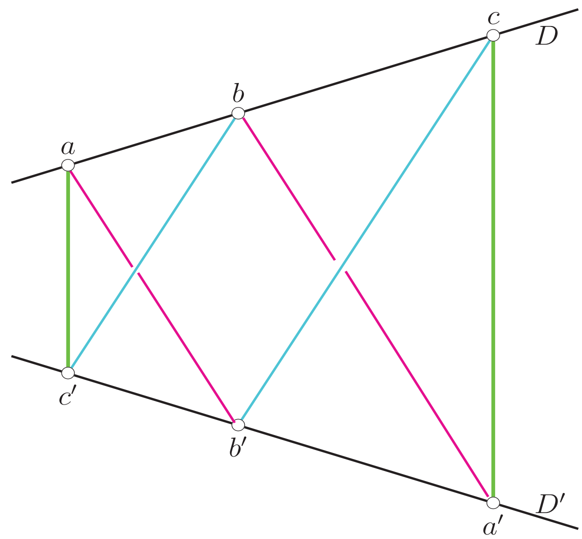


Figure 24.23: Pappus's theorem (affine version).

parallel, the points a, b, a', b' are coplanar. Thus, either $\langle a, a' \rangle$ and $\langle b, b' \rangle$ are parallel, or they have some intersection d. We consider the second case where they intersect, leaving the other case as an easy exercise. Let f be the dilatation of center d such that $f(a) = a'$. By Proposition 24.11, we get $f(b) = b'$. If $f(c) = c''$, again by Proposition 24.11 twice, the lines $\langle b, c \rangle$ and $\langle b', c'' \rangle$ are parallel, and the lines $\langle a, c \rangle$ and $\langle a', c'' \rangle$ are parallel. From this it follows that $c'' = c'$. Indeed, recall that $\langle b, c \rangle$ and $\langle b', c' \rangle$ are parallel, and similarly $\langle a, c \rangle$ and $\langle a', c' \rangle$ are parallel. Thus, the lines $\langle b', c'' \rangle$ and $\langle b', c' \rangle$ are identical, and similarly the lines $\langle a', c'' \rangle$ and $\langle a', c' \rangle$ are identical. Since $\overrightarrow{a'c'}$ and $\overrightarrow{b'c'}$ are linearly independent, these lines have a unique intersection, which must be $c'' = c'$.

The direction where it is assumed that the lines $\langle a, a' \rangle$, $\langle b, b' \rangle$ and $\langle c, c' \rangle$, are either parallel or concurrent is left as an exercise (in fact, the proof is quite similar). $\square$

Desargues's theorem is illustrated in Figure 24.24.

There is a fancier version of Desargues's theorem, but it is easier to prove it using projective geometry. It should be noted that in axiomatic presentations of projective geometry, Desargues's theorem is related to the associativity of the ground field K (in the present case, $K = \mathbb{R}$). Also, Desargues's theorem yields a geometric characterization of the affine dilatations. An affine dilatation f on an affine space E is a bijection that maps every line D to a line $f(D)$ parallel to D. We leave the proof as an exercise.

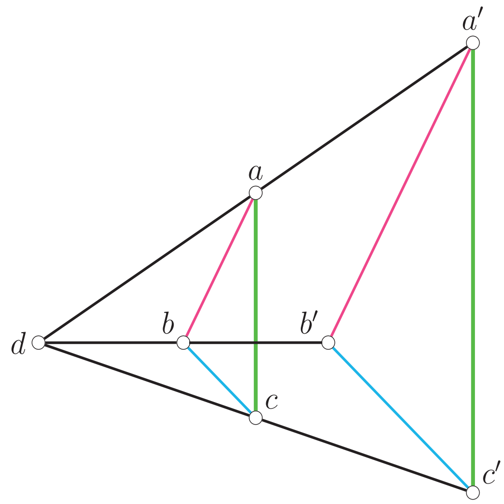


Figure 24.24: Desargues's theorem (affine version).

24.10 Affine Hyperplanes

We now consider affine forms and affine hyperplanes. In Section 24.5 we observed that the set L of solutions of an equation

$$ax + by = c$$

is an affine subspace of $\mathbb{A}^2$ of dimension 1, in fact, a line (provided that a and b are not both null). It would be equally easy to show that the set P of solutions of an equation

$$ax + by + cz = d$$

is an affine subspace of $\mathbb{A}^3$ of dimension 2, in fact, a plane (provided that a, b, c are not all null). More generally, the set H of solutions of an equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

is an affine subspace of $\mathbb{A}^m$, and if $\lambda_1, \ldots, \lambda_m$ are not all null, it turns out that it is a subspace of dimension $m - 1$ called a *hyperplane*.

We can interpret the equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

in terms of the map $f\colon \mathbb{R}^m \to \mathbb{R}$ defined such that

$$f(x_1, \ldots, x_m) = \lambda_1 x_1 + \cdots + \lambda_m x_m - \mu$$

for all $(x_1, \ldots, x_m) \in \mathbb{R}^m$. It is immediately verified that this map is affine, and the set H of solutions of the equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

is the *null set*, or *kernel,* of the affine map $f\colon \mathbb{A}^m \to \mathbb{R}$, in the sense that

$$H = f^{-1}(0) = \{x \in \mathbb{A}^m \mid f(x) = 0\},$$

where $x = (x_1, \ldots, x_m)$.

Thus, it is interesting to consider *affine forms*, which are just affine maps $f\colon E \to \mathbb{R}$ from an affine space to $\mathbb{R}$. Unlike linear forms f^*, for which $\operatorname{Ker} f^*$ is never empty (since it always contains the vector 0), it is possible that $f^{-1}(0) = \emptyset$ for an affine form f. Given an affine map $f\colon E \to \mathbb{R}$, we also denote $f^{-1}(0)$ by $\operatorname{Ker} f$, and we call it the *kernel* of f. Recall that an (affine) hyperplane is an affine subspace of codimension 1. The relationship between affine hyperplanes and affine forms is given by the following proposition.

Proposition 24.14. *Let E be an affine space. The following properties hold:*

(a) Given any nonconstant affine form $f\colon E \to \mathbb{R}$, its kernel $H = \operatorname{Ker} f$ is a hyperplane.

(b) For any hyperplane H in E, there is a nonconstant affine form $f\colon E \to \mathbb{R}$ such that $H = \operatorname{Ker} f$. For any other affine form $g\colon E \to \mathbb{R}$ such that $H = \operatorname{Ker} g$, there is some $\lambda \in \mathbb{R}$ such that $g = \lambda f$ (with $\lambda \neq 0$).

(c) Given any hyperplane H in E and any (nonconstant) affine form $f\colon E \to \mathbb{R}$ such that $H = \operatorname{Ker} f$, every hyperplane H' parallel to H is defined by a nonconstant affine form g such that $g(a) = f(a) - \lambda$, for all $a \in E$ and some $\lambda \in \mathbb{R}$.

Proof. The proof is straightforward, and is omitted. It is also given in Gallier [70]. $\square$

When E is of dimension n, given an affine frame $(a_0, (u_1, \ldots, u_n))$ of E with origin a_0, recall from Definition 24.5 that every point of E can be expressed uniquely as $x = a_0 + x_1u_1 + \cdots + x_nu_n$, where $(x_1, \ldots, x_n)$ are the *coordinates* of x with respect to the affine frame $(a_0, (u_1, \ldots, u_n))$.

Also recall that every linear form f^* is such that $f^*(x) = \lambda_1x_1 + \cdots + \lambda_nx_n$, for every $x = x_1u_1 + \cdots + x_nu_n$ and some $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Since an affine form $f\colon E \to \mathbb{R}$ satisfies the property $f(a_0 + x) = f(a_0) + \overrightarrow{f}(x)$, denoting $f(a_0 + x)$ by $f(x_1, \ldots, x_n)$, we see that we have

$$f(x_1, \ldots, x_n) = \lambda_1x_1 + \cdots + \lambda_nx_n + \mu,$$

where $\mu = f(a_0) \in \mathbb{R}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Thus, a hyperplane is the set of points whose coordinates $(x_1, \ldots, x_n)$ satisfy the (affine) equation

$$\lambda_1x_1 + \cdots + \lambda_nx_n + \mu = 0.$$

24.11 Intersection of Affine Spaces

In this section we take a closer look at the intersection of affine subspaces. This subsection can be omitted at first reading.

First, we need a result of linear algebra. Given a vector space E and any two subspaces M and N, there are several interesting linear maps. We have the canonical injections $i\colon M \to M+N$ and $j\colon N \to M+N$, the canonical injections $in_1\colon M \to M\oplus N$ and $in_2\colon N \to M\oplus N$, and thus, injections $f\colon M\cap N \to M\oplus N$ and $g\colon M\cap N \to M\oplus N$, where f is the composition of the inclusion map from $M \cap N$ to M with in_1, and g is the composition of the inclusion map from $M \cap N$ to N with in_2. Then, we have the maps $f + g\colon M \cap N \to M \oplus N$, and $i - j\colon M \oplus N \to M + N$.

Proposition 24.15. *Given a vector space E and any two subspaces M and N, with the definitions above,*

$$0 \longrightarrow M \cap N \xrightarrow{f+g} M \oplus N \xrightarrow{i-j} M + N \longrightarrow 0$$

is a short exact sequence, which means that $f + g$ is injective, $i - j$ is surjective, and that $\mathrm{Im}\,(f + g) = \mathrm{Ker}\,(i - j)$. *As a consequence, we have the Grassmann relation*

$$\dim(M) + \dim(N) = \dim(M + N) + \dim\,(M \cap N).$$

Proof. It is obvious that $i - j$ is surjective and that $f + g$ is injective. Assume that $(i - j)(u + v) = 0$, where $u \in M$, and $v \in N$. Then, $i(u) = j(v)$, and thus, by definition of i and j, there is some $w \in M \cap N$, such that $i(u) = j(v) = w \in M \cap N$. By definition of f and g, $u = f(w)$ and $v = g(w)$, and thus $\mathrm{Im}\,(f + g) = \mathrm{Ker}\,(i - j)$, as desired. The second part of the proposition follows from standard results of linear algebra (see Artin [7], Strang [170], or Lang [109]). $\square$

We now prove a simple proposition about the intersection of affine subspaces.

Proposition 24.16. *Given any affine space E, for any two nonempty affine subspaces M and N, the following facts hold:*

(1) $M \cap N \neq \emptyset$ iff $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$.

(2) $M \cap N$ consists of a single point iff $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$, and $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$.

(3) If S is the least affine subspace containing M and N, then $\overrightarrow{S} = \overrightarrow{M} + \overrightarrow{N} + K\overrightarrow{ab}$ (the vector space $\overrightarrow{E}$ is defined over the field K).

Proof. (1) Pick any $a \in M$ and any $b \in N$, which is possible, since M and N are nonempty. Since $\overrightarrow{M} = \{\overrightarrow{ax} \mid x \in M\}$ and $\overrightarrow{N} = \{\overrightarrow{by} \mid y \in N\}$, if $M \cap N \neq \emptyset$, for any $c \in M \cap N$ we have $\overrightarrow{ab} = \overrightarrow{ac} - \overrightarrow{bc}$, with $\overrightarrow{ac} \in \overrightarrow{M}$ and $\overrightarrow{bc} \in \overrightarrow{N}$, and thus, $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$. Conversely, assume that $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$. Then $\overrightarrow{ab} = \overrightarrow{ax} + \overrightarrow{by}$, for some $x \in M$ and some $y \in N$. But we also have

$$\overrightarrow{ab} = \overrightarrow{ax} + \overrightarrow{xy} + \overrightarrow{yb},$$

and thus we get $0 = \overrightarrow{xy} + \overrightarrow{yb} - \overrightarrow{by}$, that is, $\overrightarrow{xy} = 2\overrightarrow{by}$. Thus, b is the middle of the segment $[x, y]$, and since $\overrightarrow{yx} = 2\overrightarrow{yb}$, $x = 2b - y$ is the barycenter of the weighted points $(b, 2)$ and $(y, -1)$. Thus x also belongs to N, since N being an affine subspace, it is closed under barycenters. Thus, $x \in M \cap N$, and $M \cap N \neq \emptyset$.

(2) Note that in general, if $M \cap N \neq \emptyset$, then

$$\overrightarrow{M \cap N} = \overrightarrow{M} \cap \overrightarrow{N},$$

because

$$\overrightarrow{M \cap N} = \{\overrightarrow{ab} \mid a, b \in M \cap N\} = \{\overrightarrow{ab} \mid a, b \in M\} \cap \{\overrightarrow{ab} \mid a, b \in N\} = \overrightarrow{M} \cap \overrightarrow{N}.$$

Since $M \cap N = c + \overrightarrow{M \cap N}$ for any $c \in M \cap N$, we have

$$M \cap N = c + \overrightarrow{M} \cap \overrightarrow{N} \quad \text{for any } c \in M \cap N.$$

From this it follows that if $M \cap N \neq \emptyset$, then $M \cap N$ consists of a single point iff $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$. This fact together with what we proved in (1) proves (2).

(3) This is left as an easy exercise. ☐

Remarks:

(1) The proof of Proposition 24.16 shows that if $M \cap N \neq \emptyset$, then $\overrightarrow{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for all $a \in M$ and all $b \in N$.

(2) Proposition 24.16 implies that for any two nonempty affine subspaces M and N, if $\overrightarrow{E} = \overrightarrow{M} \oplus \overrightarrow{N}$, then $M \cap N$ consists of a single point. Indeed, if $\overrightarrow{E} = \overrightarrow{M} \oplus \overrightarrow{N}$, then $\overrightarrow{ab} \in \overrightarrow{E}$ for all $a \in M$ and all $b \in N$, and since $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$, the result follows from part (2) of the proposition.

We can now state the following proposition.

Proposition 24.17. *Given an affine space E and any two nonempty affine subspaces M and N, if S is the least affine subspace containing M and N, then the following properties hold:*

(1) If $M \cap N = \emptyset$, then

$$\dim(M) + \dim(N) < \dim(E) + \dim(\overrightarrow{M} + \overrightarrow{N})$$

and

$$\dim(S) = \dim(M) + \dim(N) + 1 - \dim(\overrightarrow{M} \cap \overrightarrow{N}).$$

(2) If $M \cap N \neq \emptyset$, then

$$\dim(S) = \dim(M) + \dim(N) - \dim(M \cap N).$$

Proof. The proof is not difficult, using Proposition 24.16 and Proposition 24.15, but we leave it as an exercise. $\square$

Chapter 25

Embedding an Affine Space in a Vector Space

25.1 The "Hat Construction," or Homogenizing

For all practical purposes, most geometric objects, including curves and surfaces, live in affine spaces. A disadvantage of the affine world is that points and vectors live in disjoint universes. It is often more convenient, at least mathematically, to deal with linear objects (vector spaces, linear combinations, linear maps), rather than affine objects (affine spaces, affine combinations, affine maps). Actually, it would also be advantageous if we could manipulate points and vectors as if they lived in a common universe, using perhaps an extra bit of information to distinguish between them if necessary.

Such a "homogenization" (or "hat construction") can be achieved. As a matter of fact, such a homogenization of an affine space and its associated vector space will be very useful to define and manipulate rational curves and surfaces. Indeed, the hat construction yields a canonical construction of the projective completion of an affine space. It also leads to a very elegant method for obtaining the various formulae giving the derivatives of a polynomial curve, or the directional derivatives of polynomial surfaces. However, these formulae are not needed here. Thus we omit this topic, referring the readers to Gallier [70].

This chapter proceeds as follows. First, the construction of a vector space $\widehat{E}$ in which both E and $\overrightarrow{E}$ are embedded as (affine) hyperplanes is described. It is shown how affine frames in E become bases in $\widehat{E}$. It turns out that $\widehat{E}$ is characterized by a universality property: Affine maps to vector spaces extend uniquely to linear maps. As a consequence, affine maps between affine spaces E and F extend to linear maps between $\widehat{E}$ and $\widehat{F}$.

Let us first explain how to distinguish between points and vectors practically, using what amounts to a "hacking trick." Then, we will show that such a procedure can be put on firm mathematical grounds.

Assume that we consider the real affine space E of dimension 3, and that we have some affine frame $(a_0, (v_1, v_2, v_2))$. With respect to this affine frame, every point $x \in E$ is represented by its coordinates (x_1, x_2, x_3), where $a = a_0 + x_1 v_1 + x_2 v_2 + x_3 v_3$. A vector $u \in \overrightarrow{E}$ is also represented by its coordinates (u_1, u_2, u_3) over the basis (v_1, v_2, v_2). One way to distinguish between points and vectors is to add a fourth coordinate, and to agree that points are represented by (row) vectors $(x_1, x_2, x_3, 1)$ whose fourth coordinate is 1, and that vectors are represented by (row) vectors $(v_1, v_2, v_3, 0)$ whose fourth coordinate is 0. This "programming trick" actually works very well. Of course, we are opening the door for strange elements such as $(x_1, x_2, x_3, 5)$, where the fourth coordinate is neither 1 nor 0.

The question is, can we make sense of such elements, and of such a construction? The answer is yes. We will present a construction in which an affine space $\big(E, \overrightarrow{E}\big)$ is embedded in a vector space $\widehat{E}$, in which $\overrightarrow{E}$ is embedded as a hyperplane passing through the origin, and E itself is embedded as an affine hyperplane, defined as $\omega^{-1}(1)$, for some linear form $\omega\colon \widehat{E} \to \mathbb{R}$. In the case of an affine space E of dimension 2, we can think of $\widehat{E}$ as the vector space $\mathbb{R}^3$ of dimension 3 in which $\overrightarrow{E}$ corresponds to the xy-plane, and E corresponds to the plane of equation $z = 1$, parallel to the xy-plane and passing through the point on the z-axis of coordinates $(0, 0, 1)$. The construction of the vector space $\widehat{E}$ is presented in some detail in Berger [11]. Berger explains the construction in terms of vector fields. We prefer a more geometric and simpler description in terms of simple geometric transformations, translations, and dilatations.

Remark: Readers with a good knowledge of geometry will recognize the first step in embedding an affine space into a projective space. We will also show that the homogenization $\widehat{E}$ of an affine space $\big(E, \overrightarrow{E}\big)$, satisfies a universal property with respect to the extension of affine maps to linear maps. As a consequence, the vector space $\widehat{E}$ is unique up to isomorphism, and its actual construction is not so important. However, it is quite useful to visualize the space $\widehat{E}$, in order to understand well rational curves and rational surfaces.

As usual, for simplicity, it is assumed that all vector spaces are defined over the field $\mathbb{R}$ of real numbers, and that all families of scalars (points and vectors) are finite. The extension to arbitrary fields and to families of finite support is immediate. We begin by defining two very simple kinds of geometric (affine) transformations. Given an affine space $\big(E, \overrightarrow{E}\big)$, every $u \in \overrightarrow{E}$ induces a mapping $t_u\colon E \to E$, called a *translation*, and defined such that $t_u(a) = a + u$ for every $a \in E$. Clearly, the set of translations is a vector space isomorphic to $\overrightarrow{E}$. Thus, we will use the same notation u for both the vector u and the translation t_u. Given any point a and any scalar $\lambda \in \mathbb{R}$, we define the mapping $H_{a,\lambda}\colon E \to E$, called *dilatation (or central dilatation, or homothety) of center a and ratio λ*, and defined such that

$$H_{a,\lambda}(x) = a + \lambda\overrightarrow{ax},$$

for every $x \in E$. We have $H_{a,\lambda}(a) = a$, and when $\lambda \neq 0$ and $x \neq a$, $H_{a,\lambda}(x)$ is on the line defined by a and x, and is obtained by "scaling" $\overrightarrow{ax}$ by λ. The effect is a uniform dilatation (or contraction, if $\lambda < 1$). When $\lambda = 0$, $H_{a,0}(x) = a$ for all $x \in E$, and $H_{a,0}$ is the constant affine map sending every point to a. If we assume $\lambda \neq 1$, note that $H_{a,\lambda}$ is never the identity, and since a is a fixed point, $H_{a,\lambda}$ is never a translation.

We now consider the set $\widehat{E}$ of geometric transformations from E to E, consisting of the union of the (disjoint) sets of translations and dilatations of ratio $\lambda \neq 1$. We would like to give this set the structure of a vector space, in such a way that both E and $\overrightarrow{E}$ can be naturally embedded into $\widehat{E}$. In fact, it will turn out that barycenters show up quite naturally too!

In order to “add” two dilatations H_{a_1,λ_1} and H_{a_2,λ_2}, it turns out that it is more convenient to consider dilatations of the form $H_{a,1-\lambda}$, where $\lambda \neq 0$. To see this, let us see the effect of such a dilatation on a point $x \in E$: We have

$$H_{a,1-\lambda}(x) = a + (1-\lambda)\overrightarrow{ax} = a + \overrightarrow{ax} - \lambda\overrightarrow{ax} = x + \lambda\overrightarrow{xa}.$$

For simplicity of notation, let us denote $H_{a,1-\lambda}$ by $\langle a, \lambda\rangle$. Then, we have

$$\langle a, \lambda\rangle(x) = x + \lambda\overrightarrow{xa}.$$

Remarks:

(1) Note that $H_{a,1-\lambda}(x) = H_{x,\lambda}(a)$.

(2) Berger defines a map $h\colon E \to \overrightarrow{E}$ as a *vector field*. Thus, each $\langle a, \lambda\rangle$ can be viewed as the vector field $x \mapsto \lambda\overrightarrow{xa}$. Similarly, a translation u can be viewed as the constant vector field $x \mapsto u$. Thus, we could define $\widehat{E}$ as the (disjoint) union of these two vector fields. We prefer our view in terms of geometric transformations.

Then, since

$$\langle a_1, \lambda_1\rangle(x) = x + \lambda_1\overrightarrow{xa_1} \quad \text{and} \quad \langle a_2, \lambda_2\rangle(x) = x + \lambda_2\overrightarrow{xa_2},$$

if we want to define $\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle$, we see that we have to distinguish between two cases:

(1) $\lambda_1 + \lambda_2 = 0$. In this case, since

$$\lambda_1\overrightarrow{xa_1} + \lambda_2\overrightarrow{xa_2} = \lambda_1\overrightarrow{xa_1} - \lambda_1\overrightarrow{xa_2} = \lambda_1\overrightarrow{a_2a_1},$$

we let

$$\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle = \lambda_1\overrightarrow{a_2a_1},$$

where $\lambda_1\overrightarrow{a_2a_1}$ denotes the translation associated with the vector $\lambda_1\overrightarrow{a_2a_1}$.

(2) $\lambda_1 + \lambda_2 \neq 0$. In this case, the points a_1 and a_2 assigned the weights $\lambda_1/(\lambda_1 + \lambda_2)$ and $\lambda_2/(\lambda_1 + \lambda_2)$ have a barycenter

$$b = \frac{\lambda_1}{\lambda_1 + \lambda_2}a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2}a_2,$$

such that

$$\overrightarrow{xb} = \frac{\lambda_1}{\lambda_1 + \lambda_2}\overrightarrow{xa_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2}\overrightarrow{xa_2}.$$

Since

$$\lambda_1\overrightarrow{xa_1} + \lambda_2\overrightarrow{xa_2} = (\lambda_1 + \lambda_2)\overrightarrow{xb},$$

we let

$$\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle = \left\langle \frac{\lambda_1}{\lambda_1 + \lambda_2} a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} a_2, \lambda_1 + \lambda_2 \right\rangle,$$

the dilatation associated with the point b and the scalar $\lambda_1 + \lambda_2$.

Given a translation defined by u and a dilatation $\langle a, \lambda\rangle$, since $\lambda \neq 0$, we have

$$\lambda\overrightarrow{xa} + u = \lambda(\overrightarrow{xa} + \lambda^{-1}u),$$

and so, letting $b = a + \lambda^{-1}u$, since $\overrightarrow{ab} = \lambda^{-1}u$, we have

$$\lambda\overrightarrow{xa} + u = \lambda(\overrightarrow{xa} + \lambda^{-1}u) = \lambda(\overrightarrow{xa} + \overrightarrow{ab}) = \lambda\overrightarrow{xb},$$

and we let

$$\langle a, \lambda\rangle \mathbin{\widehat{+}} u = \langle a + \lambda^{-1}u, \lambda\rangle,$$

the dilatation of center $a + \lambda^{-1}u$ and ratio λ.

The sum of two translations u and v is of course defined as the translation $u + v$. It is also natural to define multiplication by a scalar as follows:

$$\mu \cdot \langle a, \lambda\rangle = \langle a, \lambda\mu\rangle,$$

and

$$\lambda \cdot u = \lambda u,$$

where λu is the product by a scalar in $\overrightarrow{E}$.

We can now use the definition of the above operations to state the following proposition, showing that the "hat construction" described above has allowed us to achieve our goal of embedding both E and $\overrightarrow{E}$ in the vector space $\widehat{E}$.

Proposition 25.1. *The set $\widehat{E}$ consisting of the disjoint union of the translations and the dilatations $H_{a,1-\lambda} = \langle a, \lambda\rangle$, $\lambda \in \mathbb{R}, \lambda \neq 0$, is a vector space under the following operations of addition and multiplication by a scalar: If $\lambda_1 + \lambda_2 = 0$, then*

$$\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle = \lambda_1\overrightarrow{a_2a_1};$$

if $\lambda_1 + \lambda_2 \neq 0$, then

$$\begin{aligned}
\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle &= \left\langle \frac{\lambda_1}{\lambda_1 + \lambda_2} a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} a_2, \lambda_1 + \lambda_2 \right\rangle, \\
\langle a, \lambda\rangle \mathbin{\widehat{+}} u &= u \mathbin{\widehat{+}} \langle a, \lambda\rangle = \langle a + \lambda^{-1}u, \lambda\rangle, \\
u \mathbin{\widehat{+}} v &= u + v;
\end{aligned}$$

if $\mu \neq 0$, then

$$\begin{aligned} \mu \cdot \langle a, \lambda \rangle &= \langle a, \lambda\mu \rangle, \\ 0 \cdot \langle a, \lambda \rangle &= 0; \end{aligned}$$

and

$$\lambda \cdot u = \lambda u.$$

Furthermore, the map $\omega\colon \widehat{E} \to \mathbb{R}$ defined such that

$$\begin{aligned} \omega(\langle a, \lambda \rangle) &= \lambda, \\ \omega(u) &= 0, \end{aligned}$$

is a linear form, $\omega^{-1}(0)$ is a hyperplane isomorphic to $\overrightarrow{E}$ under the injective linear map $i\colon \overrightarrow{E} \to \widehat{E}$ such that $i(u) = t_u$ (the translation associated with u), and $\omega^{-1}(1)$ is an affine hyperplane isomorphic to E with direction $i(\overrightarrow{E})$, under the injective affine map $j\colon E \to \widehat{E}$, where $j(a) = \langle a, 1 \rangle$ for every $a \in E$. Finally, for every $a \in E$, we have

$$\widehat{E} = i\big(\overrightarrow{E}\big) \oplus \mathbb{R}j(a).$$

Proof. The verification that $\widehat{E}$ is a vector space is straightforward. The linear map mapping a vector u to the translation defined by u is clearly an injection $i\colon \overrightarrow{E} \to \widehat{E}$ embedding $\overrightarrow{E}$ as an hyperplane in $\widehat{E}$. It is also clear that ω is a linear form. Note that

$$j(a + u) = \langle a + u, 1 \rangle = \langle a, 1 \rangle \mathbin{\widehat{+}} u,$$

where u stands for the translation associated with the vector u, and thus j is an affine injection with associated linear map i. Thus, $\omega^{-1}(1)$ is indeed an affine hyperplane isomorphic to E with direction $i\big(\overrightarrow{E}\big)$, under the map $j\colon E \to \widehat{E}$. Finally, from the definition of $\widehat{+}$, for every $a \in E$ and every $u \in \overrightarrow{E}$, since

$$i(u) \mathbin{\widehat{+}} \lambda \cdot j(a) = u \mathbin{\widehat{+}} \langle a, \lambda \rangle = \langle a + \lambda^{-1}u, \lambda \rangle,$$

when $\lambda \neq 0$, we get any arbitrary $v \in \widehat{E}$ by picking $\lambda = 0$ and $u = v$, and we get any arbitrary element $\langle b, \mu \rangle$, $\mu \neq 0$, by picking $\lambda = \mu$ and $u = \mu\overrightarrow{ab}$. Thus,

$$\widehat{E} = i\big(\overrightarrow{E}\big) + \mathbb{R}j(a),$$

and since $i\big(\overrightarrow{E}\big) \cap \mathbb{R}j(a) = \{0\}$, we have

$$\widehat{E} = i\big(\overrightarrow{E}\big) \oplus \mathbb{R}j(a),$$

for every $a \in E$. $\square$

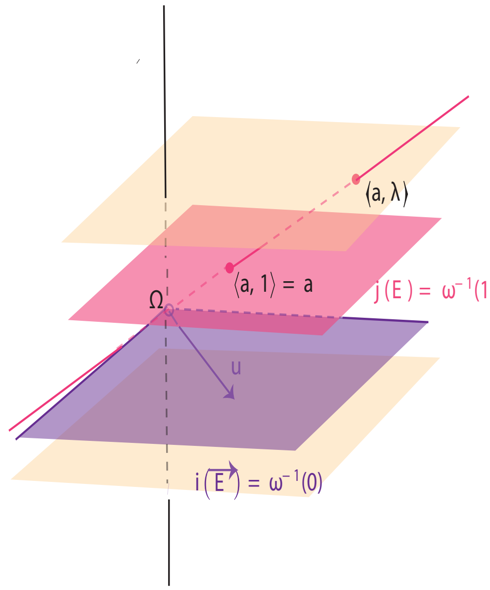


Figure 25.1: Embedding an affine space $(E, \overrightarrow{E})$ into a vector space $\widehat{E}$.

Figure 25.1 illustrates the embedding of the affine space E into the vector space $\widehat{E}$, when E is an affine plane.

Note that $\widehat{E}$ is isomorphic to $\overrightarrow{E} \cup (E \times \mathbb{R}^*)$. Intuitively, we can think of $\widehat{E}$ as a stack of parallel hyperplanes, one for each λ, a little bit like an infinite stack of very thin pancakes! There are two privileged pancakes: one corresponding to E, for $\lambda = 1$, and one corresponding to $\overrightarrow{E}$, for $\lambda = 0$.

From now on, we will identify $j(E)$ and E, and $i(\overrightarrow{E})$ and $\overrightarrow{E}$. We will also write λa instead of $\langle a, \lambda \rangle$, which we will call a *weighted point*, and write $1a$ just as a. When we want to be more precise, we may also write $\langle a, 1 \rangle$ as $\overline{a}$. In particular, when we consider the homogenized version $\widehat{\mathbb{A}}$ of the affine space $\mathbb{A}$ associated with the field $\mathbb{R}$ considered as an affine space, we write $\overline{\lambda}$ for $\langle \lambda, 1 \rangle$, when viewing λ as a point in both $\mathbb{A}$ and $\widehat{\mathbb{A}}$, and simply λ, when viewing λ as a vector in $\mathbb{R}$ and in $\widehat{\mathbb{A}}$. As an example, the expression $2 + 3$ denotes the real number 5, in $\mathbb{A}$, $(\overline{2} + \overline{3})/2$ denotes the midpoint of the segment $[\overline{2}, \overline{3}]$, which can be denoted by $\overline{2.5}$, and $\overline{2} + \overline{3}$ does not make sense in $\mathbb{A}$, since it is not a barycentric combination. However, in $\widehat{\mathbb{A}}$, the expression $\overline{2} + \overline{3}$ makes sense: It is the weighted point $\langle \overline{2.5}, 2 \rangle$.

Then, in view of the fact that

$$\langle a + u, 1 \rangle = \langle a, 1 \rangle \mathbin{\widehat{+}} u,$$

and since we are identifying $a + u$ with $\langle a + u, 1 \rangle$ (under the injection j), in the simplified notation the above reads as $a + u = a \mathbin{\widehat{+}} u$. Thus, we go one step further, and denote $a \mathbin{\widehat{+}} u$

by $a + u$. However, since

$$\langle a, \lambda \rangle \mathbin{\widehat{+}} u = \langle a + \lambda^{-1} u, \lambda \rangle,$$

we will refrain from writing $\lambda a \mathbin{\widehat{+}} u$ as $\lambda a + u$, because we find it too confusing. From Proposition 25.1, for every $a \in E$, every element of $\widehat{E}$ can be written uniquely as $u \mathbin{\widehat{+}} \lambda a$. We also denote

$$\lambda a \mathbin{\widehat{+}} (-\mu) b$$

by

$$\lambda a \mathbin{\widehat{-}} \mu b.$$

We can now justify rigorously the programming trick of the introduction of an extra coordinate to distinguish between points and vectors. First, we make a few observations. Given any family $(a_i)_{i \in I}$ of points in E, and any family $(\lambda_i)_{i \in I}$ of scalars in $\mathbb{R}$, it is easily shown by induction on the size of I that the following holds:

(1) If $\sum_{i \in I} \lambda_i = 0$, then

$$\sum_{i \in I} \langle a_i, \lambda_i \rangle = \overrightarrow{\sum_{i \in I} \lambda_i a_i},$$

where

$$\overrightarrow{\sum_{i \in I} \lambda_i a_i} = \sum_{i \in I} \lambda_i \overrightarrow{b a_i}$$

for any $b \in E$, which, by Proposition 24.1, is a vector independent of b, or

(2) If $\sum_{i \in I} \lambda_i \neq 0$, then

$$\sum_{i \in I} \langle a_i, \lambda_i \rangle = \left\langle \sum_{i \in I} \frac{\lambda_i}{\sum_{i \in I} \lambda_i} a_i, \sum_{i \in I} \lambda_i \right\rangle.$$

Thus, we see how barycenters reenter the scene quite naturally, and that in $\widehat{E}$, we can make sense of $\sum_{i \in I} \langle a_i, \lambda_i \rangle$, regardless of the value of $\sum_{i \in I} \lambda_i$. When $\sum_{i \in I} \lambda_i = 1$, the element $\sum_{i \in I} \langle a_i, \lambda_i \rangle$ belongs to the hyperplane $\omega^{-1}(1)$, and thus it is a point. When $\sum_{i \in I} \lambda_i = 0$, the linear combination of points $\sum_{i \in I} \lambda_i a_i$ is a vector, and when $I = \{1, \ldots, n\}$, we allow ourselves to write

$$\lambda_1 a_1 \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} \lambda_n a_n,$$

where some of the occurrences of $\widehat{+}$ can be replaced by $\widehat{-}$, as

$$\lambda_1 a_1 + \cdots + \lambda_n a_n,$$

where the occurrences of $\widehat{-}$ (if any) are replaced by $-$.

In fact, we have the following slightly more general property, which is left as an exercise.

Proposition 25.2. *Given any affine space $\big(E, \overrightarrow{E}\big)$, for any family $(a_i)_{i\in I}$ of points in E, any family $(\lambda_i)_{i\in I}$ of scalars in $\mathbb{R}$, and any family $(v_j)_{j\in J}$ of vectors in $\overrightarrow{E}$, with $I \cap J = \emptyset$, the following properties hold:*

(1) If $\sum_{i\in I} \lambda_i = 0$, then

$$\sum_{i\in I} \langle a_i, \lambda_i\rangle \mathbin{\widehat{+}} \sum_{j\in J} v_j = \overrightarrow{\sum_{i\in I} \lambda_i a_i} + \sum_{j\in J} v_j,$$

where

$$\overrightarrow{\sum_{i\in I} \lambda_i a_i} = \sum_{i\in I} \lambda_i \overrightarrow{b a_i}$$

for any $b \in E$, which, by Proposition 24.1, is a vector independent of b, or

(2) If $\sum_{i\in I} \lambda_i \neq 0$, then

$$\sum_{i\in I} \langle a_i, \lambda_i\rangle \mathbin{\widehat{+}} \sum_{j\in J} v_j = \left\langle \sum_{i\in I} \frac{\lambda_i}{\sum_{i\in I} \lambda_i} a_i + \sum_{j\in J} \frac{v_j}{\sum_{i\in I} \lambda_i}, \sum_{i\in I} \lambda_i \right\rangle.$$

Proof. By induction on the size of I and the size of J. ☐

The above formulae show that we have some kind of extended barycentric calculus. Operations on weighted points and vectors were introduced by H. Grassmann, in his book published in 1844! This calculus will be helpful in dealing with rational curves.

25.2 Affine Frames of E and Bases of $\widehat{E}$

There is also a nice relationship between affine frames in $\big(E, \overrightarrow{E}\big)$ and bases of $\widehat{E}$, stated in the following proposition.

Proposition 25.3. *Given any affine space $\big(E, \overrightarrow{E}\big)$, for any affine frame $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ for E, the family $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}, a_0)$ is a basis for $\widehat{E}$, and for any affine frame $(a_0, \ldots, a_m)$ for E, the family $(a_0, \ldots, a_m)$ is a basis for $\widehat{E}$. Furthermore, given any element $\langle x, \lambda\rangle \in \widehat{E}$, if*

$$x = a_0 + x_1\overrightarrow{a_0a_1} + \cdots + x_m\overrightarrow{a_0a_m}$$

over the affine frame $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ in E, then the coordinates of $\langle x, \lambda\rangle$ over the basis $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}, a_0)$ in $\widehat{E}$ are

$$(\lambda x_1, \ldots, \lambda x_m, \lambda).$$

For any vector $v \in \overrightarrow{E}$, if

$$v = v_1\overrightarrow{a_0a_1} + \cdots + v_m\overrightarrow{a_0a_m}$$

over the basis $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ *in* $\overrightarrow{E}$*, then over the basis* $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m},\ a_0)$ *in* $\widehat{E}$*, the coordinates of* v *are*

$$(v_1, \ldots, v_m, 0).$$

For any element $\langle a, \lambda\rangle$*, where* $\lambda \neq 0$*, if the barycentric coordinates of* a *w.r.t. the affine basis* $(a_0, \ldots, a_m)$ *in* E *are* $(\lambda_0, \ldots, \lambda_m)$ *with* $\lambda_0 + \cdots + \lambda_m = 1$*, then the coordinates of* $\langle a, \lambda\rangle$ *w.r.t. the basis* $(a_0, \ldots, a_m)$ *in* $\widehat{E}$ *are*

$$(\lambda\lambda_0, \ldots, \lambda\lambda_m).$$

If a vector $v \in \overrightarrow{E}$ *is expressed as*

$$v = v_1\overrightarrow{a_0a_1} + \cdots + v_m\overrightarrow{a_0a_m} = -(v_1 + \cdots + v_m)a_0 + v_1a_1 + \cdots + v_ma_m,$$

with respect to the affine basis $(a_0, \ldots, a_m)$ *in* E*, then its coordinates w.r.t. the basis* $(a_0, \ldots, a_m)$ *in* $\widehat{E}$ *are*

$$(-(v_1 + \cdots + v_m), v_1, \ldots, v_m).$$

Proof. We sketch parts of the proof, leaving the details as an exercise. Figure 25.2 shows the basis $(\overrightarrow{a_0a_1}, \overrightarrow{a_0a_2}, a_0)$ corresponding to the affine frame (a_0, a_1, a_2) in E.

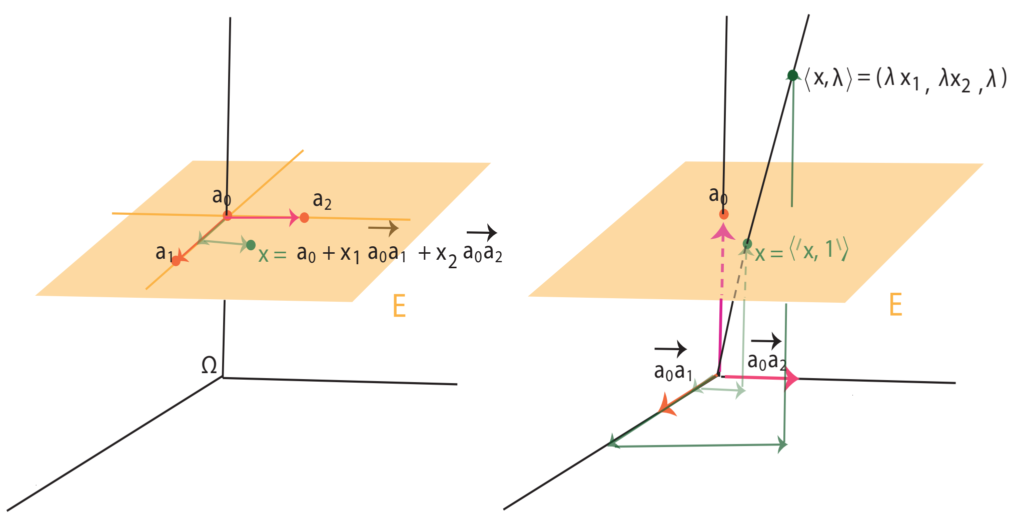


Figure 25.2: The affine frame (a_0, a_1, a_2) of E and the basis $(\overrightarrow{a_0a_1}, \overrightarrow{a_0a_2}, a_0)$ in $\widehat{E}$.

If we assume that we have a nontrivial linear combination

$$\lambda_1\overrightarrow{a_0a_1} \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} \lambda_m\overrightarrow{a_0a_m} \mathbin{\widehat{+}} \mu a_0 = 0,$$

if $\mu \neq 0$, then we have

$$\lambda_1\overrightarrow{a_0a_1} \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} \lambda_m\overrightarrow{a_0a_m} \mathbin{\widehat{+}} \mu a_0 = \langle a_0 + \mu^{-1}\lambda_1\overrightarrow{a_0a_1} + \cdots + \mu^{-1}\lambda_m\overrightarrow{a_0a_m}, \mu\rangle,$$

which is never null, and thus, $\mu = 0$, but since $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ is a basis of $\overrightarrow{E}$, we must also have $\lambda_i = 0$ for all $i, 1 \leq i \leq m$.

Given any element $\langle x, \lambda\rangle \in \widehat{E}$, if

$$x = a_0 + x_1\overrightarrow{a_0a_1} + \cdots + x_m\overrightarrow{a_0a_m}$$

over the affine frame $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ in E, in view of the definition of $\widehat{+}$, we have

$$\begin{aligned}\langle x, \lambda\rangle &= \langle a_0 + x_1\overrightarrow{a_0a_1} + \cdots + x_m\overrightarrow{a_0a_m}, \lambda\rangle \\ &= \langle a_0, \lambda\rangle \mathbin{\widehat{+}} \lambda x_1\overrightarrow{a_0a_1} \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} \lambda x_m\overrightarrow{a_0a_m},\end{aligned}$$

which shows that over the basis $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}, a_0)$ in $\widehat{E}$, the coordinates of $\langle x, \lambda\rangle$ are

$$(\lambda x_1, \ldots, \lambda x_m, \lambda).$$

$\square$

If $(x_1, \ldots, x_m)$ are the coordinates of x w.r.t. the affine frame $(a_0, (\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m}))$ in E, then $(x_1, \ldots, x_m, 1)$ are the coordinates of x in $\widehat{E}$, i.e., the last coordinate is 1, and if u has coordinates $(u_1, \ldots, u_m)$ with respect to the basis $(\overrightarrow{a_0a_1}, \ldots, \overrightarrow{a_0a_m})$ in $\overrightarrow{E}$, then u has coordinates $(u_1, \ldots, u_m, 0)$ in $\widehat{E}$, i.e., the last coordinate is 0. Figure 25.3 shows the affine frame (a_0, a_1, a_2) in E viewed as a basis in $\widehat{E}$.

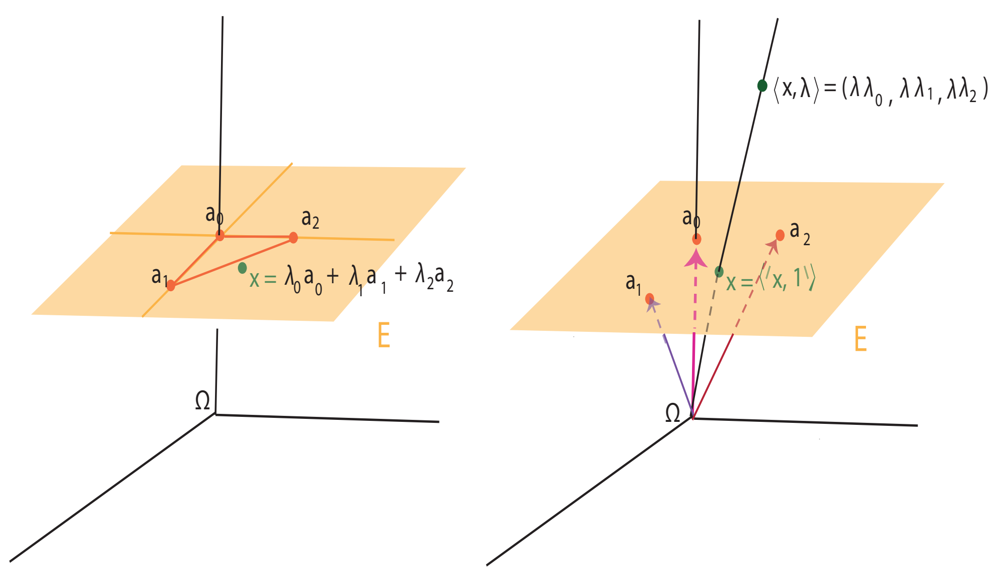


Figure 25.3: The basis (a_0, a_1, a_2) in $\widehat{E}$.

Now that we have defined $\widehat{E}$ and investigated the relationship between affine frames in E and bases in $\widehat{E}$, we can give another construction of a vector space $\mathcal{F}$ from E and $\overrightarrow{E}$ that will allow us to "visualize" in a much more intuitive fashion the structure of $\widehat{E}$ and of its operations $\widehat{+}$ and $\cdot$.

25.3 Another Construction of $\widehat{E}$

One would probably wish that we could start with this construction of $\mathcal{F}$ first, and then define $\widehat{E}$ using the isomorphism $\widehat{\Omega}\colon \widehat{E} \to \mathcal{F}$ defined below. Unfortunately, we first need the vector space structure on $\widehat{E}$ to show that $\widehat{\Omega}$ is linear!

Definition 25.1. Given any affine space $\big(E, \overrightarrow{E}\big)$, we define the vector space $\mathcal{F}$ as the direct sum $\overrightarrow{E} \oplus \mathbb{R}$, where $\mathbb{R}$ denotes the field $\mathbb{R}$ considered as a vector space (over itself). Denoting the unit vector in $\mathbb{R}$ by 1, since $\mathcal{F} = \overrightarrow{E} \oplus \mathbb{R}$, every vector $v \in \mathcal{F}$ can be written as $v = u + \lambda 1$, for some unique $u \in \overrightarrow{E}$ and some unique $\lambda \in \mathbb{R}$. Then, for any choice of an origin Ω_1 in E, we define the map $\widehat{\Omega}\colon \widehat{E} \to \mathcal{F}$, as follows:

$$\widehat{\Omega}(\theta) = \begin{cases} \lambda(1 + \overrightarrow{\Omega_1 a}) & \text{if } \theta = \langle a, \lambda\rangle, \text{ where } a \in E \text{ and } \lambda \neq 0; \\ u & \text{if } \theta = u, \text{ where } u \in \overrightarrow{E}. \end{cases}$$

The idea is that, once again, viewing $\mathcal{F}$ as an affine space under its canonical structure, E is embedded in $\mathcal{F}$ as the hyperplane $H = 1 + \overrightarrow{E}$, with direction $\overrightarrow{E}$, the hyperplane $\overrightarrow{E}$ in $\mathcal{F}$. Then, every point $a \in E$ is in bijection with the point $A = 1 + \overrightarrow{\Omega_1 a}$, in the hyperplane H. If we denote the origin 0 of the canonical affine space $\mathcal{F}$ by Ω, the map $\widehat{\Omega}$ maps a point $\langle a, \lambda\rangle \in E$ to a point in $\mathcal{F}$, as follows: $\widehat{\Omega}(\langle a, \lambda\rangle)$ is the point on the line passing through both the origin Ω of $\mathcal{F}$ and the point $A = 1 + \overrightarrow{\Omega_1 a}$ in the hyperplane $H = 1 + \overrightarrow{E}$, such that

$$\widehat{\Omega}(\langle a, \lambda\rangle) = \lambda\overrightarrow{\Omega A} = \lambda(1 + \overrightarrow{\Omega_1 a}).$$

The following proposition shows that $\widehat{\Omega}$ is an isomorphism of vector spaces.

Proposition 25.4. *Given any affine space $(E, \overrightarrow{E})$, for any choice Ω_1 of an origin in E, the map $\widehat{\Omega}\colon \widehat{E} \to \mathcal{F}$ is a linear isomorphism between $\widehat{E}$ and the vector space $\mathcal{F}$ of Definition 25.1. The inverse of $\widehat{\Omega}$ is given by*

$$\widehat{\Omega}^{-1}(u + \lambda 1) = \begin{cases} \langle \Omega_1 + \lambda^{-1}u, \lambda\rangle) & \text{if } \lambda \neq 0; \\ u & \text{if } \lambda = 0. \end{cases}$$

Proof. It is a straightforward verification. We check that $\widehat{\Omega}$ is invertible, leaving the verification that it is linear as an exercise. We have

$$\langle a, \lambda\rangle \mapsto \lambda 1 + \lambda\overrightarrow{\Omega_1 a} \mapsto \langle \Omega_1 + \overrightarrow{\Omega_1 a}, \lambda\rangle = \langle a, \lambda\rangle$$

and

$$u + \lambda 1 \mapsto \langle \Omega_1 + \lambda^{-1} u, \lambda \rangle \mapsto u + \lambda 1,$$

and since $\widehat{\Omega}$ is the identity on $\overrightarrow{E}$, we have shown that $\widehat{\Omega} \circ \widehat{\Omega}^{-1} = \mathrm{id}$, and $\widehat{\Omega}^{-1} \circ \widehat{\Omega} = \mathrm{id}$. This shows that $\widehat{\Omega}$ is a bijection. $\square$

Figure 25.4 illustrates the embedding of the affine space E into the vector space $\mathcal{F}$, when E is an affine plane.

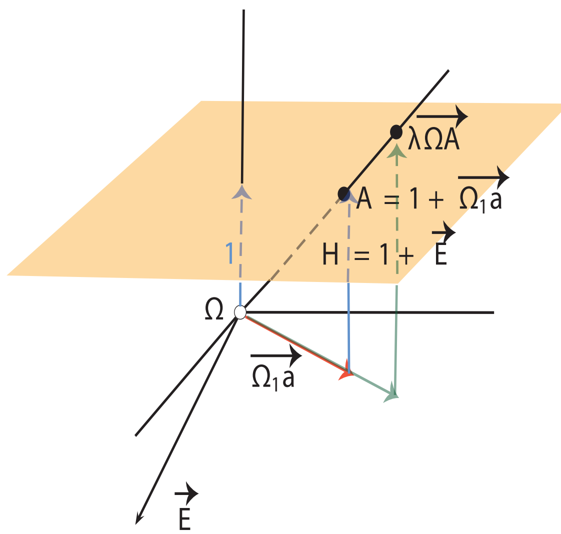


Figure 25.4: Embedding an affine space $(E, \overrightarrow{E})$ into a vector space $\mathcal{F}$.

Proposition 25.4 gives a nice interpretation of the sum operation $\widehat{+}$ of $\widehat{E}$. Given two weighted points $\langle a_1, \lambda_1 \rangle$ and $\langle a_2, \lambda_2 \rangle$, we have

$$\langle a_1, \lambda_1 \rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2 \rangle = \widehat{\Omega}^{-1}(\widehat{\Omega}(\langle a_1, \lambda_1 \rangle) + \widehat{\Omega}(\langle a_2, \lambda_2 \rangle)).$$

The operation $\widehat{\Omega}(\langle a_1, \lambda_1 \rangle) + \widehat{\Omega}(\langle a_2, \lambda_2 \rangle)$ has a simple geometric interpretation. If $\lambda_1 + \lambda_2 \neq 0$, then find the points M_1 and M_2 on the lines passing through the origin Ω of $\mathcal{F}$ and the points $A_1 = \widehat{\Omega}(a_1)$ and $A_2 = \widehat{\Omega}(a_2)$ in the hyperplane H, such that $\overrightarrow{\Omega M_1} = \lambda_1 \overrightarrow{\Omega A_1}$ and $\overrightarrow{\Omega M_2} = \lambda_2 \overrightarrow{\Omega A_2}$, add the vectors $\overrightarrow{\Omega M_1}$ and $\overrightarrow{\Omega M_2}$, getting a point N such that $\overrightarrow{\Omega N} = \overrightarrow{\Omega M_1} + \overrightarrow{\Omega M_2}$, and consider the intersection G of the line passing through Ω and N with the hyperplane H. Then, G is the barycenter of A_1 and A_2 assigned the weights $\lambda_1/(\lambda_1 + \lambda_2)$ and $\lambda_2/(\lambda_1 + \lambda_2)$, and if $g = \widehat{\Omega}^{-1}(\overrightarrow{\Omega G})$, then $\widehat{\Omega}^{-1}(\overrightarrow{\Omega N}) = \langle g, \lambda_1 + \lambda_2 \rangle$. See Figure 25.5.

Instead of adding the vectors $\overrightarrow{\Omega M_1}$ and $\overrightarrow{\Omega M_2}$, we can take the middle N' of the segment $M_1 M_2$, and G is the intersection of the line passing through Ω and N' with the hyperplane H as illustrated in Figure 25.5.

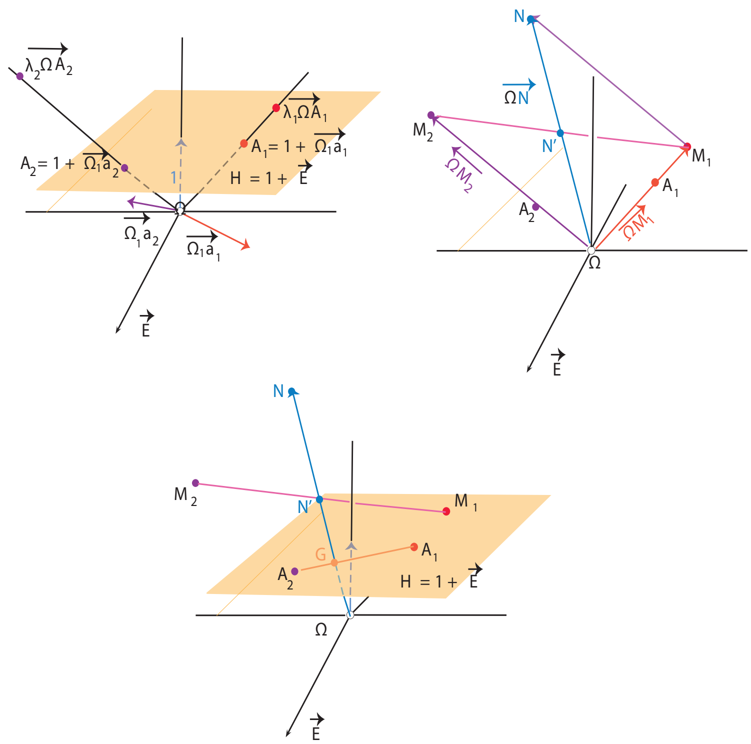


Figure 25.5: The geometric construction of $\widehat{\Omega}(\langle a_1, \lambda_1\rangle) + \widehat{\Omega}(\langle a_2, \lambda_2\rangle)$ for $\lambda_1 + \lambda_2 \neq 0$.

If $\lambda_1 + \lambda_2 = 0$, then $\langle a_1, \lambda_1\rangle \mathbin{\widehat{+}} \langle a_2, \lambda_2\rangle$ is a vector determined as follows. Again, find the points M_1 and M_2 on the lines passing through the origin Ω of $\mathcal{F}$ and the points $A_1 = \widehat{\Omega}(a_1)$ and $A_2 = \widehat{\Omega}(a_2)$ in the hyperplane H, such that $\overrightarrow{\Omega M_1} = \lambda_1 \overrightarrow{\Omega A_1}$ and $\overrightarrow{\Omega M_2} = \lambda_2 \overrightarrow{\Omega A_2}$, and add the vectors $\overrightarrow{\Omega M_1}$ and $\overrightarrow{\Omega M_2}$, getting a point N such that $\overrightarrow{\Omega N} = \overrightarrow{\Omega M_1} + \overrightarrow{\Omega M_2}$. The desired vector is $\overrightarrow{\Omega N}$, which is parallel to the line A_1A_2. Equivalently, let N' be the middle of the segment M_1M_2, and the desired vector is $2\overrightarrow{\Omega N'}$. See Figure 25.6.

We can also give a geometric interpretation of $\langle a, \lambda\rangle + u$. Let $A = \widehat{\Omega}(a)$ in the hyperplane H, let D be the line determined by A and u, let M_1 be the point such that $\overrightarrow{\Omega M_1} = \lambda \overrightarrow{\Omega A}$, and let M_2 be the point such that $\overrightarrow{\Omega M_2} = u$, that is, $M_2 = \Omega + u$. By construction, the line D is in the hyperplane H, and it is parallel to $\overrightarrow{\Omega M_2}$, so that D, M_1, and M_2 are coplanar. Then, add the vectors $\overrightarrow{\Omega M_1}$ and $\overrightarrow{\Omega M_2}$, getting a point N such that $\overrightarrow{\Omega N} = \overrightarrow{\Omega M_1} + \overrightarrow{\Omega M_2}$, and let G be the intersection of the line determined by Ω and N with the line D. If $g = \widehat{\Omega}^{-1}(\overrightarrow{\Omega G})$, then, $\widehat{\Omega}^{-1}(\overrightarrow{\Omega N}) = \langle g, \lambda\rangle$. Equivalently, if N' is the middle of the segment M_1M_2, then G is the intersection of the line determined by Ω and N', with the line D; see Figure 25.7.

We now consider the universal property of $\widehat{E}$ mentioned at the beginning of this section.

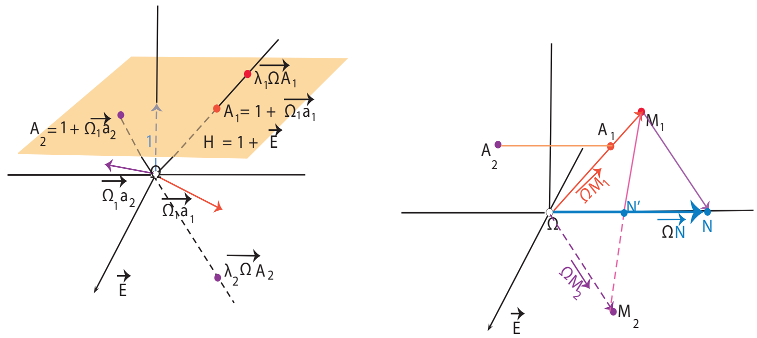


Figure 25.6: The geometric construction of $\widehat{\Omega}(\langle a_1, \lambda_1\rangle) + \widehat{\Omega}(\langle a_2, \lambda_2\rangle)$ for $\lambda_1 + \lambda_2 = 0$.

25.4 Extending Affine Maps to Linear Maps

Roughly, the vector space $\widehat{E}$ has the property that for any vector space $\overrightarrow{F}$ and any affine map $f\colon E \to \overrightarrow{F}$, there is a unique linear map $\widehat{f}\colon \widehat{E} \to \overrightarrow{F}$ extending $f\colon E \to \overrightarrow{F}$. As a consequence, given two affine spaces E and F, every affine map $f\colon E \to F$ extends uniquely to a linear map $\widehat{f}\colon \widehat{E} \to \widehat{F}$. First, we define rigorously the notion of homogenization of an affine space.

Definition 25.2. Given any affine space $\bigl(E, \overrightarrow{E}\bigr)$, a *homogenization (or linearization) of* $(E, \overrightarrow{E})$ is a triple $\langle \mathcal{E}, j, \omega\rangle$, where $\mathcal{E}$ is a vector space, $j\colon E \to \mathcal{E}$ is an injective affine map with associated injective linear map $i\colon \overrightarrow{E} \to \mathcal{E}$, $\omega\colon \mathcal{E} \to \mathbb{R}$ is a linear form such that $\omega^{-1}(0) = i\bigl(\overrightarrow{E}\bigr)$, $\omega^{-1}(1) = j(E)$, and for every vector space $\overrightarrow{F}$ and every affine map $f\colon E \to \overrightarrow{F}$ there is a unique linear map $\widehat{f}\colon \mathcal{E} \to \overrightarrow{F}$ extending f, i.e., $f = \widehat{f} \circ j$, as in the following diagram:

$$\begin{tikzcd} E \arrow[r, "j"] \arrow[dr, "f"'] & \mathcal{E} \arrow[d, "\widehat{f}"] \\ & \overrightarrow{F} \end{tikzcd}$$

Thus, $j(E) = \omega^{-1}(1)$ is an affine hyperplane with direction $i\bigl(\overrightarrow{E}\bigr) = \omega^{-1}(0)$. Note that we could have defined a homogenization of an affine space $(E, \overrightarrow{E})$, as a triple $\langle \mathcal{E}, j, H\rangle$, where $\mathcal{E}$ is a vector space, H is an affine hyperplane in $\mathcal{E}$, and $j\colon E \to \mathcal{E}$ is an injective affine map such that $j(E) = H$, and such that the universal property stated above holds. However, Definition 25.2 is more convenient for our purposes, since it makes the notion of weight more evident.

The obvious candidate for $\mathcal{E}$ is the vector space $\widehat{E}$ that we just constructed. The next proposition will show that $\widehat{E}$ indeed has the required extension property. As usual, objects

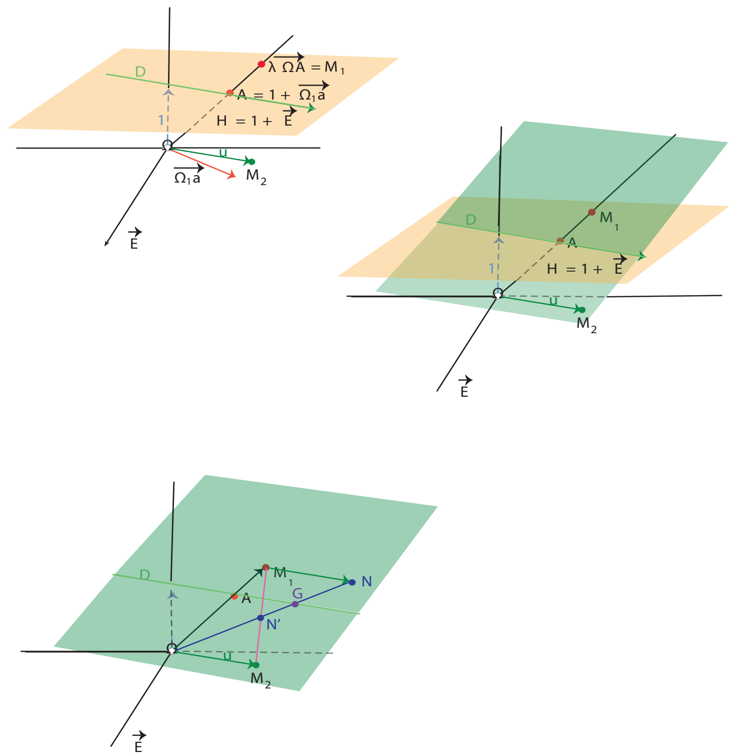


Figure 25.7: The geometric construction of $\langle a, \lambda \rangle + u$.

defined by a universal property are unique up to isomorphism. This property is left as an exercise.

Proposition 25.5. *Given any affine space $\big(E, \overrightarrow{E}\big)$ and any vector space $\overrightarrow{F}$, for any affine map $f \colon E \to \overrightarrow{F}$, there is a unique linear map $\widehat{f} \colon \widehat{E} \to \overrightarrow{F}$ extending f such that*

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a) + \overrightarrow{f}(u)$$

for all $a \in E$, all $u \in \overrightarrow{E}$, and all $\lambda \in \mathbb{R}$, where $\overrightarrow{f}$ is the linear map associated with f. In particular, when $\lambda \neq 0$, we have

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a + \lambda^{-1} u).$$

Proof. Assuming that $\widehat{f}$ exists, recall that from Proposition 25.1, for every $a \in E$, every element of $\widehat{E}$ can be written uniquely as $u \mathbin{\widehat{+}} \lambda a$. By linearity of $\widehat{f}$ and since $\widehat{f}$ extends f, we have

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \widehat{f}(u) + \lambda \widehat{f}(a) = \widehat{f}(u) + \lambda f(a) = \lambda f(a) + \widehat{f}(u).$$

If $\lambda = 1$, since $a \mathbin{\widehat{+}} u$ and $a + u$ are identified, and since $\widehat{f}$ extends f, we must have

$$f(a) + \widehat{f}(u) = \widehat{f}(a) + \widehat{f}(u) = \widehat{f}(a \mathbin{\widehat{+}} u) = f(a + u) = f(a) + \overrightarrow{f}(u),$$

and thus $\widehat{f}(u) = \overrightarrow{f}(u)$ for all $u \in \overrightarrow{E}$. Then we have

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a) + \overrightarrow{f}(u),$$

which proves the uniqueness of $\widehat{f}$. On the other hand, the map $\widehat{f}$ defined as above is clearly a linear map extending f.

When $\lambda \neq 0$, we have

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \widehat{f}(\lambda(a + \lambda^{-1}u)) = \lambda \widehat{f}(a + \lambda^{-1}u) = \lambda f(a + \lambda^{-1}u).$$

$\square$

Proposition 25.5 shows that $\langle \widehat{E}, j, \omega \rangle$, is a homogenization of $(E, \overrightarrow{E})$. As a corollary, we obtain the following proposition.

Proposition 25.6. *Given two affine spaces E and F and an affine map $f \colon E \to F$, there is a unique linear map $\widehat{f} \colon \widehat{E} \to \widehat{F}$ extending f, as in the diagram below,*

$$\begin{tikzcd} E \arrow[r, "f"] \arrow[d, "j"'] & F \arrow[d, "j"] \\ \widehat{E} \arrow[r, "\widehat{f}"'] & \widehat{F} \end{tikzcd}$$

such that

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \overrightarrow{f}(u) \mathbin{\widehat{+}} \lambda f(a),$$

for all $a \in E$, all $u \in \overrightarrow{E}$, and all $\lambda \in \mathbb{R}$, where $\overrightarrow{f}$ is the linear map associated with f. In particular, when $\lambda \neq 0$, we have

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a + \lambda^{-1}u).$$

Proof. Consider the vector space $\widehat{F}$ and the affine map $j \circ f \colon E \to \widehat{F}$. By Proposition 25.5, there is a unique linear map $\widehat{f} \colon \widehat{E} \to \widehat{F}$ extending $j \circ f$, and thus extending f. $\square$

Note that $\widehat{f} \colon \widehat{E} \to \widehat{F}$ has the property that $\widehat{f}(\overrightarrow{E}) \subseteq \overrightarrow{F}$. More generally, since

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \overrightarrow{f}(u) \mathbin{\widehat{+}} \lambda f(a),$$

the linear map $\widehat{f}$ is weight-preserving. Also observe that we recover f from $\widehat{f}$, by letting $\lambda = 1$ in $\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a + \lambda^{-1}u)$, that is, we have

$$f(a + u) = \widehat{f}(u \mathbin{\widehat{+}} a).$$

From a practical point of view, Proposition 25.6 shows us how to homogenize an affine map to turn it into a linear map between the two homogenized spaces. Assume that E and F are of finite dimension, that $(a_0, (u_1, \ldots, u_n))$ is an affine frame of E with origin a_0, and $(b_0, (v_1, \ldots, v_m))$ is an affine frame of F with origin b_0. Then, with respect to the two bases $(u_1, \ldots, u_n, a_0)$ in $\widehat{E}$ and $(v_1, \ldots, v_m, b_0)$ in $\widehat{F}$, a linear map $h \colon \widehat{E} \to \widehat{F}$ is given by an $(m+1) \times (n+1)$ matrix A. Assume that this linear map h is equal to the homogenized version $\widehat{f}$ of an affine map f. Since

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \overrightarrow{f}(u) \mathbin{\widehat{+}} \lambda f(a),$$

and since over the basis $(u_1, \ldots, u_n, a_0)$ in $\widehat{E}$, points are represented by vectors whose last coordinate is 1 and vectors are represented by vectors whose last coordinate is 0, the following properties hold.

1. The last row of the matrix $A = M(\widehat{f})$ with respect to the given bases is
$$(0, 0, \ldots, 0, 1)$$
with n occurrences of 0.

2. The last column of A contains the coordinates
$$(\mu_1, \ldots, \mu_m, 1)$$
of $f(a_0)$ with respect to the basis $(v_1, \ldots, v_m, b_0)$.

3. The submatrix of A obtained by deleting the last row and the last column is the matrix of the linear map $\overrightarrow{f}$ with respect to the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$,

Finally, since
$$f(a_0 + u) = \widehat{f}(u \mathbin{\widehat{+}} a_0),$$
given any $x \in E$ and $y \in F$ with coordinates $(x_1, \ldots, x_n, 1)$ and $(y_1, \ldots, y_m, 1)$, for $X = (x_1, \ldots, x_n, 1)^\top$ and $Y = (y_1, \ldots, y_m, 1)^\top$, we have $y = f(x)$ iff

$$Y = AX.$$

For example, consider the following affine map $f \colon \mathbb{A}^2 \to \mathbb{A}^2$ defined as follows:

$$\begin{aligned} y_1 &= ax_1 + bx_2 + \mu_1, \\ y_2 &= cx_1 + dx_2 + \mu_2. \end{aligned}$$

The matrix of $\widehat{f}$ is

$$\begin{pmatrix} a & b & \mu_1 \\ c & d & \mu_2 \\ 0 & 0 & 1 \end{pmatrix},$$

and we have

$$\begin{pmatrix} y_1 \\ y_2 \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & \mu_1 \\ c & d & \mu_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}.$$

In $\widehat{E}$, we have

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} a & b & \mu_1 \\ c & d & \mu_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

which means that the homogeneous map $\widehat{f}$ is is obtained from f by "adding the variable of homogeneity x_3:"

$$\begin{aligned} y_1 &= ax_1 + bx_2 + \mu_1 x_3, \\ y_2 &= cx_1 + dx_2 + \mu_2 x_3, \\ y_3 &= x_3. \end{aligned}$$

Chapter 26

Basics of Projective Geometry

Think geometrically, prove algebraically.
—John Tate

26.1 Why Projective Spaces?

For a novice, projective geometry usually appears to be a bit odd, and it is not obvious to motivate why its introduction is inevitable and in fact fruitful. One of the main motivations arises from algebraic geometry.

The main goal of algebraic geometry is to study the properties of geometric objects, such as curves and surfaces, defined implicitly in terms of algebraic equations. For instance, the equation

$$x^2 + y^2 - 1 = 0$$

defines a circle in $\mathbb{R}^2$. More generally, we can consider the curves defined by general equations

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

of degree 2, known as *conics*. It is then natural to ask whether it is possible to classify these curves according to their generic geometric shape. This is indeed possible. Except for so-called singular cases, we get ellipses, parabolas, and hyperbolas. The same question can be asked for surfaces defined by quadratic equations, known as *quadrics*, and again, a classification is possible. However, these classifications are a bit artificial. For example, an ellipse and a hyperbola differ by the fact that a hyperbola has points at infinity, and yet, their geometric properties are identical, provided that points at infinity are handled properly.

Another important problem is the study of intersection of geometric objects (defined algebraically). For example, given two curves C_1 and C_2 of degree m and n, respectively, what is the number of intersection points of C_1 and C_2? (by degree of the curve we mean the total degree of the defining polynomial).

Well, it depends! Even in the case of lines (when $m = n = 1$), there are three possibilities: either the lines coincide, or they are parallel, or there is a single intersection point. In general, we expect mn intersection points, but some of these points may be missing because they are at infinity, because they coincide, or because they are imaginary.

What begins to transpire is that "points at infinity" cause trouble. They cause exceptions that invalidate geometric theorems (for example, consider the more general versions of the theorems of Pappus and Desargues), and make it difficult to classify geometric objects. Projective geometry is designed to deal with "points at infinity" and regular points in a uniform way, without making a distinction. Points at infinity are now just ordinary points, and many things become simpler. For example, the classification of conics and quadrics becomes simpler, and intersection theory becomes cleaner (although, to be honest, we need to consider complex projective spaces).

Technically, projective geometry can be defined axiomatically, or by building upon linear algebra. Historically, the axiomatic approach came first (see Veblen and Young [183, 184], Emil Artin [6], and Coxeter [45, 46, 43, 44]). Although very beautiful and elegant, we believe that it is a harder approach than the linear algebraic approach. In the linear algebraic approach, all notions are considered up to a scalar. For example, a projective point is really a line through the origin. In terms of coordinates, this corresponds to "homogenizing." For example, the homogeneous equation of a conic is

$$ax^2 + by^2 + cxy + dxz + eyz + fz^2 = 0.$$

Now, regular points are points of coordinates (x, y, z) with $z \neq 0$, and points at infinity are points of coordinates $(x, y, 0)$ (with x, y, z not all null, and up to a scalar). There is a useful model (interpretation) of plane projective geometry in terms of the central projection in $\mathbb{R}^3$ from the origin onto the plane $z = 1$. Another useful model is the spherical (or the half-spherical) model. In the spherical model, a projective point corresponds to a pair of antipodal points on the sphere.

As affine geometry is the study of properties invariant under affine bijections, projective geometry is the study of properties invariant under bijective projective maps. Roughly speaking, projective maps are linear maps up to a scalar. In analogy with our presentation of affine geometry, we will define projective spaces, projective subspaces, projective frames, and projective maps. The analogy will fade away when we define the projective completion of an affine space, and when we define duality.

One of the virtues of projective geometry is that it yields a very clean presentation of rational curves and rational surfaces. The general idea is that a plane rational curve is the projection of a simpler curve in a larger space, a polynomial curve in $\mathbb{R}^3$, onto the plane $z = 1$, as we now explain.

Polynomial curves are curves defined parametrically in terms of polynomials. More specifically, if $\mathcal{E}$ is an affine space of finite dimension $n \geq 2$ and $(a_0, (e_1, \ldots, e_n))$ is an affine frame for $\mathcal{E}$, a polynomial curve of degree m is a map $F\colon \mathbb{A} \to \mathcal{E}$ such that

$$F(t) = a_0 + F_1(t)e_1 + \cdots + F_n(t)e_n,$$

for all $t \in \mathbb{A}$, where $F_1(t), \ldots, F_n(t)$ are polynomials of degree at most m.

Although many curves can be defined, it is somewhat embarassing that a circle cannot be defined in such a way. In fact, many interesting curves cannot be defined this way, for example, ellipses and hyperbolas. A rather simple way to extend the class of curves defined parametrically is to allow rational functions instead of polynomials. A *parametric rational curve* of degree m is a function $F\colon \mathbb{A} \to \mathcal{E}$ such that

$$F(t) = a_0 + \frac{F_1(t)}{F_{n+1}(t)}e_1 + \cdots + \frac{F_n(t)}{F_{n+1}(t)}e_n,$$

for all $t \in \mathbb{A}$, where $F_1(t), \ldots, F_n(t), F_{n+1}(t)$ are polynomials of degree at most m. For example, a circle in $\mathbb{A}^2$ can be defined by the rational map

$$F(t) = a_0 + \frac{1-t^2}{1+t^2}e_1 + \frac{2t}{1+t^2}e_2.$$

In terms of coordinates, the above curve is given by

$$\begin{aligned} x &= \frac{1-t^2}{1+t^2} \\ y &= \frac{2t}{1+t^2}, \end{aligned}$$

and it is easily checked that $x^2 + y^2 = 1$. Note that the point $(-1, 0)$ is not achieved for any finite value of t, but it is for $t = \infty$.

In the above example, the denominator $F_3(t) = 1 + t^2$ never takes the value 0 when t ranges over $\mathbb{A}$, but consider the following curve in $\mathbb{A}^2$:

$$G(t) = a_0 + \frac{t^2}{t}e_1 + \frac{1}{t}e_2.$$

Observe that $G(0)$ is undefined. In terms of coordinates, the above curve is given by

$$\begin{aligned} x &= \frac{t^2}{t} = t \\ y &= \frac{1}{t}, \end{aligned}$$

so we have $y = 1/x$. The curve defined above is a hyperbola, and for t close to 0, the point on the curve goes toward infinity in one of the two asymptotic directions.

A clean way to handle the situation in which the denominator vanishes is to work in a projective space. Intuitively, this means viewing a rational curve in $\mathbb{A}^n$ as some appropriate projection of a polynomial curve in $\mathbb{A}^{n+1}$, back onto $\mathbb{A}^n$.

Given an affine space $\mathcal{E}$, for any hyperplane H in $\mathcal{E}$ and any point a_0 not in H, the *central projection (or conic projection, or perspective projection) of center a_0 onto H*, is the partial map p defined as follows: For every point x not in the hyperplane passing through a_0 and parallel to H, we define $p(x)$ as the intersection of the line defined by a_0 and x with the hyperplane H; see Figure 26.1.

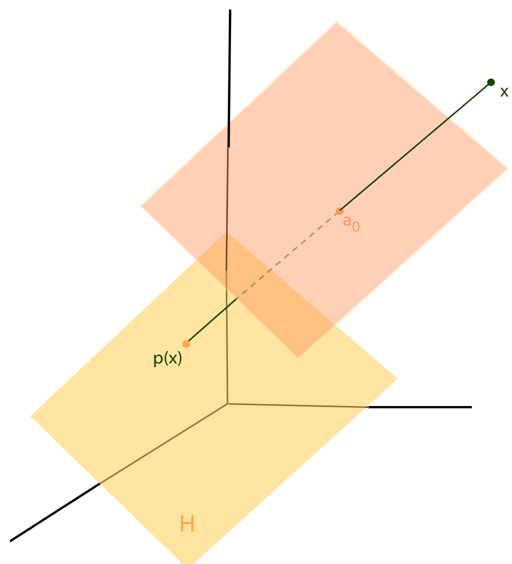


Figure 26.1: A central projection in $\mathbb{A}^3$ through a_0 onto the yellow hyperplane H. This central projection is not defined for any points in the peach hyperplane.

For example, we can view G as a rational curve in $\mathbb{A}^3$ given by

$$G_1(t) = a_0 + t^2 e_1 + e_2 + t e_3.$$

If we project this curve G_1 (in fact, a parabola in $\mathbb{A}^3$) using the central projection (perspective projection) of center a_0 onto the plane of equation $x_3 = 1$, we get the previous hyperbola; see Figure 26.2. For $t = 0$, the point $G_1(0) = a_0 + e_2$ in $\mathbb{A}^3$ is in the plane of equation $x_3 = 0$, and its projection is undefined. We can consider that $G_1(0) = a_0 + e_2$ in $\mathbb{A}^3$ is projected to infinity in the direction of e_2 in the plane $x_3 = 0$. In the setting of projective spaces, this direction corresponds rigorously to a point at infinity; see Figure 26.2.

Let us verify that the central projection used in the previous example has the desired effect. Let us assume that $\mathcal{E}$ has dimension $n + 1$ and that $(a_0, (e_1, \ldots, e_{n+1}))$ is an affine

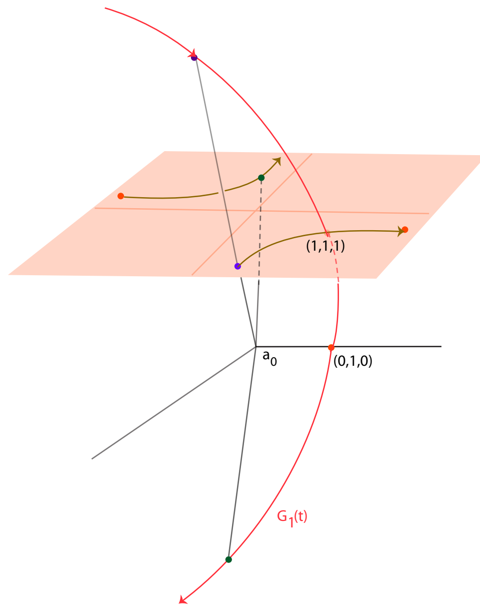


Figure 26.2: A central projection in $\mathbb{A}^3$ through a_0 of the parabola $G_1(t)$ onto the hyperplane $x_3 = 1$.

frame for $\mathcal{E}$. We want to determine the coordinates of the central projection $p(x)$ of a point $x \in \mathcal{E}$ onto the hyperplane H of equation $x_{n+1} = 1$ (the center of projection being a_0). If

$$x = a_0 + x_1e_1 + \cdots + x_ne_n + x_{n+1}e_{n+1},$$

assuming that $x_{n+1} \neq 0$; a point on the line passing through a_0 and x has coordinates of the form $(\lambda x_1, \ldots, \lambda x_{n+1})$; and $p(x)$, the central projection of x onto the hyperplane H of equation $x_{n+1} = 1$, is the intersection of the line from a_0 to x and this hyperplane H. Thus we must have $\lambda x_{n+1} = 1$, and the coordinates of $p(x)$ are

$$\left(\frac{x_1}{x_{n+1}}, \ldots, \frac{x_n}{x_{n+1}}, 1\right).$$

Note that $p(x)$ is undefined when $x_{n+1} = 0$. In projective spaces, we can make sense of such points.

The above calculation confirms that $G(t)$ is a central projection of $G_1(t)$. Similarly, if we define the curve F_1 in $\mathbb{A}^3$ by

$$F_1(t) = a_0 + (1 - t^2)e_1 + 2te_2 + (1 + t^2)e_3,$$

the central projection of the polynomial curve F_1 (again, a parabola in $\mathbb{A}^3$) onto the plane of equation $x_3 = 1$ is the circle F.

What we just sketched is a general method to deal with rational curves. We can use our "hat construction" to embed an affine space $\mathcal{E}$ into a vector space $\widehat{\mathcal{E}}$ having one more dimension, then construct the projective space $\mathbf{P}\big(\widehat{\mathcal{E}}\big)$. This turns out to be the "projective completion" of the affine space $\mathcal{E}$. Then we can define a rational curve in $\mathbf{P}\big(\widehat{\mathcal{E}}\big)$, basically as the central projection of a polynomial curve in $\widehat{\mathcal{E}}$ back onto $\mathbf{P}\big(\widehat{\mathcal{E}}\big)$. The same approach can be used to deal with rational surfaces. Due to the lack of space, such a presentation is omitted. However, it can be found on the web; see http://www.cis.upenn.edu/~jean/gbooks/geom2.html.

More generally, the projective completion of an affine space is a very convenient tool to handle "points at infinity" in a clean fashion.

This chapter contains a brief presentation of concepts of projective geometry. The following concepts are presented: projective spaces, projective frames, homogeneous coordinates, projective maps, projective hyperplanes, multiprojective maps, affine patches. The projective completion of an affine space is presented using the "hat construction." The theorems of Pappus and Desargues are proved, using the method in which points are "sent to infinity." We also discuss the cross-ratio and duality. The chapter ends with a very brief explanation of the use of the complexification of a projective space in order to define the notion of angle and orthogonality in a projective setting. We also include a short section on applications of projective geometry, notably to computer vision (camera calibration), efficient communication, and error-correcting codes.

26.2 Projective Spaces

As in the case of affine geometry, our presentation of projective geometry is rather sketchy. For a systematic treatment of projective geometry, we recommend Berger [11, 12], Samuel [142], Pedoe [136], Coxeter [45, 46, 43, 44], Beutelspacher and Rosenbaum [22], Fresnel [65], Sidler [161], Tisseron [175], Lehmann and Bkouche [115], Vienne [185], and the classical treatise by Veblen and Young [183, 184], which, although slightly old-fashioned, is definitely worth reading. Emil Artin's famous book [6] contains, among other things, an axiomatic presentation of projective geometry, and a wealth of geometric material presented from an algebraic point of view. Other "oldies but goodies" include the beautiful books by Darboux [47] and Klein [103]. For a development of projective geometry addressing the delicate problem of orientation, see Stolfi [167], and for an approach geared towards computer graphics, see Penna and Patterson [137].

First, we define projective spaces, allowing the field K to be arbitrary (which does no harm, and is needed to allow finite and complex projective spaces). Roughly speaking, every projective concept is a linear–algebraic concept "up to a scalar." For spaces, this is made precise as follows.

Definition 26.1. Given a vector space E over a field K, the *projective space* $\mathbf{P}(E)$ *induced by* E is the set $(E - \{0\})/\sim$ of equivalence classes of nonzero vectors in E under the equivalence relation $\sim$ defined such that for all $u, v \in E - \{0\}$,

$$u \sim v \quad \text{iff} \quad v = \lambda u, \text{ for some } \lambda \in K - \{0\}.$$

The *canonical projection* $p\colon (E - \{0\}) \to \mathbf{P}(E)$ is the function associating the equivalence class $[u]_\sim$ modulo $\sim$ to $u \neq 0$. The *dimension* $\dim(\mathbf{P}(E))$ of $\mathbf{P}(E)$ is defined as follows: If E is of infinite dimension, then $\dim(\mathbf{P}(E)) = \dim(E)$, and if E has finite dimension, $\dim(E) = n \geq 1$ then $\dim(\mathbf{P}(E)) = n - 1$.

Mathematically, a projective space $\mathbf{P}(E)$ is a set of equivalence classes of vectors in E. The spirit of projective geometry is to view an equivalence class $p(u) = [u]_\sim$ as an "atomic" object, forgetting the internal structure of the equivalence class. For this reason, it is customary to call an equivalence class $a = [u]_\sim$ a *point* (the entire equivalence class $[u]_\sim$ is collapsed into a single object viewed as a point).

Remarks:

(1) If we view E as an affine space, then for any nonnull vector $u \in E$, since

$$[u]_\sim = \{\lambda u \mid \lambda \in K,\, \lambda \neq 0\},$$

letting

$$Ku = \{\lambda u \mid \lambda \in K\}$$

denote the subspace of dimension 1 spanned by u, the map

$$[u]_\sim \mapsto Ku$$

from $\mathbf{P}(E)$ to the set of one-dimensional subspaces of E is clearly a bijection, and since subspaces of dimension 1 correspond to lines through the origin in E, we can view $\mathbf{P}(E)$ as the set of lines in E passing through the origin. So, the projective space $\mathbf{P}(E)$ can be viewed as the set obtained from E when lines through the origin are treated as points.

However, this is a somewhat deceptive view. Indeed, depending on the structure of the vector space E, a line (through the origin) in E may be a fairly complex object, and treating a line just as a point is really a mental game. For example, E may be the vector space of real homogeneous polynomials $P(x, y, z)$ of degree 2 in three variables x, y, z (plus the null polynomial), and a "line" (through the origin) in E corresponds to an algebraic curve of degree 2. Lots of details need to be filled in, but roughly speaking, the curve defined by P is the "zero locus of P," i.e., the set of points $(x, y, z) \in \mathbf{P}(\mathbb{R}^3)$ (or perhaps in $\mathbf{P}(\mathbb{C}^3)$) for which $P(x, y, z) = 0$. We will come back to this point in Section 26.4 after having introduced homogeneous coordinates.

More generally, E may be a vector space of homogeneous polynomials of degree m in 3 or more variables (plus the null polynomial), and the lines in E correspond to such objects as algebraic curves, algebraic surfaces, and algebraic varieties. The point of view where a complex object such as a curve or a surface is treated as a point in a (projective) space is actually very fruitful and is one of the themes of algebraic geometry (see Fulton [66] or Harris [87]).

(2) When $\dim(E) = 1$, we have $\dim(\mathbf{P}(E)) = 0$. When $E = \{0\}$, we have $\mathbf{P}(E) = \emptyset$. By convention, we give it the dimension -1.

We denote the projective space $\mathbf{P}(K^{n+1})$ by $\mathbb{P}^n_K$. When $K = \mathbb{R}$, we also denote $\mathbb{P}^n_{\mathbb{R}}$ by $\mathbb{RP}^n$, and when $K = \mathbb{C}$, we denote $\mathbb{P}^n_{\mathbb{C}}$ by $\mathbb{CP}^n$. The projective space $\mathbb{P}^0_K$ is a (projective) point. The projective space $\mathbb{P}^1_K$ is called a *projective line*. The projective space $\mathbb{P}^2_K$ is called a *projective plane*.

The projective space $\mathbf{P}(E)$ can be visualized in the following way. For simplicity, assume that $E = \mathbb{R}^{n+1}$, and thus $\mathbf{P}(E) = \mathbb{RP}^n$ (the same reasoning applies to $E = K^{n+1}$, where K is any field).

Let H be the affine hyperplane consisting of all points $(x_1, \ldots, x_{n+1})$ such that $x_{n+1} = 1$. Every nonzero vector u in E determines a line D passing through the origin, and this line intersects the hyperplane H in a unique point a, unless D is parallel to H. When D is parallel to H, the line corresponding to the equivalence class of u can be thought of as a point at infinity, often denoted by u_∞. Thus, the projective space $\mathbf{P}(E)$ can be viewed as the set of points in the hyperplane H, together with points at infinity associated with lines in the hyperplane H_∞ of equation $x_{n+1} = 0$. We will come back to this point of view when we consider the projective completion of an affine space. Figure 26.3 illustrates the above representation of the projective space for $E = \mathbb{R}^2$ and $E = \mathbb{R}^3$.

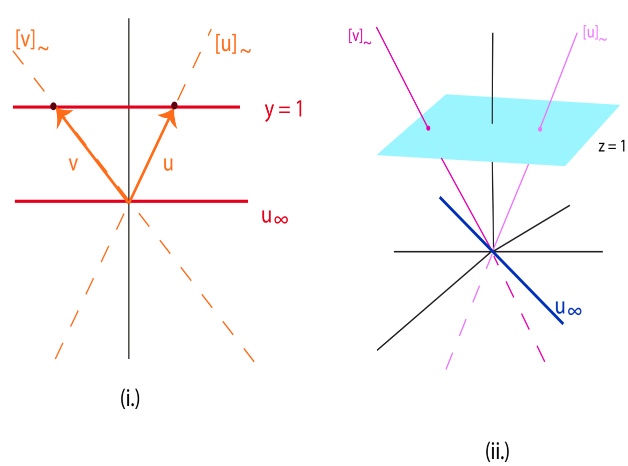


Figure 26.3: The hyperplane model representations of $\mathbb{RP}^1$ and $\mathbb{RP}^2$.

We refer to the above model of $\mathbf{P}(E)$ as the *hyperplane model*. In this model some hyperplane H_∞ (through the origin) in $\mathbb{R}^{n+1}$ is singled out, and the points of $\mathbf{P}(E)$ arising from the hyperplane H_∞ are declared to be "points at infinity." The purpose of the affine hyperplane H parallel to H_∞ and distinct from H_∞ is to get images for the other points in $\mathbf{P}(E)$ (i.e., those that arise from lines not contained in H_∞). It should be noted that the choice of which points should be considered as infinite is relative to the choice of H_∞. Viewing certain points of $\mathbf{P}(E)$ as points at infinity is convenient for getting a mental picture of $\mathbf{P}(E)$, but there is nothing intrinsic about that. Points of $\mathbf{P}(E)$ are all equal, and unless some additional structure in introduced in $\mathbf{P}(E)$ (such as a hyperplane), a point in $\mathbf{P}(E)$ doesn't know whether it is infinite! The notion of point at infinity is really an affine notion. This point will be made precise in Section 26.8.

Again, for $\mathbb{RP}^n = \mathbf{P}(\mathbb{R}^{n+1})$, instead of considering the hyperplane H, we can consider the n-sphere S^n of center 0 and radius 1, i.e., the set of points $(x_1, \ldots, x_{n+1})$ such that

$$x_1^2 + \cdots + x_n^2 + x_{n+1}^2 = 1.$$

In this case, every line D through the center of the sphere intersects the sphere S^n in two antipodal points a_+ and a_-. The projective space $\mathbb{RP}^n$ is the quotient space obtained from the sphere S^n by identifying antipodal points a_+ and a_-. It is hard to visualize such an object! We call this model of $\mathbf{P}(E)$ the *spherical model*. See Figure 26.4.

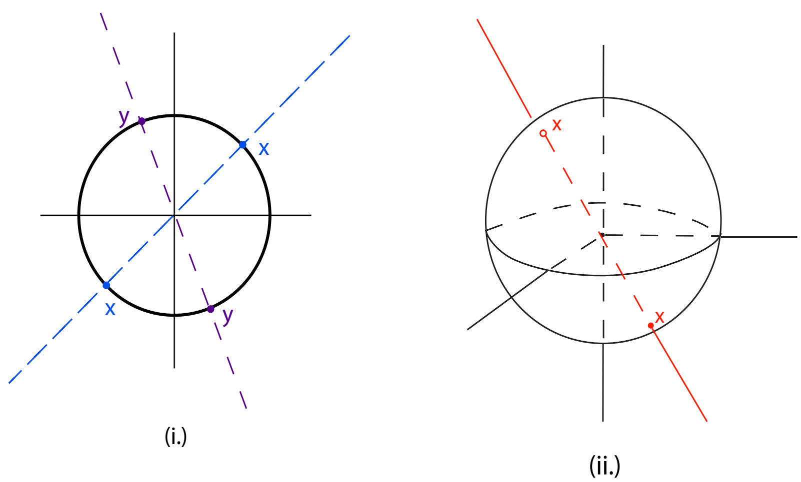


Figure 26.4: The spherical model representations of $\mathbb{RP}^1$ and $\mathbb{RP}^2$.

A more subtle construction consists in considering the (upper) half-sphere instead of the sphere, where the upper half-sphere S^n_+ is set of points on the sphere S^n such that $x_{n+1} \geq 0$. This time, every line through the center intersects the (upper) half-sphere in a single point, except on the boundary of the half-sphere, where it intersects in two antipodal points a_+ and a_-. Thus, the projective space $\mathbb{RP}^n$ is the quotient space obtained from the (upper) half-sphere S^n_+ by identifying antipodal points a_+ and a_- on the boundary of the half-sphere. We call this model of $\mathbf{P}(E)$ the *half-spherical model*; see Figure 26.5.

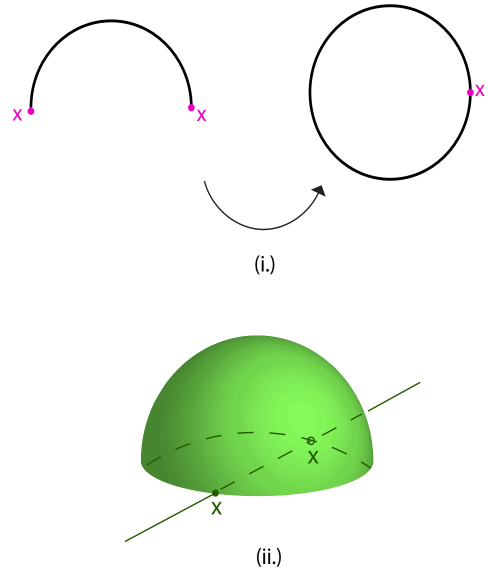


Figure 26.5: The half-spherical model representations of $\mathbb{RP}^1$ and $\mathbb{RP}^2$.

When $n = 2$, we get a circle. When $n = 3$, the upper half-sphere is homeomorphic to a closed disk (say, by orthogonal projection onto the xy-plane), and $\mathbb{RP}^2$ is in bijection with a closed disk in which antipodal points on its boundary (a unit circle) have been identified. This is hard to visualize! In this model of the real projective space, projective lines are great semicircles on the upper half-sphere, with antipodal points on the boundary identified. Boundary points correspond to points at infinity. By orthogonal projection, these great semicircles correspond to semiellipses, with antipodal points on the boundary identified. Traveling along such a projective "line," when we reach a boundary point, we "wrap around"! In general, the upper half-sphere S^n_+ is homeomorphic to the closed unit ball in $\mathbb{R}^n$, whose boundary is the $(n-1)$-sphere S^{n-1}. For example, the projective space $\mathbb{RP}^3$ is in bijection with the closed unit ball in $\mathbb{R}^3$, with antipodal points on its boundary (the sphere S^2) identified!

Remarks:

(1) A projective space $\mathbf{P}(E)$ has been defined as a *set* without any topological structure. When the field K is either the field $\mathbb{R}$ of reals or the field $\mathbb{C}$ of complex numbers, the vector space E is a topological space. Thus, the projection map $p\colon (E - \{0\}) \to \mathbf{P}(E)$ induces a topology on the projective space $\mathbf{P}(E)$, namely the quotient topology. This means that a subset V of $\mathbf{P}(E)$ is open iff $p^{-1}(V)$ is an open set in E. Then, for example, it turns out that the real projective space $\mathbb{RP}^n$ is homeomorphic to the space

obtained by taking the quotient of the (upper) half-sphere S^n_+, by the equivalence relation identifying antipodal points a_+ and a_- on the boundary of the half-sphere. Another interesting fact is that the complex projective line $\mathbb{CP}^1 = \mathbf{P}(\mathbb{C}^2)$ is homeomorphic to the (real) 2-sphere S^2, and that the real projective space $\mathbb{RP}^3$ is homeomorphic to the group of rotations $\mathbf{SO}(3)$ of $\mathbb{R}^3$.

(2) If H is a hyperplane in E, recall from Proposition 11.7 that there is some nonnull linear form $f \in E^*$ such that $H = \mathrm{Ker}\, f$. Also, given any nonnull linear form $f \in E^*$, its kernel $H = \mathrm{Ker}\, f = f^{-1}(0)$ is a hyperplane, and if $\mathrm{Ker}\, f = \mathrm{Ker}\, g = H$, then $g = \lambda f$ for some $\lambda \neq 0$. These facts can be concisely stated by saying that the map

$$[f]_\sim \mapsto \mathrm{Ker}\, f$$

mapping the equivalence class $[f]_\sim = \{\lambda f \mid \lambda \neq 0\}$ of a nonnull linear form $f \in E^*$ to the hyperplane $H = \mathrm{Ker}\, f$ in E is a bijection between the projective space $\mathbf{P}(E^*)$ and the set of hyperplanes in E. When E is of finite dimension, this bijection yields a useful duality, which will be investigated in Section 26.12.

We now define projective subspaces.

26.3 Projective Subspaces

Projective subspaces of a projective space $\mathbf{P}(E)$ are induced by subspaces of the vector space E.

Definition 26.2. Given a nontrivial vector space E, a *projective subspace (or linear projective variety) of* $\mathbf{P}(E)$ is any subset W of $\mathbf{P}(E)$ such that there is some subspace $V \neq \{0\}$ of E with $W = p(V - \{0\})$. The dimension $\dim(W)$ of W is defined as follows: If V is of infinite dimension, then $\dim(W) = \dim(V)$, and if $\dim(V) = p \geq 1$, then $\dim(W) = p - 1$. We say that a family $(a_i)_{i\in I}$ of points of $\mathbf{P}(E)$ is *projectively independent* if there is a linearly independent family $(u_i)_{i\in I}$ in E such that $a_i = p(u_i)$ for every $i \in I$.

Remark: If we allow the empty subset to be a projective subspace, then if assign the empty subset to the trivial subspace $\{0\}$, we obtain a bijection between the subspaces of E and the projective subspaces of $\mathbf{P}(E)$. If $\mathbf{P}(V)$ is the projective space induced by the vector space V, we also denote $p(V - \{0\})$ by $\mathbf{P}(V)$, or even by $p(V)$, even though $p(0)$ is undefined.

A projective subspace of dimension 0 is a called a *(projective) point*. A projective subspace of dimension 1 is called a *(projective) line*, and a projective subspace of dimension 2 is called a *(projective) plane*. If H is a hyperplane in E, then $\mathbf{P}(H)$ is called a *projective hyperplane*. It is easily verified that any arbitrary intersection of projective subspaces is a projective subspace.

A single point is projectively independent. Two points a, b are projectively independent if $a \neq b$. Two distinct points define a (unique) projective line. Three points a, b, c are projectively independent if they are distinct, and neither belongs to the projective line defined by the other two. Three projectively independent points define a (unique) projective plane.

A closer look at projective subspaces will show some of the advantages of projective geometry: In considering intersection properties, there are no exceptions due to parallelism, as in affine spaces.

Let E be a nontrivial vector space. Given any nontrivial subset S of E, the subset S defines a subset $U = p(S - \{0\})$ of the projective space $\mathbf{P}(E)$, and if $\langle S \rangle$ denotes the subspace of E spanned by S, it is immediately verified that $\mathbf{P}(\langle S \rangle)$ is the intersection of all projective subspaces containing U, and this projective subspace is denoted by $\langle U \rangle$. Then $n \geq 2$ point $a_1, \ldots, a_n \in \mathbf{P}(E)$ are projectively independent iff for all $i = 1, \ldots, n$ the point a_i does not belong to the projective subspace $\langle a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n \rangle$ spanned by $\{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\}$.

Given any subspaces M and N of E, recall from Proposition 24.15 that we have the *Grassmann relation*

$$\dim(M) + \dim(N) = \dim(M + N) + \dim(M \cap N).$$

Then the following proposition is easily shown.

Proposition 26.1. *Given a projective space $\mathbf{P}(E)$, for any two projective subspaces U, V of $\mathbf{P}(E)$, we have*

$$\dim(U) + \dim(V) = \dim(\langle U \cup V \rangle) + \dim(U \cap V).$$

Furthermore, if $\dim(U) + \dim(V) \geq \dim(\mathbf{P}(E))$, then $U \cap V$ is nonempty and if $\dim(\mathbf{P}(E)) = n$, then:

(i) The intersection of any n hyperplanes is nonempty.

(ii) For every hyperplane H and every point $a \notin H$, every line D containing a intersects H in a unique point.

(iii) In a projective plane, every two distinct lines intersect in a unique point.

As a corollary, in 3D projective space ($\dim(\mathbf{P}(E)) = 3$), for every plane H, every line not contained in H intersects H in a unique point.

It is often useful to deal with projective hyperplanes in terms of nonnull linear forms and equations. Recall that the map

$$[f]_\sim \mapsto \operatorname{Ker} f$$

is a bijection between $\mathbf{P}(E^*)$ and the set of hyperplanes in E, mapping the equivalence class $[f]_\sim = \{\lambda f \mid \lambda \neq 0\}$ of a nonnull linear form $f \in E^*$ to the hyperplane $H = \mathrm{Ker}\, f$. Furthermore, if $u \sim v$, which means that $u = \lambda v$ for some $\lambda \neq 0$, we have

$$f(u) = 0 \quad \text{iff} \quad f(v) = 0,$$

since $f(v) = \lambda f(u)$ and $\lambda \neq 0$. Thus, there is a bijection

$$\{\lambda f \mid \lambda \neq 0\} \mapsto \mathbf{P}(\mathrm{Ker}\, f)$$

mapping points in $\mathbf{P}(E^*)$ to hyperplanes in $\mathbf{P}(E)$. Any nonnull linear form f associated with some hyperplane $\mathbf{P}(H)$ in the above bijection (i.e., $H = \mathrm{Ker}\, f$) is called an *equation of the projective hyperplane* $\mathbf{P}(H)$. We also say that $f = 0$ *is the equation of the hyperplane* $\mathbf{P}(H)$.

Before ending this section, we give an example of a projective space where lines have a nontrivial geometric interpretation, namely as "pencils of lines." If $E = \mathbb{R}^3$, recall that the dual space E^* is the set of all linear maps $f\colon \mathbb{R}^3 \to \mathbb{R}$. As we have just explained, there is a bijection

$$p(f) \mapsto \mathbf{P}(\mathrm{Ker}\, f)$$

between $\mathbf{P}(E^*)$ and the set of lines in $\mathbf{P}(E)$, mapping every point $a^* = p(f)$ to the line $D_{a^*} = \mathbf{P}(\mathrm{Ker}\, f)$.

Is there a way to give a geometric interpretation in $\mathbf{P}(E)$ of a line Δ in $\mathbf{P}(E^*)$? Well, a line Δ in $\mathbf{P}(E^*)$ is defined by two distinct points $a^* = p(f)$ and $b^* = p(g)$, where $f, g \in E^*$ are two linearly independent linear forms. But f and g define two distinct planes $H_1 = \mathrm{Ker}\, f$ and $H_2 = \mathrm{Ker}\, g$ through the origin (in $E = \mathbb{R}^3$), and H_1 and H_2 define two distinct lines $D_1 = p(H_1)$ and $D_2 = p(H_2)$ in $\mathbf{P}(E)$. The line Δ in $\mathbf{P}(E^*)$ is of the form $\Delta = p(V)$, where

$$V = \{\lambda f + \mu g \mid \lambda, \mu \in \mathbb{R}\}$$

is the plane in E^* spanned by f, g. Every nonnull linear form $\lambda f + \mu g \in V$ defines a plane $H = \mathrm{Ker}\,(\lambda f + \mu g)$ in E, and since H_1 and H_2 (in E) are distinct, they intersect in a line L that is also contained in every plane H as above. Thus, the set of planes in E associated with nonnull linear forms in V is just the set of all planes containing the line L. Passing to $\mathbf{P}(E)$ using the projection p, the line L in E corresponds to the point $c = p(L)$ in $\mathbf{P}(E)$, which is just the intersection of the lines D_1 and D_2. Thus, every point of the line Δ in $\mathbf{P}(E^*)$ corresponds to a line in $\mathbf{P}(E)$ passing through c (the intersection of the lines D_1 and D_2), and this correspondence is bijective.

In summary, a line Δ in $\mathbf{P}(E^*)$ corresponds to the set of all lines in $\mathbf{P}(E)$ through some given point. Such sets of lines are called *pencils of lines* and are illustrated in Figure 26.6.

The above discussion can be generalized to higher dimensions and is discussed quite extensively in Section 26.12. In brief, letting $E = \mathbb{R}^{n+1}$, there is a bijection mapping points in $\mathbf{P}(E^*)$ to hyperplanes in $\mathbf{P}(E)$. A line in $\mathbf{P}(E^*)$ corresponds to a *pencil of hyperplanes* in

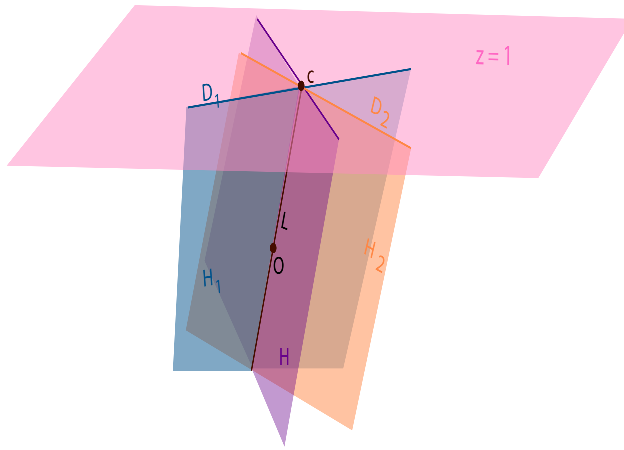


Figure 26.6: A pencil of lines through c in the hyperplane model of $\mathbb{RP}^2$

$\mathbf{P}(E)$, i.e., the set of all hyperplanes containing some given projective subspace $W = p(V)$ of dimension $n-2$. For $n = 3$, a pencil of planes in $\mathbb{RP}^3 = \mathbf{P}(\mathbb{R}^4)$ is the set of all planes (in $\mathbb{RP}^3$) containing some given line W. Other examples of unusual projective spaces and pencils will be given in Section 26.4.

Next, we define the projective analogues of bases (or frames) and linear maps.

26.4 Projective Frames

As all good notions in projective geometry, the concept of a projective frame turns out to be uniquely defined up to a scalar.

Definition 26.3. Given a nontrivial vector space E of dimension $n+1$, a family $(a_i)_{1\leq i\leq n+2}$ of $n+2$ points of the projective space $\mathbf{P}(E)$ is a *projective frame (or basis) of* $\mathbf{P}(E)$ if there exists some basis $(e_1,\ldots,e_{n+1})$ of E such that $a_i = p(e_i)$ for $1 \leq i \leq n+1$, and $a_{n+2} = p(e_1+\cdots+e_{n+1})$. Any basis with the above property is said to be *associated with the projective frame* $(a_i)_{1\leq i\leq n+2}$.

The justification of Definition 26.3 is given by the following proposition.

Proposition 26.2. *If $(a_i)_{1\leq i\leq n+2}$ is a projective frame of $\mathbf{P}(E)$, for any two bases $(u_1,\ldots,u_{n+1})$, $(v_1,\ldots,v_{n+1})$ of E such that $a_i = p(u_i) = p(v_i)$ for $1 \leq i \leq n+1$, and $a_{n+2} = p(u_1+\cdots+u_{n+1}) = p(v_1+\cdots+v_{n+1})$, there is a nonzero scalar $\lambda \in K$ such that $v_i = \lambda u_i$, for all i, $1 \leq i \leq n+1$.*

Proof. Since $p(u_i) = p(v_i)$ for $1 \leq i \leq n+1$, there exist some nonzero scalars $\lambda_i \in K$ such that $v_i = \lambda_i u_i$ for all i, $1 \leq i \leq n+1$. Since we must have

$$p(u_1+\cdots+u_{n+1}) = p(v_1+\cdots+v_{n+1}),$$

there is some $\lambda \neq 0$ such that

$$\lambda(u_1 + \cdots + u_{n+1}) = v_1 + \cdots + v_{n+1} = \lambda_1 u_1 + \cdots + \lambda_{n+1} u_{n+1},$$

and thus we have

$$(\lambda - \lambda_1)u_1 + \cdots + (\lambda - \lambda_{n+1})u_{n+1} = 0,$$

and since $(u_1, \ldots, u_{n+1})$ is a basis, we have $\lambda_i = \lambda$ for all i, $1 \leq i \leq n+1$, which implies $\lambda_1 = \cdots = \lambda_{n+1} = \lambda$. $\square$

Proposition 26.2 shows that a projective frame determines a unique basis of E, up to a (nonzero) scalar. This would not necessarily be the case if we did not have a point a_{n+2} such that $a_{n+2} = p(u_1 + \cdots + u_{n+1})$.

When $n = 0$, the projective space consists of a single point a, and there is only one projective frame, the pair (a, a). When $n = 1$, the projective space is a line, and a projective frame consists of any three pairwise distinct points a, b, c on this line. When $n = 2$, the projective space is a plane, and a projective frame consists of any four distinct points a, b, c, d such that a, b, c are the vertices of a nondegenerate triangle and d is not on any of the lines determined by the sides of this triangle. These examples of projective frames are illustrated in Figure 26.7. The reader can easily generalize to higher dimensions.

Given a projective frame $(a_i)_{1\leq i\leq n+2}$ of $\mathbf{P}(E)$, let $(u_1, \ldots, u_{n+1})$ be a basis of E associated with $(a_i)_{1\leq i\leq n+2}$. For every $a \in \mathbf{P}(E)$, there is some $u \in E - \{0\}$ such that

$$a = [u]_\sim = \{\lambda u \mid \lambda \in K - \{0\}\},$$

the equivalence class of u, and the set

$$\{(x_1, \ldots, x_{n+1}) \in K^{n+1} \mid v = x_1 u_1 + \cdots + x_{n+1} u_{n+1},\ \ v \in [u]_\sim = a\}$$

of coordinates of all the vectors in the equivalence class $[u]_\sim$ is called the *set of homogeneous coordinates of a over the basis* $(u_1, \ldots, u_{n+1})$.

Note that for each homogeneous coordinate $(x_1, \ldots, x_{n+1})$ we must have $x_i \neq 0$ for some i, $1 \leq i \leq n+1$, and any two homogeneous coordinates $(x_1, \ldots, x_{n+1})$ and $(y_1, \ldots, y_{n+1})$ for a differ by a nonzero scalar, i.e., there is some $\lambda \neq 0$ such that $y_i = \lambda x_i$, $1 \leq i \leq n+1$. Homogeneous coordinates $(x_1, \ldots, x_{n+1})$ are sometimes denoted by $(x_1 \colon \cdots : x_{n+1})$, for instance in algebraic geometry.

By Proposition 26.2, any other basis $(v_1, \ldots, v_{n+1})$ associated with the projective frame $(a_i)_{1\leq i\leq n+2}$ differs from $(u_1, \ldots, u_{n+1})$ by a nonzero scalar, which implies that the set of homogeneous coordinates of $a \in \mathbf{P}(E)$ over the basis $(v_1, \ldots, v_{n+1})$ is identical to the set of homogeneous coordinates of $a \in \mathbf{P}(E)$ over the basis $(u_1, \ldots, u_{n+1})$. Consequently, we can associate a unique set of homogeneous coordinates to every point $a \in \mathbf{P}(E)$ with respect to the projective frame $(a_i)_{1\leq i\leq n+2}$. With respect to this projective frame, note that a_{n+2} has homogeneous coordinates $(1, \ldots, 1)$, and that a_i has homogeneous coordinates $(0, \ldots, 1, \ldots, 0)$, where the 1 is in the ith position, where $1 \leq i \leq n+1$. We summarize the above discussion in the following definition.

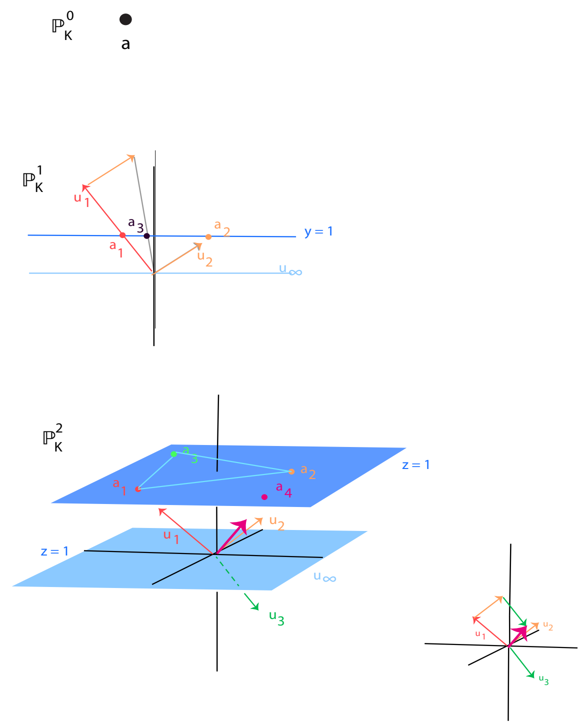


Figure 26.7: The projective frames for projective spaces of dimension 1, 2, and 3.

Definition 26.4. Given a nontrivial vector space E of dimension $n+1$, for any projective frame $(a_i)_{1\leq i\leq n+2}$ of $\mathbf{P}(E)$ and for any point $a \in \mathbf{P}(E)$, the *set of homogeneous coordinates of a with respect to* $(a_i)_{1\leq i\leq n+2}$ is the set of $(n+1)$-tuples

$$\{(\lambda x_1, \ldots, \lambda x_{n+1}) \in K^{n+1} \mid x_i \neq 0 \;\text{ for some } i,\; \lambda \neq 0,\; a = p(x_1u_1 + \cdots + x_{n+1}u_{n+1})\},$$

where $(u_1, \ldots, u_{n+1})$ is any basis of E associated with $(a_i)_{1\leq i\leq n+2}$.

Given a projective frame $(a_i)_{1\leq i\leq n+2}$ for $\mathbf{P}(E)$, if $(x_1, \ldots, x_{n+1})$ are homogeneous coordinates of a point $a \in \mathbf{P}(E)$, we write $a = (x_1, \ldots, x_{n+1})$, and with a slight abuse of language, we may even talk about a point $(x_1, \ldots, x_{n+1})$ in $\mathbf{P}(E)$ and write $(x_1, \ldots, x_{n+1}) \in \mathbf{P}(E)$.

The special case of the projective line $\mathbb{P}^1_K$ is worth examining. The projective line $\mathbb{P}^1_K$ consists of all equivalence classes $[x, y]$ of pairs $(x, y) \in K^2$ such that $(x, y) \neq (0, 0)$, under the equivalence relation $\sim$ defined such that

$$(x_1, y_1) \sim (x_2, y_2) \quad \text{iff} \quad x_2 = \lambda x_1 \quad \text{and} \quad y_2 = \lambda y_1,$$

for some $\lambda \in K - \{0\}$. When $y \neq 0$, the equivalence class of (x, y) contains the representative $(xy^{-1}, 1)$, and when $y = 0$, the equivalence class of $(x, 0)$ contains the representative $(1, 0)$.

Thus, there is a bijection between K and the set of equivalence classes containing some representative of the form $(x, 1)$, and we denote the class $[x, 1]$ by x. The equivalence class $[1, 0]$ is denoted by ∞ and it is called the point at infinity. Thus, the projective line $\mathbb{P}^1_K$ is in bijection with $K \cup \{\infty\}$. The three points $\infty = [1, 0]$, $0 = [0, 1]$, and $1 = [1, 1]$, form a projective frame for $\mathbb{P}^1_K$. The projective frame $(\infty, 0, 1)$ is often called the *canonical frame of* $\mathbb{P}^1_K$.

Homogeneous coordinates are also very useful to handle hyperplanes in terms of equations. If $(a_i)_{1\leq i\leq n+2}$ is a projective frame for $\mathbf{P}(E)$ associated with a basis $(u_1, \ldots, u_{n+1})$ for E, a nonnull linear form f is determined by $n+1$ scalars $\alpha_1, \ldots, \alpha_{n+1}$ (not all null), and a point $x \in \mathbf{P}(E)$ of homogeneous coordinates $(x_1, \ldots, x_{n+1})$ belongs to the projective hyperplane $\mathbf{P}(H)$ of equation f iff

$$\alpha_1 x_1 + \cdots + \alpha_{n+1} x_{n+1} = 0.$$

In particular, if $\mathbf{P}(E)$ is a projective plane, a line is defined by an equation of the form $\alpha x + \beta y + \gamma z = 0$. If $\mathbf{P}(E)$ is a projective space, a plane is defined by an equation of the form $\alpha x + \beta y + \gamma z + \delta w = 0$.

As an application, let us find the coordinates of the intersection point of two distinct lines in a projective plane $\mathbf{P}(E)$ (with respect to some projective frame (a_1, a_2, a_3, a_4)). If D and D' are two lines of equations

$$\alpha x + \beta y + \gamma z = 0 \quad \text{and} \quad \alpha' x + \beta' y + \gamma' z = 0, \tag{$*$}$$

then D and D' are distinct lines iff the matrix

$$\begin{pmatrix} \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{pmatrix}$$

has rank 2. We claim that the intersection Q of the lines D and D' has homogeneous coordinates

$$(\beta\gamma' - \beta'\gamma \colon \gamma\alpha' - \gamma'\alpha \colon \alpha\beta' - \alpha'\beta); \tag{$\dagger$}$$

in other words, it is the projective point corresponding to the cross-product

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \times \begin{pmatrix} \alpha' \\ \beta' \\ \gamma' \end{pmatrix},$$

as illustrated in Figure 26.8.

Indeed, the homogeneous coordinates of the intersection Q of D and D' must satisfy simultaneously the two equations $(*)$, and since the two determinants

$$\begin{vmatrix} \alpha & \beta & \gamma \\ \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} \alpha' & \beta' & \gamma' \\ \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{vmatrix}$$

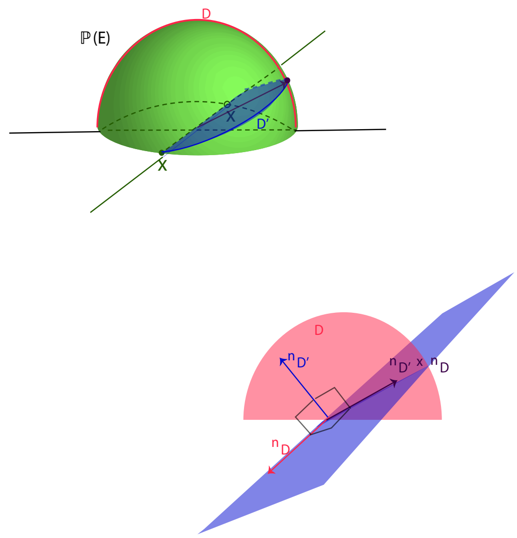


Figure 26.8: The intersection of two projective lines in the projective plane $\mathbf{P}(E)$ is the cross product of the normals for the two corresponding planes in $\mathbb{R}^3$.

are zero because they have two equal rows, and since by expanding these determinants with respect to their first row using the Laplace expansion formula we get

$$0 = \begin{vmatrix} \alpha & \beta & \gamma \\ \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{vmatrix} = \alpha(\beta\gamma' - \beta'\gamma) + \beta(\gamma\alpha' - \gamma'\alpha) + \gamma(\alpha\beta' - \alpha'\beta)$$

and

$$0 = \begin{vmatrix} \alpha' & \beta' & \gamma' \\ \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{vmatrix} = \alpha'(\beta\gamma' - \beta'\gamma) + \beta'(\gamma\alpha' - \gamma'\alpha) + \gamma'(\alpha\beta' - \alpha'\beta),$$

which confirms that the point

$$Q = (\beta\gamma' - \beta'\gamma \colon \gamma\alpha' - \gamma'\alpha \colon \alpha\beta' - \alpha'\beta)$$

satisfies both equations in $(*)$, and thus belongs to both lines D and D'. Since the matrix

$$\begin{pmatrix} \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \end{pmatrix}$$

has rank 2, at least one of the coordinates of Q is nonzero, so Q is indeed a point in the projective plane, and it is the intersection of the lines D and D'.

The result that we just proved yields the following criterion for three lines D, D', D'' in a projective plane to pass through a common point (to be concurrent). In a projective plane, three lines D, D', D'' of equations

$$\begin{aligned} \alpha x + \beta y + \gamma z &= 0 \\ \alpha' x + \beta' y + \gamma' z &= 0 \\ \alpha'' x + \beta'' y + \gamma'' z &= 0 \end{aligned}$$

are concurrent iff

$$\begin{vmatrix} \alpha & \beta & \gamma \\ \alpha' & \beta' & \gamma' \\ \alpha'' & \beta'' & \gamma'' \end{vmatrix} = 0.$$

We can also find the equation of the unique line $D = \langle P, P' \rangle$ passing through two distinct points $P = (u\colon v\colon w)$ and $P' = (u'\colon v'\colon w')$ of a projective plane. This line is given by the equation

$$(vw' - v'w)x + (wu' - w'u) + (uv' - u'v)z = 0, \qquad (\dagger\dagger)$$

and since

$$\begin{pmatrix} u & v & w \\ u' & v' & w' \end{pmatrix}$$

has rank 2 because $P \neq P'$, at least one of the coordinates of the equation $(\dagger\dagger)$ is nonzero. Observe that the coefficients of the equation $(\dagger\dagger)$ correspond to the cross-product

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} \times \begin{pmatrix} u' \\ v' \\ w'' \end{pmatrix}.$$

The equation of the line $D = \langle P, P' \rangle$ must be satisfied by the homogeneous coordinates of the points P and P'. Equation $(\dagger\dagger)$ can be written as

$$\begin{vmatrix} x & y & z \\ u & v & w \\ u' & v' & w' \end{vmatrix} = 0,$$

and a reasoning as in the case of the intersection of lines shows that the equation of the line passing through P and P' is given by equation $(\dagger\dagger)$.

Then, in a projective plane, three points $P = (u\colon v\colon w)$, $P' = (u'\colon v'\colon w')$ and $P'' = (u''\colon v''\colon w'')$ belong to a common line (are collinear) iff

$$\begin{vmatrix} u & v & w \\ u' & v' & w' \\ u'' & v'' & w'' \end{vmatrix} = 0.$$

More generally, in a projective space $\mathbf{P}(E)$ of dimension $n \geq 2$, if n points $P_1, \ldots, P_n$ are projectively independent and if P_i has homogeneous coordinates $(u_1^i\colon \cdots : u_{n+1}^i)$ (with respect to some projective frame $(a_1, \ldots, a_{n+2})$), then the equation of the unique hyperplane H containing $P_1, \ldots, P_n$ is given by the equation

$$\begin{vmatrix} x_1 & x_2 & \cdots & x_n & x_{n+1} \\ u_1^1 & u_2^1 & \cdots & u_n^1 & u_{n+1}^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_1^{n-1} & u_2^{n-1} & \cdots & u_n^{n-1} & u_{n+1}^{n-1} \\ u_1^n & u_2^n & \cdots & u_n^n & u_{n+1}^n \end{vmatrix} = 0.$$

We also have the following proposition giving another characterization of projective frames.

Proposition 26.3. *A family $(a_i)_{1 \leq i \leq n+2}$ of $n+2$ points is a projective frame of $\mathbf{P}(E)$ iff for every i, $1 \leq i \leq n+2$, the subfamily $(a_j)_{j \neq i}$ is projectively independent.*

Proof. We leave as an (easy) exercise the fact that if $(a_i)_{1 \leq i \leq n+2}$ is a projective frame, then each subfamily $(a_j)_{j \neq i}$ is projectively independent. Conversely, pick some $u_i \in E - \{0\}$ such that $a_i = p(u_i)$, $1 \leq i \leq n+2$. Since $(a_j)_{j \neq n+2}$ is projectively independent, $(u_1, \ldots, u_{n+1})$ is a basis of E. Thus, we must have

$$u_{n+2} = \lambda_1 u_1 + \cdots + \lambda_{n+1} u_{n+1},$$

for some $\lambda_i \in K$. However, since for every i, $1 \leq i \leq n+1$, the family $(a_j)_{j \neq i}$ is projectively independent, we must have $\lambda_i \neq 0$, and thus $(\lambda_1 u_1, \ldots, \lambda_{n+1} u_{n+1})$ is also a basis of E, and since

$$u_{n+2} = \lambda_1 u_1 + \cdots + \lambda_{n+1} u_{n+1},$$

it induces the projective frame $(a_i)_{1 \leq i \leq n+2}$. $\square$

Figure 26.9 shows a projective frame (a, b, c, d) in a projective plane. With respect to this projective frame, the points a, b, c, d have homogeneous coordinates $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, and $(1, 1, 1)$. Let a' be the intersection of $\langle d, a \rangle$ and $\langle b, c \rangle$, b' be the intersection of $\langle d, b \rangle$ and $\langle a, c \rangle$, and c' be the intersection of $\langle d, c \rangle$ and $\langle a, b \rangle$. Then the points a', b', c' have homogeneous coordinates $(0, 1, 1)$, $(1, 0, 1)$, and $(1, 1, 0)$. The diagram formed by the line segments $\langle a, c' \rangle$, $\langle a, b' \rangle$, $\langle b, b' \rangle$, $\langle c, c' \rangle$, $\langle a, d \rangle$, and $\langle b, c \rangle$ is sometimes called a *Möbius net*; see Hilbert and Cohn-Vossen [92] (Chapter III, §15, page 96).

Recall that the equation of a line (a hyperplane in a projective plane) in terms of homogeneous coordinates with respect to the projective frame (a, b, c, d) is given by a homogeneous equation of the form

$$\alpha x + \beta y + \gamma z = 0,$$

where α, β, γ are not all zero. It is easily verified that the equations of the lines $\langle a, b \rangle$, $\langle a, c \rangle$, $\langle b, c \rangle$, are $z = 0$, $y = 0$, and $x = 0$, and the equations of the lines $\langle a, d \rangle$, $\langle b, d \rangle$, and $\langle c, d \rangle$,

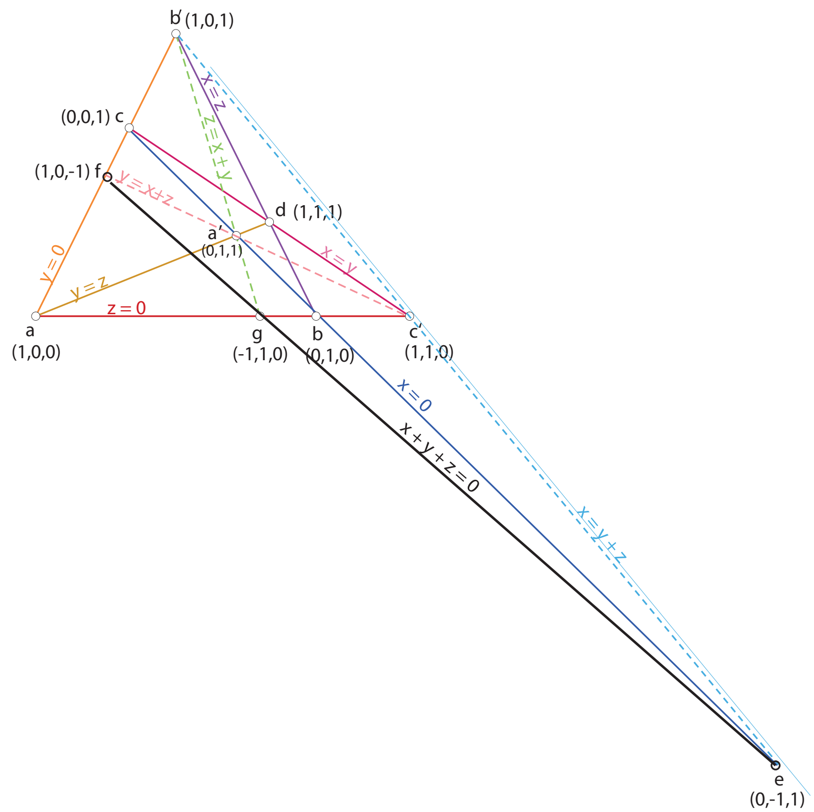


Figure 26.9: A projective frame (a, b, c, d).

are $y = z$, $x = z$, and $x = y$. The equations of the lines $\langle a', b'\rangle$, $\langle a', c'\rangle$, $\langle b', c'\rangle$ are $z = x + y$, $y = x + z$, and $x = y + z$.

If we let e be the intersection of $\langle b, c\rangle$ and $\langle b', c'\rangle$, f be the intersection of $\langle a, c\rangle$ and $\langle a', c'\rangle$, and g be the intersection of $\langle a, b\rangle$ and $\langle a', b'\rangle$, then it easily seen that e, f, g have homogeneous coordinates $(0, -1, 1)$, $(1, 0, -1)$, and $(-1, 1, 0)$. For example, since the equation of the line $\langle b, c\rangle$ is $x = 0$ and the equation of the line $\langle b', c'\rangle$ is $x = y + z$, for $x = 0$, we get $z = -y$, which correspond to the homogeneous coordinates $(0, -1, 1)$ for e.

The coordinates of the points e, f, g satisfy the equation $x + y + z = 0$, which shows that they are collinear.

As pointed out in Coxeter [45] (Proposition 2.41), this is a special case of the projective version of Desargues's theorem (Proposition 26.7) applied to the triangles (a, b, c) and (a', b', c'). Indeed, by construction, the lines $\langle a, a'\rangle$, $\langle b, b'\rangle$, and $\langle c, c'\rangle$ intersect in the common point d. The line containing the points e, f, g is called the *polar line (or fundamental line)* of d with respect to the triangle (a, b, c) (see Pedoe [136]). The diagram also shows the intersection g of $\langle a, b\rangle$ and $\langle a', b'\rangle$.

The projective space of circles provides a nice illustration of homogeneous coordinates.

Let E be the vector space (over $\mathbb{R}$) consisting of all homogeneous polynomials of degree 2 in x, y, z of the form

$$ax^2 + ay^2 + bxz + cyz + dz^2$$

(plus the null polynomial). The projective space $\mathbf{P}(E)$ consists of all equivalence classes

$$[P]_{\sim} = \{\lambda P \mid \lambda \neq 0\},$$

where $P(x, y, z)$ is a nonnull homogeneous polynomial in E. We want to give a geometric interpretation of the points of the projective space $\mathbf{P}(E)$. In order to do so, pick some projective frame (a_1, a_2, a_3, a_4) for the projective plane $\mathbb{RP}^2$, and associate to every $[P] \in \mathbf{P}(E)$ the subset of $\mathbb{RP}^2$ known as its its *zero locus (or zero set, or variety)* $V([P])$, and defined such that

$$V([P]) = \{a \in \mathbb{RP}^2 \mid P(x, y, z) = 0\},$$

where (x, y, z) are homogeneous coordinates for a.

As explained earlier, we also use the simpler notation

$$V([P]) = \{(x, y, z) \in \mathbb{RP}^2 \mid P(x, y, z) = 0\}.$$

Actually, in order for $V([P])$ to make sense, we have to check that $V([P])$ does not depend on the representative chosen in the equivalence class $[P] = \{\lambda P \mid \lambda \neq 0\}$. This is because

$$P(x, y, z) = 0 \quad \text{iff} \quad \lambda P(x, y, z) = 0 \quad \text{when } \lambda \neq 0.$$

For simplicity of notation, we also denote $V([P])$ by $V(P)$. We also have to check that if $(\lambda x, \lambda y, \lambda z)$ are other homogeneous coordinates for $a \in \mathbb{RP}^2$, where $\lambda \neq 0$, then

$$P(x, y, z) = 0 \quad \text{iff} \quad P(\lambda x, \lambda y, \lambda z) = 0.$$

However, since $P(x, y, z)$ is homogeneous of degree 2, we have

$$P(\lambda x, \lambda y, \lambda z) = \lambda^2 P(x, y, z),$$

and since $\lambda \neq 0$,

$$P(x, y, z) = 0 \quad \text{iff} \quad \lambda^2 P(x, y, z) = 0.$$

The above argument applies to any homogeneous polynomial $P(x_1, \ldots, x_n)$ in n variables of any degree m, since

$$P(\lambda x_1, \ldots, \lambda x_n) = \lambda^m P(x_1, \ldots, x_n).$$

Thus, we can associate to every $[P] \in \mathbf{P}(E)$ the curve $V(P)$ in $\mathbb{RP}^2$. One might wonder why we are considering only homogeneous polynomials of degree 2, and not arbitrary polynomials of degree 2? The first reason is that the polynomials in x, y, z of degree 2 do **not** form a vector space. For example, if $P = x^2 + x$ and $Q = -x^2 + y$, the polynomial $P + Q = x + y$ is not of degree 2. We could consider the set of polynomials of degree ≤ 2, which is a vector space, but now the problem is that $V(P)$ is not necessarily well defined!. For example, if $P(x, y, z) = -x^2 + 1$, we have

$$P(1, 0, 0) = 0 \quad \text{and} \quad P(2, 0, 0) = -3,$$

and yet $(2, 0, 0) = 2(1, 0, 0)$, so that $P(x, y, z)$ takes different values depending on the representative chosen in the equivalence class $[1, 0, 0]$. Thus, we are led to restrict ourselves to homogeneous polynomials. Actually, this is usually an advantage more than a disadvantage, because homogeneous polynomials tend to be well behaved.

What are the curves $V(P)$? One way to "see" such curves is to go back to the hyperplane model of $\mathbb{RP}^2$ in terms of the plane H of equation $z = 1$ in $\mathbb{R}^3$. Then the trace of $V(P)$ on H is the circle of equation

$$ax^2 + ay^2 + bx + cy + d = 0.$$

Thus, we may think of $\mathbf{P}(E)$ as a projective space of circles. However, there are some problems. For example, $V(P)$ may be empty! This happens, for instance, for $P(x, y, z) = x^2 + y^2 + z^2$, since the equation

$$x^2 + y^2 + z^2 = 0$$

has only the trivial solution $(0, 0, 0)$, which does not correspond to any point in $\mathbb{RP}^2$. Indeed, only nonnull vectors in $\mathbb{R}^3$ yield points in $\mathbb{RP}^2$. It is also possible that $V(P)$ is reduced to a single point, for instance when $P(x, y, z) = x^2 + y^2$, since the only homogeneous solution of

$$x^2 + y^2 = 0$$

is $(0, 0, 1)$. Also, note that the map

$$[P] \mapsto V(P)$$

is not injective. For instance, $P = x^2 + y^2$ and $Q = x^2 + 2y^2$ define the same degenerate circle reduced to the point $(0, 0, 1)$. We also accept as circles the union of two lines, as in the case

$$(bx + cy + dz)z = 0,$$

where $a = 0$, and even a double line, as in the case

$$z^2 = 0,$$

where $a = b = c = 0$.

A clean way to resolve most of these problems is to switch to homogeneous polynomials over the complex field $\mathbb{C}$ and to consider curves in $\mathbb{CP}^2$. This is what is done in algebraic geometry (see Fulton [66] or Harris [87]). If $P(x, y, z)$ is a homogeneous polynomial over $\mathbb{C}$ of degree 2 (plus the null polynomial), it is easy to show that $V(P)$ is always nonempty, and in fact infinite. It can also be shown that $V(P) = V(Q)$ implies that $Q = \lambda P$ for some $\lambda \in \mathbb{C}$, with $\lambda \neq 0$ (see Samuel [142], Section 1.6, Theorem 10). Another advantage of switching to the complex field $\mathbb{C}$ is that the theory of intersection is cleaner. Thus, any two circles that do not contain a common line always intersect in four points, some of which might be multiple points (as in the case of tangent circles). This may seem surprising, since in the real plane, two circles intersect in at most two points. Where are the other two points? They turn out to be the points $(1, i, 0)$ and $(1, -i, 0)$, as one can immediately verify. We can think of them as complex points at infinity! Not only are they at infinity, but they are not real. No wonder we cannot see them! We will come back to these points, called the *circular points*, in Section 26.14.

Going back to the vector space E of circles over $\mathbb{R}$, it is worth saying that it can be shown that if $V(P) = V(Q)$ contains at least two points (in which case, $V(P)$ is actually infinite), then $Q = \lambda P$ for some $\lambda \in \mathbb{R}$ with $\lambda \neq 0$ (see Tisseron [175], Theorem 3.6.1 and Theorem 4.7). Thus, even over $\mathbb{R}$, the mapping

$$[P] \mapsto V(P)$$

is injective whenever $V(P)$ is neither empty nor reduced to a single point. Note that the projective space $\mathbf{P}(E)$ of circles has dimension 3. In fact, it is easy to show that three distinct points that are not collinear determine a unique circle (see Samuel [142], Section 1.6).

In a similar vein, we can define the *projective space of conics* $\mathbf{P}(E)$ where E is the vector space (over $\mathbb{R}$) consisting of all homogeneous polynomials of degree 2 in x, y, z,

$$ax^2 + by^2 + cxy + dxz + eyz + fz^2$$

(plus the null polynomial). The curves $V(P)$ are indeed conics, perhaps degenerate. To see this, we can use the hyperplane model of $\mathbb{RP}^2$. The trace of $V(P)$ on the plane of equation $z = 1$ is the conic of equation

$$ax^2 + by^2 + cxy + dx + ey + f = 0.$$

Another way to see that $V(P)$ is a conic is to observe that in $\mathbb{R}^3$,

$$ax^2 + by^2 + cxy + dxz + eyz + fz^2 = 0$$

defines a cone with vertex $(0, 0, 0)$, and since its section by the plane $z = 1$ is a conic, all of its sections by planes are conics. See Figure 26.10 for schematic illustration of a projective conic embedded in $\mathbb{RP}^2$.

The mapping

$$[P] \mapsto V(P)$$

is still injective when E is defined over the ground field $\mathbb{C}$ (Samuel [142], Section 1.6, Theorem 10), or if $V(P)$ has at least two points when E is defined over $\mathbb{R}$ (Tisseron [175], Theorem 3.6.1 and Theorem 4.7). Note that the projective space $\mathbf{P}(E)$ of conics has dimension 5. In fact, it can be shown that five distinct points, no four of which are collinear, determine a

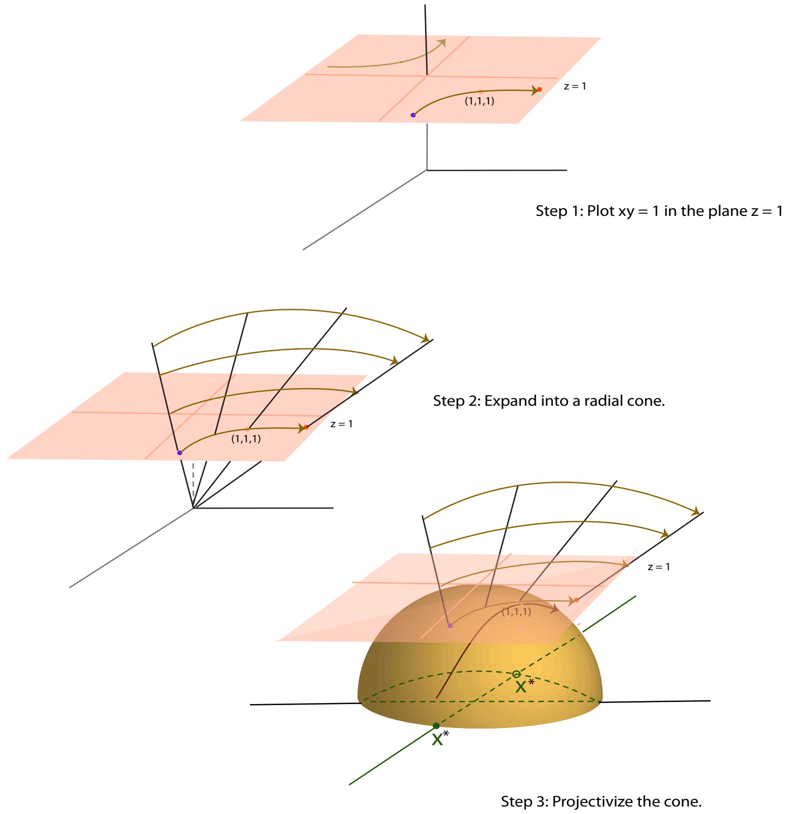


Figure 26.10: A three step process for constructing $V(P)$ where P is the homogenous conic $xy = z$. In Step 2, we convert to homogenous coordinates via the transformation $x \to x/z$, $y \to y/z$.

unique conic (among many sources, see Samuel [142], Section 1.7, Theorem 17, or Coxeter [45], Theorem 6.56, where a geometric construction is given in Section 6.6).

In fact, if we pick a projective frame (a_1, a_2, a_3, a_4) in $\mathbb{CP}^2$ (or $\mathbb{RP}^2$), and if the five points p_1, p_2, p_3, p_4, p_5 have homogeneous coordinates $p_i = (x_i, y_i, z_i)$ for $i = 1, \ldots, 5$ and (x, y, z) are variables, then it is an easy exercise to show that the equation of the unique conic C passing through the points p_1, p_2, p_3, p_4, p_5 is given by

$$\begin{vmatrix} x^2 & xy & y^2 & xz & yz & z^2 \\ x_1^2 & x_1y_1 & y_1^2 & x_1z_1 & y_1z_1 & z_1^2 \\ x_2^2 & x_2y_2 & y_2^2 & x_2z_2 & y_2z_2 & z_2^2 \\ x_3^2 & x_3y_3 & y_3^2 & x_3z_3 & y_3z_3 & z_3^2 \\ x_4^2 & x_4y_4 & y_4^2 & x_4z_4 & y_4z_4 & z_4^2 \\ x_5^2 & x_5y_5 & y_5^2 & x_5z_5 & y_5z_5 & z_5^2 \end{vmatrix} = 0.$$

The polynomial obtained by expanding the above determinant according to the first row is a homogeneous polynomial of degree 2 in the variables x, y, z, and it is not the zero polynomial because the 5×6 matrix obtained by deleting the first row in the matrix of the determinant has rank 5. Indeed, this is the matrix of the linear system determining the six coefficients of the conic passign through p_1, p_2, p_3, p_4, p_5 (up to a scalar), and since this conic is unique, this matrix must have rank 5.

It is also interesting to see what are lines in the space of circles or in the space of conics. In both cases we get pencils (of circles and conics, respectively). For more details, see Samuel [142], Sidler [161], Tisseron [175], Lehmann and Bkouche [115], Pedoe [136], Coxeter [45, 46], and Veblen and Young [183, 184].

The generalization of the space of projective conics is the space of *projective quadrics* $\mathbf{P}(E)$, where E is the vector space (over a field K, typically $K = \mathbb{R}$ or $K = \mathbb{C}$) consisting of all homogeneous polynomials $P(x_1, \ldots, x_{N+1})$ of degree 2 in the variables $x_1, \ldots, x_{N+1}$, with $N \geq 3$ (plus the null polynomial). The zero locus $V(P)$ of P is defined just as before as

$$V(P) = \{(x_1\colon \cdots : x_{N+1}) \in \mathbb{P}^N_K \mid P(x_1, \ldots, x_{N+1}) = 0\}.$$

If the field K is algebraically closed, in particular if $K = \mathbb{C}$, then $V(P) = V(Q)$ implies that there is some nonzero $\lambda \in K$ such that $Q = \lambda P$; see Berger [12] (Chapter 14, Theorem 14.1.6.2).

Another situation where the map $[P] \mapsto V(P)$ is injective involves the notion of simple (or regular) point of a quadric. For any $a = (a_1\colon \cdots : a_{N+1}) \in \mathbb{P}^N_K$, let $P_{x_i}(a)$ be the partial derivative of P at a given by

$$P_{x_i}(a) = \frac{\partial P}{\partial x_i}(a_1, \ldots, a_{N+1}).$$

Strictly speaking, $P_{x_i}(a)$ depends on the representative $(a_1, \ldots, a_{N+1}) \in K^{N+1}$ chosen for the point a, but since P is homogeneous of degree 2, for any nonzero $\lambda \in K$,

$$\frac{\partial P}{\partial x_i}(\lambda a_1, \ldots, \lambda a_{N+1}) = \lambda \frac{\partial P}{\partial x_i}(a_1, \ldots, a_{N+1}).$$

Thus $P_{x_i}(a)$ is defined up to a nonzero scalar. In particular, whether or not $P_{x_i}(a) = 0$ depends only the point $a = (a_1\colon \cdots : a_{N+1}) \in \mathbb{P}^N_K$. Then the point $a \in V(P)$ is said to be *simple* (or *regular*) if

$$P_{x_i}(a) \neq 0 \quad \text{for some } i,\ 1 \leq i \leq N+1.$$

Otherwise, if $P_{x_1}(a) = \cdots = P_{x_{N+1}}(a) = 0$, we say that $a \in V(P)$ is a *singular* point. If $a \in V(P)$ is a regular point, then the *tangent hyperplane* $T_a V(P)$ to $V(P)$ at a is the hyperplane given by the equation

$$P_{x_1}(a)x_1 + \cdots + P_{x_{N+1}}(a)x_{N+1} = 0.$$

It can be shown that if the field K is not the field $\mathbf{F}_2 = \{0, 1\}$ and if the quadric $V(P)$ contains some regular point, then $V(P) = V(Q)$ implies that there is some nonzero $\lambda \in K$ such that $Q = \lambda P$; see Samuel [142] (Chapter 3, Theorem 46).

Quadrics, projective, affine, and Euclidean, have been thoroughly investigated. Among many sources, the reader is referred to Berger [11], Samuel [142], Tisseron [175], Fresnel [65], and Vienne [185].

We could also investigate algebraic plane curves of any degree m, by letting E be the vector space of homogeneous polynomials of degree m in x, y, z (plus the null polynomial). The zero locus $V(P)$ of P is defined just as before as

$$V(P) = \{(x\colon y\colon z) \in \mathbb{RP}^2 \mid P(x, y, z) = 0\}.$$

Observe that when $m = 1$, since homogeneous polynomials of degree 1 are linear forms, we are back to the case where $E = (\mathbb{R}^3)^*$, the dual space of $\mathbb{R}^3$, and $\mathbf{P}(E)$ can be identified with the set of lines in $\mathbb{RP}^2$. But when $m \geq 3$, things are even worse regarding the injectivity of the map $[P] \mapsto V(P)$. For instance, both $P = xy^2$ and $Q = x^2y$ define the same union of two lines. It is necessary to consider *irreducible* curves, i.e., curves that are defined by irreducible polynomials, and to work over the field $\mathbb{C}$ of complex numbers (recall that a polynomial P is irreducible if it cannot be written as the product $P = Q_1Q_2$ of two polynomials Q_1, Q_2 of degree ≥ 1). We refer the reader to Fischer's book for a beautiful (and very clear) introduction to algebraic curves [62]. The next step is Fulton [66].

We can also investigate algebraic surfaces in $\mathbb{RP}^3$ (or $\mathbb{CP}^3$), by letting E be the vector space of homogeneous polynomials of degree m in four variables x, y, z, t (plus the null polynomial). We can also consider the zero locus of a set of equations

$$\mathcal{E} = \{P_1 = 0,\ P_2 = 0,\ \ldots,\ P_n = 0\},$$

where $P_1, \ldots, P_n$ are homogeneous polynomials of degree m in x, y, z, t, defined as

$$V(\mathcal{E}) = \{(x\colon y\colon z\colon t) \in \mathbb{RP}^3 \mid P_i(x, y, z, t) = 0,\ 1 \leq i \leq n\}.$$

This way, we can also deal with space curves.

Finally, we can consider homogeneous polynomials $P(x_1, \ldots, x_{N+1})$ in $N + 1$ variables and of degree m (plus the null polynomial), and study the subsets of $\mathbb{RP}^N$ or $\mathbb{CP}^N$ (or more generally of $\mathbb{P}^N_K$, for an arbitrary field K), defined as the zero locus of a set of equations

$$\mathcal{E} = \{P_1 = 0,\ P_2 = 0,\ \ldots,\ P_n = 0\},$$

where $P_1, \ldots, P_n$ are homogeneous polynomials of degree m in the variables $x_1, \ldots, x_{N+1}$. For example, it turns out that the set of lines in $\mathbb{RP}^3$ forms a surface of degree 2 in $\mathbb{RP}^5$ (the Klein quadric). However, all this would really take us too far into algebraic geometry, and we simply refer the interested reader to Hulek [97], Fulton [66], and Harris [87].

We now consider projective maps.

26.5 Projective Maps

Given two nontrivial vector spaces E and F and a linear map $f\colon E \to F$, observe that for every $u, v \in (E - \operatorname{Ker} f)$, if $v = \lambda u$ for some $\lambda \in K - \{0\}$, then $f(v) = \lambda f(u)$, and thus f restricted to $(E - \operatorname{Ker} f)$ induces a function $\mathbf{P}(f)\colon (\mathbf{P}(E) - \mathbf{P}(\operatorname{Ker} f)) \to \mathbf{P}(F)$ defined such that

$$\mathbf{P}(f)([u]_\sim) = [f(u)]_\sim,$$

as in the following commutative diagram:

$$\begin{tikzcd} E - \operatorname{Ker} f \arrow[r, "f"] \arrow[d, "p"'] & F - \{0\} \arrow[d, "p"] \\ \mathbf{P}(E) - \mathbf{P}(\operatorname{Ker} f) \arrow[r, "\mathbf{P}(f)"'] & \mathbf{P}(F) \end{tikzcd}$$

When f is injective, i.e., when $\operatorname{Ker} f = \{0\}$, then $\mathbf{P}(f)\colon \mathbf{P}(E) \to \mathbf{P}(F)$ is indeed a well-defined function. The above discussion motivates the following definition.

Definition 26.5. Given two nontrivial vector spaces E and F, any linear map $f\colon E \to F$ induces a partial map $\mathbf{P}(f)\colon \mathbf{P}(E) \to \mathbf{P}(F)$ called a *projective map*, such that if $\operatorname{Ker} f = \{u \in E \mid f(u) = 0\}$ is the kernel of f, then $\mathbf{P}(f)\colon (\mathbf{P}(E) - \mathbf{P}(\operatorname{Ker} f)) \to \mathbf{P}(F)$ is a total map defined such that

$$\mathbf{P}(f)([u]_\sim) = [f(u)]_\sim,$$

as in the following commutative diagram:

$$\begin{tikzcd} E - \operatorname{Ker} f \arrow[r, "f"] \arrow[d, "p"'] & F - \{0\} \arrow[d, "p"] \\ \mathbf{P}(E) - \mathbf{P}(\operatorname{Ker} f) \arrow[r, "\mathbf{P}(f)"'] & \mathbf{P}(F) \end{tikzcd}$$

If f is injective, i.e., when $\operatorname{Ker} f = \{0\}$, then $\mathbf{P}(f)\colon \mathbf{P}(E) \to \mathbf{P}(F)$ is a total function called a *projective transformation*, and when f is bijective, we call $\mathbf{P}(f)$ a *projectivity, or projective isomorphism, or homography*. The set of projectivities $\mathbf{P}(f)\colon \mathbf{P}(E) \to \mathbf{P}(E)$ is a group called the *projective (linear) group*, and is denoted by $\mathbf{PGL}(E)$.

One should realize that if a linear map $f\colon E \to F$ is not injective, then the projective map $\mathbf{P}(f)\colon \mathbf{P}(E) \to \mathbf{P}(F)$ is only a *partial map*, i.e., it is undefined on $\mathbf{P}(\operatorname{Ker} f)$. In particular, if $f\colon E \to F$ is the null map (i.e., $\operatorname{Ker} f = E$), the domain of $\mathbf{P}(f)$ is empty and $\mathbf{P}(f)$ is the partial function undefined everywhere. We might want to require in Definition 26.5 that f not be the null map to avoid this degenerate case. Projective maps are often defined only when they are induced by bijective linear maps.

We take a closer look at the projectivities of the projective line $\mathbb{P}^1_K$, since they play a role in the "change of parameters" for projective curves. A projectivity $f\colon \mathbb{P}^1_K \to \mathbb{P}^1_K$ is induced by some bijective linear map $g\colon K^2 \to K^2$ given by some invertible matrix

$$M(g) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $ad - bc \neq 0$. Since the projective line $\mathbb{P}^1_K$ is isomorphic to $K \cup \{\infty\}$, it is easily verified that f is defined as follows:

$$c \neq 0 \begin{cases} z \mapsto \dfrac{az+b}{cz+d} & \text{if } z \neq -\dfrac{d}{c}, \\ -\dfrac{d}{c} \mapsto \infty, \\ \infty \mapsto \dfrac{a}{c}; \end{cases} \qquad c = 0 \begin{cases} z \mapsto \dfrac{az+b}{d}, \\ \infty \mapsto \infty. \end{cases}$$

From Section 26.4, we know that the points not at infinity are repesented by vectors of the form $(z, 1)$ where $z \in K$ and that ∞ is represented by $(1, 0)$. First, assume $c \neq 0$. Since

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} z \\ 1 \end{pmatrix} = \begin{pmatrix} az+b \\ cz+d \end{pmatrix},$$

if $cz + d \neq 0$, that is, $z \neq -d/c$, then

$$(az+b, cz+d) \sim \left(\frac{az+b}{cz+d}, 1 \right),$$

so z is is mapped to $(az + d)/cz + d)$. If $cz + d = 0$, then

$$(az+d, 0) \sim (1, 0) = \infty,$$

so $-d/c$ is mapped to ∞. We also have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix},$$

and since $c \neq 0$ we have

$$(a, c) \sim (a/c, 1),$$

so ∞ is mapped to a/c. The case where $c = 0$ is handled similarly.

If $K = \mathbb{R}$ or $K = \mathbb{C}$, note that a/c is the limit of $(az + b)/(cz + d)$, as z approaches infinity, and the limit of $(az + b)/(cz + d)$ as z approaches $-d/c$ is ∞ (when $c \neq 0$).

Projections between hyperplanes form an important example of projectivities.

Definition 26.6. Given a projective space $\mathbf{P}(E)$, for any two distinct hyperplanes $\mathbf{P}(H)$ and $\mathbf{P}(H')$, for any point $c \in \mathbf{P}(E)$ neither in $\mathbf{P}(H)$ nor in $\mathbf{P}(H')$, the *projection (or perspectivity) of center c between* $\mathbf{P}(H)$ *and* $\mathbf{P}(H')$ is the map $f\colon \mathbf{P}(H) \to \mathbf{P}(H')$ defined such that for every $a \in \mathbf{P}(H)$, the point $f(a)$ is the intersection of the line $\langle c, a\rangle$ through c and a with $\mathbf{P}(H')$.

Let us verify that f is well–defined and a bijective projective transformation. Since the hyperplanes $\mathbf{P}(H)$ and $\mathbf{P}(H')$ are distinct, the hyperplanes H and H' in E are distinct, and since c is neither in $\mathbf{P}(H)$ nor in $\mathbf{P}(H')$, letting $c = p(u)$ for some nonnull vector $u \in E$, then $u \notin H$ and $u \notin H'$, and thus $E = H \oplus Ku = H' \oplus Ku$. If $\pi\colon E \to H'$ is the linear map (projection onto H' parallel to u) defined such that

$$\pi(w + \lambda u) = w,$$

for all $w \in H'$ and all $\lambda \in K$, since $E = H \oplus Ku = H' \oplus Ku$, the restriction $g\colon H \to H'$ of $\pi\colon E \to H'$ to H is a linear bijection between H and H', and clearly $f = \mathbf{P}(g)$, which shows that f is a projectivity.

Remark: Going back to the linear map $\pi\colon E \to H'$ (projection onto H' parallel to u), note that $\mathbf{P}(\pi)\colon \mathbf{P}(E) \to \mathbf{P}(H')$ is also a projective map, but it is not injective, and thus only a partial map. More generally, given a direct sum $E = V \oplus W$, the projection $\pi\colon E \to V$ onto V parallel to W induces a projective map $\mathbf{P}(\pi)\colon \mathbf{P}(E) \to \mathbf{P}(V)$, and given another direct sum $E = U \oplus W$, the restriction of π to U induces a perspectivity f between $\mathbf{P}(U)$ and $\mathbf{P}(V)$. Geometrically, f is defined as follows: Given any point $a \in \mathbf{P}(U)$, if $\langle \mathbf{P}(W), a\rangle$ is the smallest projective subspace containing $\mathbf{P}(W)$ and a, the point $f(a)$ is the intersection of $\langle \mathbf{P}(W), a\rangle$ with $\mathbf{P}(V)$.

Figure 26.11 illustrates a projection f of center c between two projective lines Δ and Δ' (in the real projective plane).

If we consider three distinct points d_1, d_2, d_3 on Δ and their images d_1', d_2', d_3' on Δ' under the projection f, then ratios are not preserved, that is,

$$\frac{\overrightarrow{d_3d_1}}{\overrightarrow{d_3d_2}} \neq \frac{\overrightarrow{d_3'd_1'}}{\overrightarrow{d_3'd_2'}}.$$

However, if we consider four distinct points d_1, d_2, d_3, d_4 on Δ and their images d_1', d_2', d_3', d_4' on Δ' under the projection f, we will show later that we have the following preservation of the so-called "cross-ratio"

$$\frac{\overrightarrow{d_3d_1}}{\overrightarrow{d_3d_2}} \Bigg/ \frac{\overrightarrow{d_4d_1}}{\overrightarrow{d_4d_2}} = \frac{\overrightarrow{d_3'd_1'}}{\overrightarrow{d_3'd_2'}} \Bigg/ \frac{\overrightarrow{d_4'd_1'}}{\overrightarrow{d_4'd_2'}}.$$

Cross-ratios and projections play an important role in geometry (for some very elegant illustrations of this fact, see Sidler [161]).

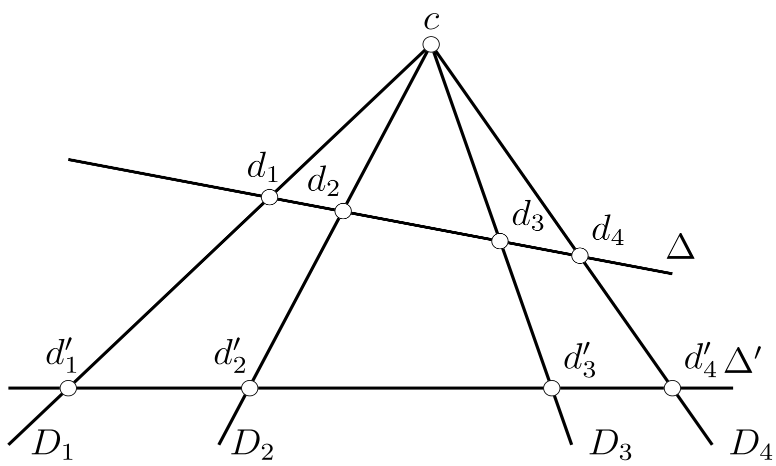


Figure 26.11: A projection of center c between two lines Δ and Δ'.

We now turn to the issue of determining when two linear maps f, g determine the same projective map, i.e., when $\mathbf{P}(f) = \mathbf{P}(g)$. The following proposition gives us a complete answer.

Proposition 26.4. *Given two nontrivial vector spaces E and F, for any two linear maps $f\colon E \to F$ and $g\colon E \to F$, we have $\mathbf{P}(f) = \mathbf{P}(g)$ iff there is some scalar $\lambda \in K - \{0\}$ such that $g = \lambda f$.*

Proof. If $g = \lambda f$, it is clear that $\mathbf{P}(f) = \mathbf{P}(g)$. Conversely, in order to have $\mathbf{P}(f) = \mathbf{P}(g)$, we must have $\operatorname{Ker} f = \operatorname{Ker} g$. If $\operatorname{Ker} f = \operatorname{Ker} g = E$, then f and g are both the null map, and this case is trivial. If $E - \operatorname{Ker} f \neq \emptyset$, by taking a basis of $\operatorname{Im} f$ and some inverse image of this basis, we obtain a basis B of a subspace G of E such that $E = \operatorname{Ker} f \oplus G$. If $\dim(G) = 1$, the restriction of any linear map $f\colon E \to F$ to G is determined by some nonzero vector $u \in E$ and some scalar $\lambda \in K$, and the proposition is obvious. Thus, assume that $\dim(G) \geq 2$. For any two distinct basis vectors $u, v \in B$, since $\mathbf{P}(f) = \mathbf{P}(g)$, there must be some nonzero scalars $\lambda(u)$, $\lambda(v)$, and $\lambda(u+v)$ such that

$$g(u) = \lambda(u)f(u), \quad g(v) = \lambda(v)f(v), \quad g(u+v) = \lambda(u+v)f(u+v).$$

Since f and g are linear, we get

$$g(u) + g(v) = \lambda(u)f(u) + \lambda(v)f(v) = \lambda(u+v)(f(u) + f(v)),$$

that is,

$$(\lambda(u+v) - \lambda(u))f(u) + (\lambda(u+v) - \lambda(v))f(v) = 0.$$

Since f is injective on G and $u, v \in B \subseteq G$ are linearly independent, $f(u)$ and $f(v)$ are also linearly independent, and thus we have

$$\lambda(u+v) = \lambda(u) = \lambda(v).$$

Now we have shown that $\lambda(u) = \lambda(v)$, for any two distinct basis vectors in B, which proves that $\lambda(u)$ is independent of $u \in G$, and proves that $g = \lambda f$. $\square$

Proposition 26.4 shows that the projective linear group $\mathbf{PGL}(E)$ is isomorphic to the quotient group of the linear group $\mathbf{GL}(E)$ modulo the subgroup $K^*\mathrm{id}_E$ (where $K^* = K - \{0\}$). Using projective frames, we prove the following useful result.

Proposition 26.5. *Given two nontrivial vector spaces E and F of the same dimension $n+1$, for any two projective frames $(a_i)_{1\leq i\leq n+2}$ for $\mathbf{P}(E)$ and $(b_i)_{1\leq i\leq n+2}$ for $\mathbf{P}(F)$, there is a unique projectivity $h\colon \mathbf{P}(E) \to \mathbf{P}(F)$ such that $h(a_i) = b_i$ for $1 \leq i \leq n+2$.*

Proof. Let $(u_1, \ldots, u_{n+1})$ be a basis of E associated with the projective frame $(a_i)_{1\leq i\leq n+2}$, and let $(v_1, \ldots, v_{n+1})$ be a basis of F associated with the projective frame $(b_i)_{1\leq i\leq n+2}$. Since $(u_1, \ldots, u_{n+1})$ is a basis, there is a unique linear bijection $g\colon E \to F$ such that $g(u_i) = v_i$, for $1 \leq i \leq n+1$. Clearly, $h = \mathbf{P}(g)$ is a projectivity such that $h(a_i) = b_i$, for $1 \leq i \leq n+2$. Let $h'\colon \mathbf{P}(E) \to \mathbf{P}(F)$ be any projectivity such that $h'(a_i) = b_i$, for $1 \leq i \leq n+2$. By definition, there is a linear isomorphism $f\colon E \to F$ such that $h' = \mathbf{P}(f)$. Since $h'(a_i) = b_i$, for $1 \leq i \leq n+2$, we must have $f(u_i) = \lambda_i v_i$, for some $\lambda_i \in K - \{0\}$, where $1 \leq i \leq n+1$, and

$$f(u_1 + \cdots + u_{n+1}) = \lambda(v_1 + \cdots + v_{n+1}),$$

for some $\lambda \in K - \{0\}$. By linearity of f, we have

$$\lambda_1 v_1 + \cdots + \lambda_{n+1} v_{n+1} = \lambda v_1 + \cdots + \lambda v_{n+1},$$

and since $(v_1, \ldots, v_{n+1})$ is a basis of F, we must have

$$\lambda_1 = \cdots = \lambda_{n+1} = \lambda.$$

This shows that $f = \lambda g$, and thus that

$$h' = \mathbf{P}(f) = \mathbf{P}(g) = h,$$

and h is uniquely determined. $\square$

The above proposition and Proposition 26.4 are false if K is a skew field. Also, Proposition 26.5 fails if $(b_i)_{1\leq i\leq n+2}$ is not a projective frame, or if a_{n+2} is dropped.

As a corollary of Proposition 26.5, given a projective space $\mathbf{P}(E)$, two distinct projective lines D and D' in $\mathbf{P}(E)$, three distinct points a, b, c on D, and any three distinct points a', b', c' on D', there is a unique projectivity from D to D', mapping a to a', b to b', and c to c'. This is because, as we mentioned earlier, any three distinct points on a line form a projective frame.

Remark: As in the affine case, there is "fundamental theorem of projective geometry." For simplicity, we state this theorem assuming that vector spaces are over the field $K = \mathbb{R}$. Given any two projective spaces $\mathbf{P}(E)$ and $\mathbf{P}(F)$ of the same dimension $n \geq 2$, for any bijective function $f\colon \mathbf{P}(E) \to \mathbf{P}(F)$, if f maps any three distinct collinear points a, b, c to collinear points $f(a), f(b), f(c)$, then f is a projectivity. For more general fields, $f = \mathbf{P}(g)$ for some "semilinear" bijection $g\colon E \to F$. A map such as f (preserving collinearity of any three distinct points) is often called a *collineation*. For $K = \mathbb{R}$, collineations and projectivities coincide. For more details, see Samuel [142].

Before closing this section, we illustrate the power of Proposition 26.5 by proving two interesting results. We begin by characterizing perspectivities between lines.

Proposition 26.6. *Given any two distinct lines D and D' in the real projective plane $\mathbb{RP}^2$, a projectivity $f\colon D \to D'$ is a perspectivity iff $f(O) = O$, where O is the intersection of D and D'.*

Proof. If $f\colon D \to D'$ is a perspectivity, then by the very definition of f, we have $f(O) = O$. Conversely, let $f\colon D \to D'$ be a projectivity such that $f(O) = O$. Let a, b be any two distinct points on D also distinct from O, and let $a' = f(a)$ and $b' = f(b)$ on D'. Since f is a bijection and since a, b, O are pairwise distinct, $a' \neq b'$. Let c be the intersection of the lines $\langle a, a'\rangle$ and $\langle b, b'\rangle$, which by the assumptions on a, b, O, cannot be on D or D'. Then we can define the perspectivity $g\colon D \to D'$ of center c, and by the definition of c, we have

$$g(a) = a', \quad g(b) = b', \quad g(O) = O.$$

See Figure 26.12. However, f agrees with g on O, a, b, and since (O, a, b) is a projective frame for D, by Proposition 26.5, we must have $f = g$. $\square$

Using Proposition 26.6, we can give an elegant proof of a version of Desargues's theorem (in the plane).

Proposition 26.7. *(Desargues) Given two triangles (a, b, c) and (a', b', c') in $\mathbb{RP}^2$, where the points a, b, c, a', b', c' are pairwise distinct and the lines $A = \langle b, c\rangle$, $B = \langle a, c\rangle$, $C = \langle a, b\rangle$, $A' = \langle b', c'\rangle$, $B' = \langle a', c'\rangle$, $C' = \langle a', b'\rangle$ are pairwise distinct, if the lines $\langle a, a'\rangle$, $\langle b, b'\rangle$, and $\langle c, c'\rangle$ intersect in a common point d distinct from a, b, c, a', b', c', then the intersection points $p = \langle b, c\rangle \cap \langle b', c'\rangle$, $q = \langle a, c\rangle \cap \langle a', c'\rangle$, and $r = \langle a, b\rangle \cap \langle a', b'\rangle$ belong to a common line distinct from A, B, C, A', B', C'.*

Proof. In view of the assumptions on a, b, c, a', b', c', and d, the point r is on neither $\langle a, a'\rangle$ nor $\langle b, b'\rangle$, the point p is on neither $\langle b, b'\rangle$ nor $\langle c, c'\rangle$, and the point q is on neither $\langle a, a'\rangle$ nor $\langle c, c'\rangle$. It is also immediately shown that the line $\langle p, q\rangle$ is distinct from the lines A, B, C, A', B', C'. Let $f\colon \langle a, a'\rangle \to \langle b, b'\rangle$ be the perspectivity of center r and $g\colon \langle b, b'\rangle \to \langle c, c'\rangle$ be the perspectivity of center p. Let $h = g \circ f$. Since both $f(d) = d$ and $g(d) = d$, we also have

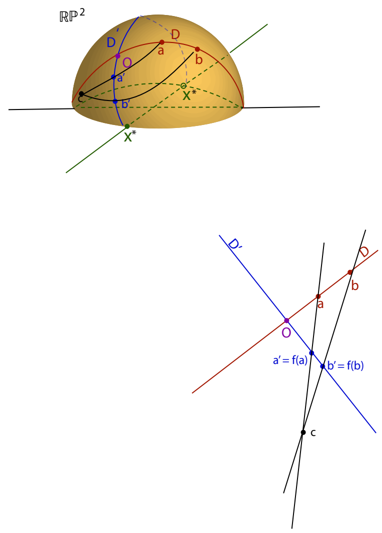


Figure 26.12: An illustration of the perspectivity construction of Proposition 26.6.

$h(d) = d$. Thus by Proposition 26.6, the projectivity $h\colon \langle a, a' \rangle \to \langle c, c' \rangle$ is a perspectivity. Since

$$\begin{aligned} h(a) &= g(f(a)) = g(b) = c, \\ h(a') &= g(f(a')) = g(b') = c', \end{aligned}$$

the intersection q of $\langle a, c \rangle$ and $\langle a', c' \rangle$ is the center of the perspectivity h. Also note that the point $m = \langle a, a' \rangle \cap \langle p, r \rangle$ and its image $h(m)$ are both on the line $\langle p, r \rangle$, since r is the center of f and p is the center of g. Since h is a perspectivity of center q, the line $\langle m, h(m) \rangle = \langle p, r \rangle$ passes through q, which proves the proposition. $\square$

Desargues's theorem is illustrated in Figure 26.13. It can also be shown that every projectivity between two distinct lines is the composition of two perspectivities (not in a unique way). An elegant proof of Pappus's theorem can also be given using perspectivities.

26.6 Finding a Homography Between Two Projective Frames

In this section we present a method for finding the matrix (up to a scalar) of the unique homography (bijective projective transformation) mapping one projective frame to an other projective frame. This problem arises notably in computer vision in the context of image morphing.

We begin with the simple case of two nondegenerate quadrilatrerals $([p_1], [p_2], [p_3], [p_4])$ and $[([q_1], [q_2], [q_3], [q_4])$ in $\mathbb{RP}^2$, that is, two projective frames, which means that (p_1, p_2, p_3)

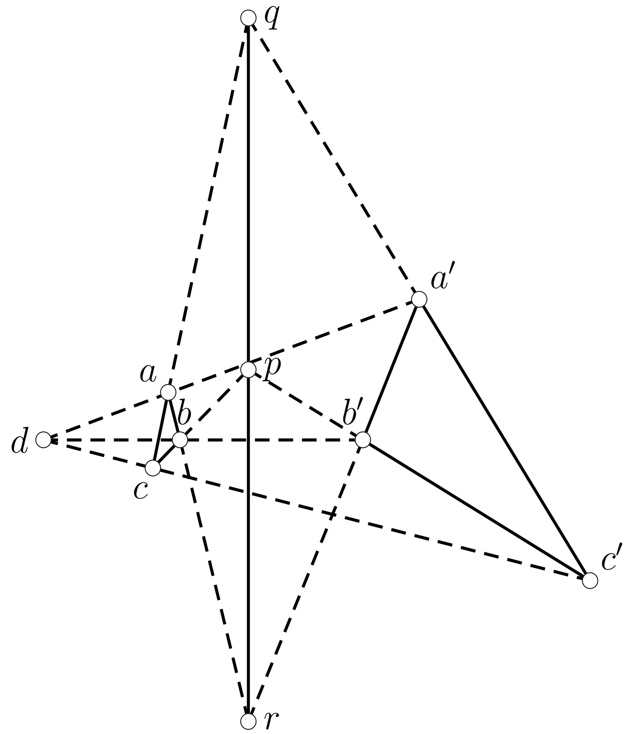


Figure 26.13: Desargues's theorem (projective version in the plane).

and (q_1, q_2, q_3) are linearly independent, and that if we write

$$p_4 = \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3$$

and

$$q_4 = \lambda_1 q_1 + \lambda_2 q_2 + \lambda_3 q_3,$$

for some unique scalars $\alpha_1, \alpha_2, \alpha_3$ and $\lambda_1, \lambda_2, \lambda_3$, then $\alpha_i \neq 0$ and $\lambda_i \neq 0$ for $i = 1, 2, 3$. The problem is to find the 3×3 matrix (up to a scalar) representing the unique homography h mapping $[p_i]$ to $[q_i]$ for $i = 1, 2, 3, 4$.

We will use the *canonical basis* $\mathcal{E} = (e_1, e_2, e_3)$ of $\mathbb{R}^3$, with $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$, and the bases $\mathcal{P} = (p_1, p_2, p_3)$ and $\mathcal{Q} = (q_1, q_2, q_3)$ of $\mathbb{R}^3$.

As a first step, it is convenient to express (q_1, q_2, q_3, q_4) over the basis $\mathcal{P} = (p_1, p_2, p_3)$, with $q_1 = (x_1, y_1, z_1)$, $q_2 = (x_2, y_2, z_2)$, $q_3 = (x_3, y_3, z_3)$, $q_4 = (x_4, y_4, z_4)$. Over the canonical basis $\mathcal{E}$, the points (p_1, p_2, p_3, p_4) are given by the coordinates $p_1 = (p_1^x, p_1^y, p_1^z)$, $p_2 = (p_2^x, p_2^y, p_2^z)$, $p_3 = (p_3^x, p_3^y, p_3^z)$, $p_4 = (p_4^x, p_4^y, p_4^z)$, and similarly, the points (q_1, q_2, q_3, q_4) are given by the coordinates $q_1 = (q_1^x, q_1^y, q_1^z)$, $q_2 = (q_2^x, q_2^y, q_2^z)$, $q_3 = (q_3^x, q_3^y, q_3^z)$, $q_4 = (q_4^x, q_4^y, q_4^z)$.

Proposition 26.8. *With respect to the basis $\mathcal{P} = (p_1, p_2, p_3)$, the matrix $A_{\mathcal{P}}$ of the unique homography h of $\mathbb{RP}^2$ mapping the projective frame $([p_1], [p_2], [p_3], [p_4])$ to the projective frame $[(q_1], [q_2], [q_3], [q_4])$ is given by*

$$A_{\mathcal{P}} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix}.$$

Proof. Let $u_1 = \alpha_1 p_1$, $u_2 = \alpha_2 p_2$, $u_3 = \alpha_3 p_3$, and let $v_1 = \lambda_1 q_1$, $v_2 = \lambda_2 q_2$, $v_3 = \lambda_3 q_3$, so that

$$p_4 = u_1 + u_2 + u_3$$

and

$$q_4 = v_1 + v_2 + v_3.$$

Because p_1, p_2, p_3 are linearly independent and since $\alpha_i \neq 0$ for $i = 1, 2, 3$, the vectors (u_1, u_2, u_2) are also linearly independent, so there is a unique linear map $f \colon \mathbb{R}^3 \to \mathbb{R}^3$, such that

$$f(u_i) = v_i \quad i = 1, \ldots, 3,$$

and by linearity

$$f(p_4) = f(u_1 + u_2 + u_3) = f(u_1) + f(u_2) + f(u_3) = v_1 + v_2 + v_3 = q_4.$$

With respect to the basis $\mathcal{P} = (p_1, p_2, p_3)$, we have

$$f(p_i) = \frac{1}{\alpha_i} v_i = \frac{\lambda_i}{\alpha_i} q_i, \quad i = 1, \ldots, 3,$$

so with respect to the basis $\mathcal{P}$, the matrix of f is

$$A_{\mathcal{P}} = \begin{pmatrix} \frac{\lambda_1}{\alpha_1} x_1 & \frac{\lambda_2}{\alpha_2} x_2 & \frac{\lambda_3}{\alpha_3} x_3 \\ \frac{\lambda_1}{\alpha_1} y_1 & \frac{\lambda_2}{\alpha_2} y_2 & \frac{\lambda_3}{\alpha_3} y_3 \\ \frac{\lambda_1}{\alpha_1} z_1 & \frac{\lambda_2}{\alpha_2} z_2 & \frac{\lambda_3}{\alpha_3} z_3 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix},$$

as claimed. ∎

If we assume that we pick the coordinates of (p_1, p_2, p_3, p_4) and (q_1, q_2, q_3, q_4) with respect to the canonical basis $\mathcal{E}$, then the coordinates $\alpha_1, \alpha_2, \alpha_3$ and $\lambda_1, \lambda_2, \lambda_3$ are solutions of the systems

$$\begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ p_1^z & p_2^z & p_3^z \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} p_4^x \\ p_4^y \\ p_4^z \end{pmatrix}$$

and
$$\begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} q_4^x \\ q_4^y \\ q_4^z \end{pmatrix},$$
and the matrix $A_{\mathcal{E}}$ of our linear map f with respect to the canonical basis is determined as follows.

Proposition 26.9. *With respect to the canonical basis $\mathcal{E} = (e_1, e_2, e_3)$, the matrix $A_{\mathcal{E}}$ of the unique homography h of $\mathbb{RP}^2$ mapping the projective frame $([p_1], [p_2], [p_3], [p_4])$ to the projective frame $[(q_1], [q_2], [q_3], [q_4])$ is given by*
$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix} \begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ p_1^z & p_2^z & p_3^z \end{pmatrix}^{-1}.$$

Proof. Since $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ is the unique linear map given by
$$f(u_i) = v_i, \quad i = 1, \ldots, 3,$$
the map $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ is equal to the composition
$$f = f_{\mathcal{Q}} \circ g,$$
where $g\colon \mathbb{R}^3 \to \mathbb{R}^3$ is the unique linear map given by
$$g(u_i) = e_i, \quad i = 1, \ldots, 3,$$
and $f_{\mathcal{Q}}\colon \mathbb{R}^3 \to \mathbb{R}^3$ is the unique linear map given by
$$f_{\mathcal{Q}}(e_i) = v_i, \quad i = 1, \ldots, 3.$$
However, $g\colon \mathbb{R}^3 \to \mathbb{R}^3$ is the inverse of the unique linear map $f_{\mathcal{P}}\colon \mathbb{R}^3 \to \mathbb{R}^3$ given by
$$f_{\mathcal{P}}(e_i) = u_i, \quad i = 1, \ldots, 3,$$
so
$$f = f_{\mathcal{Q}} \circ f_{\mathcal{P}}^{-1}.$$
The matrix $B_{\mathcal{P}}$ representing $f_{\mathcal{P}}$ over the canonical basis $\mathcal{E}$ is
$$B_{\mathcal{P}} = \begin{pmatrix} \alpha_1 p_1^x & \alpha_2 p_2^x & \alpha_3 p_3^x \\ \alpha_1 p_1^y & \alpha_2 p_2^y & \alpha_3 p_3^y \\ \alpha_1 p_1^z & \alpha_2 p_2^z & \alpha_3 p_3^z \end{pmatrix} = \begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ p_1^z & p_2^z & p_3^z \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{pmatrix},$$

and similarly the matrix $B_{\mathcal{Q}}$ representing $f_{\mathcal{Q}}$ over $\mathcal{E}$ is

$$B_{\mathcal{Q}} = \begin{pmatrix} \lambda_1 q_1^x & \lambda_2 q_2^x & \lambda_3 q_3^x \\ \lambda_1 q_1^y & \lambda_2 q_2^y & \lambda_3 q_3^y \\ \lambda_1 q_1^z & \lambda_2 q_2^z & \lambda_3 q_3^z \end{pmatrix} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix},$$

and we have

$$A_{\mathcal{E}} = B_{\mathcal{Q}} B_{\mathcal{P}}^{-1}.$$

Therefore, we have

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix} \begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ p_1^z & p_2^z & p_3^z \end{pmatrix}^{-1},$$

as claimed ☐

The above method generalizes immediately to any dimension (and any field K). If $([p_1], \ldots, [p_{n+1}], [p_{n+2}])$ and $[(q_1], \ldots, [q_{n+1}], [q_{n+2}])$ are any two projective frames in a projective space $\mathbb{P}(E)$ where E is a K-vector space of dimension $n+1$, then $(p_1, \ldots, p_{n+1})$ is a basis of E denoted by $\mathcal{P}$ and $(q_1, \ldots, q_{n+1})$ is a basis of E denoted $\mathcal{Q}$, and we can write

$$\begin{aligned} p_{n+2} &= \alpha_1 p_1 + \cdots + \alpha_{n+1} p_{n+1} \\ q_{n+2} &= \lambda_1 q_1 + \cdots + \lambda_{n+1} q_{n+1} \end{aligned}$$

for some unique $\alpha_i, \lambda_i \in K$ such that $\alpha_i \neq 0$ and $\lambda_i \neq 0$ for $i = 1, \ldots, n+1$. If we assume that $E = K^{n+1}$, then the canonical basis is $\mathcal{E} = (e_1, \ldots, e_{n+1})$.

If we express the coordinates of the q_j over the basis $\mathcal{P}$ by

$$q_j = (x_j^1, \ldots, x_j^n, x_j^{n+1}), \quad j = 1, \ldots, n+2,$$

then we have the following proposition.

Proposition 26.10. *With respect to the basis $\mathcal{P} = (p_1, \ldots, p_{n+1})$, the matrix $A_{\mathcal{P}}$ of the unique homography h of $\mathbb{P}(E)$ where E is a K-vector space of dimension $n+1$, mapping the projective frame $([p_1], \ldots, [p_{n+1}], [p_{n+2}])$ to the projective frame $[(q_1], \ldots, [q_{n+1}], [q_{n+2}])$ is given by*

$$A_{\mathcal{P}} = \begin{pmatrix} x_1^1 & \ldots & x_n^1 & x_{n+1}^1 \\ \vdots & \ddots & \vdots & \vdots \\ x_1^n & \ldots & x_n^n & x_{n+1}^n \\ x_1^{n+1} & \ldots & x_n^{n+1} & x_{n+1}^{n+1} \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & \frac{\lambda_n}{\alpha_n} & 0 \\ 0 & \ldots & 0 & \frac{\lambda_{n+1}}{\alpha_{n+1}} \end{pmatrix}.$$

If we express the coordinates of the vectors p_i and q_i over the canonical basis as

$$p_i = (p_i^1, \ldots, p_i^n, p_i^{n+1}), \qquad q_i = (q_i^1, \ldots, q_i^n, q_i^{n+1}), \quad i = 1, \ldots, n+2,$$

then we have the following result.

Proposition 26.11. *With respect to the canonical basis $\mathcal{E} = (e_1, \ldots, e_{n+1})$, the matrix $A_{\mathcal{E}}$ of the unique homography h of $\mathbb{P}(E)$ where E is a K-vector space of dimension $n+1$, mapping the projective frame $([p_1], \ldots, [p_{n+1}], [p_{n+2}])$ to the projective frame $[([q_1], \ldots, [q_{n+1}], [q_{n+2}])$ is given by*

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^1 & \cdots & q_n^1 & q_{n+1}^1 \\ \vdots & \ddots & \vdots & \vdots \\ q_1^n & \cdots & q_n^n & q_{n+1}^n \\ q_1^{n+1} & \cdots & q_n^{n+1} & q_{n+1}^{n+1} \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{\lambda_n}{\alpha_n} & 0 \\ 0 & \cdots & 0 & \frac{\lambda_{n+1}}{\alpha_{n+1}} \end{pmatrix} \begin{pmatrix} p_1^1 & \cdots & p_n^1 & p_{n+1}^1 \\ \vdots & \ddots & \vdots & \vdots \\ p_1^n & \cdots & p_n^n & p_{n+1}^n \\ p_1^{n+1} & \cdots & p_n^{n+1} & p_{n+1}^{n+1} \end{pmatrix}^{-1},$$

where $(\alpha_1, \ldots, \alpha_{n+1})$ and $(\lambda_1, \ldots, \lambda_{n+1})$ are the solutions of the systems

$$\begin{pmatrix} p_1^1 & \cdots & p_n^1 & p_{n+1}^1 \\ \vdots & \ddots & \vdots & \vdots \\ p_1^n & \cdots & p_n^n & p_{n+1}^n \\ p_1^{n+1} & \cdots & p_n^{n+1} & p_{n+1}^{n+1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \alpha_{n+1} \end{pmatrix} = \begin{pmatrix} p_{n+2}^1 \\ \vdots \\ p_{n+2}^n \\ p_{n+2}^{n+1} \end{pmatrix}$$

and

$$\begin{pmatrix} q_1^1 & \cdots & q_n^1 & q_{n+1}^1 \\ \vdots & \ddots & \vdots & \vdots \\ q_1^n & \cdots & q_n^n & q_{n+1}^n \\ q_1^{n+1} & \cdots & q_n^{n+1} & q_{n+1}^{n+1} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} q_{n+2}^1 \\ \vdots \\ q_{n+2}^n \\ q_{n+2}^{n+1} \end{pmatrix}.$$

We now consider the special case where the points $([p_1], [p_2], [p_3], [p_4])$ belong to the affine patch of $\mathbb{RP}^2$ corresponding to the plane H of equation $z = 1$. In this case, we may identify $[p_i]$ with p_i, which has coordinates $(p^i, p_i^y, 1)$ with respect to the canonical basis (the p_is are *not* points at infinity; points at infinity are of of form $(x, y, 0)$). Then, the barycentric coordinates $\alpha_1, \alpha_2, \alpha_3$ of p_4 are solutions of the systems

$$\begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} p_4^x \\ p_4^y \\ 1 \end{pmatrix}.$$

By Proposition 26.9, we obtain the following result.

Proposition 26.12. *With respect to the canonical basis $\mathcal{E} = (e_1, e_2, e_3)$, the matrix $A_{\mathcal{E}}$ of the unique homography h of $\mathbb{RP}^2$ mapping (p_1, p_2, p_4, p_4), points of the affine plane $z = 1$, to $[([q_1], [q_2], [q_3], [q_4])$ is given by*

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix} \begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ 1 & 1 & 1 \end{pmatrix}^{-1}.$$

Observe that the above homography may map some of the affine points p_1, p_2, p_3, p_4 (which are not "points at infinity") to arbitrary points in $\mathbb{RP}^2$, which may be points at infinity (in which case $q_i^z = 0$). The generalization to any dimension $n \geq 2$ is immediate.

We define the basis $\mathcal{E}^a = (e_1^a, e_2^a, e_3^a)$, with $e_1^a = (1, 0, 1)$, $e_2^a = (0, 1, 1)$, $e_3^a = (0, 0, 1)$, and call it the *affine canonical basis* (of $\mathbb{R}^2$). We also define e_4^a as $e_4^a = (1, 1, 1)$.

In the special case where (p_1, p_2, p_3, p_4) is the canonical square $(e_1^a, e_2^a, e_3^a, e_4^a)$, since

$$e_4^a = e_1^a + e_2^a - e_3^a,$$

we have $\alpha_1 = 1, \alpha_2 = 1$, and $\alpha_3 = -1$, so

$$\mathcal{B}_{\mathcal{P}} = \mathcal{B}_{\mathcal{E}^a} = P \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

where P is the change of basis matrix from the canonical basis $\mathcal{E} = (e_1, e_2, e_3)$ to the affine basis $\mathcal{E}^a = (e_1^a, e_2^a, e_3^a)$. We have

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix},$$

and its inverse is

$$P^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}.$$

In this case,

$$\mathcal{B}_{\mathcal{E}^a} = \begin{pmatrix} \alpha_1 p_1^x & \alpha_2 p_2^x & \alpha_3 p_3^x \\ \alpha_1 p_1^y & \alpha_2 p_2^y & \alpha_3 p_3^y \\ \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix},$$

and since

$$\mathcal{B}_{\mathcal{E}^a}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix} = \mathcal{B}_{\mathcal{E}^a},$$

we obtain

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & -\lambda_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix},$$

that is,

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ q_1^z & q_2^z & q_3^z \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

The generalization to any dimension $n \geq 2$ is immediate.

Finally, we consider the special case where the points $([p_1], [p_2], [p_3], [p_4])$ and the points $[([q_1], [q_2], [q_3], [q_4])$ belong to the affine patch of $\mathbb{RP}^2$ corresponding to the plane H of equation $z = 1$. In this case, we may also identify $[q_i]$ with q_i, which has coordinates $(q_i^x, q_i^y, 1)$ with respect to the canonical basis. Then, the barycentric coordinates $\lambda_1, \lambda_2, \lambda_3$ of q_4 are solutions of the systems

$$\begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} q_4^x \\ q_4^y \\ 1 \end{pmatrix}.$$

By Proposition 26.12 we obtain the following result.

Proposition 26.13. *With respect to the canonical basis $\mathcal{E} = (e_1, e_2, e_3)$, the matrix $A_{\mathcal{E}}$ of the unique homography h of $\mathbb{RP}^2$ mapping (p_1, p_2, p_4, p_4) to (q_1, q_2, q_3, q_4), all points of the affine plane $z = 1$, is given by*

$$A_{\mathcal{E}} = \begin{pmatrix} q_1^x & q_2^x & q_3^x \\ q_1^y & q_2^y & q_3^y \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\alpha_1} & 0 & 0 \\ 0 & \frac{\lambda_2}{\alpha_2} & 0 \\ 0 & 0 & \frac{\lambda_3}{\alpha_3} \end{pmatrix} \begin{pmatrix} p_1^x & p_2^x & p_3^x \\ p_1^y & p_2^y & p_3^y \\ 1 & 1 & 1 \end{pmatrix}^{-1}.$$

If

$$A_{\mathcal{E}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

the transformed point of a point $(x, y, 1)$ in the affine plane $z = 1$,

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11}x + a_{12}y + a_{13} \\ a_{21}x + a_{22}y + a_{23} \\ a_{31}x + a_{32}y + a_{33} \end{pmatrix},$$

is not a point at infinity iff $a_{31}x + a_{32}y + a_{33} \neq 0$, in which case it corresponds to the point in the affine plane $z = 1$ of coordinates

$$\begin{pmatrix} \dfrac{x'}{z'} \\[2ex] \dfrac{y'}{z'} \\[2ex] 1 \end{pmatrix} = \begin{pmatrix} \dfrac{a_{11}x + a_{12}y + a_{13}}{a_{31}x + a_{32}y + a_{33}} \\[2ex] \dfrac{a_{21}x + a_{22}y + a_{23}}{a_{31}x + a_{32}y + a_{33}} \\[2ex] 1 \end{pmatrix}.$$

The generalization to any dimension $n \geq 2$ is immediate.

Let us go back to the situation where the the points (p_1, p_2, p_3, p_4) and (q_1, q_2, q_3, q_4) are in the affine patch $z = 1$, and where the matrix of our linear map is expressed with respect to the basis $\mathcal{P} = (p_1, p_2, p_3)$ and the coordinates of (q_1, q_2, q_3, q_4) are also expressed with respect to the basis $\mathcal{P} = (p_1, p_2, p_3)$. In practical situations, for example in computer vision, it is important to find necessary and sufficient conditions for the unique projective transformation mapping (p_1, p_2, p_3, p_4) to (q_1, q_2, q_3, q_4) to be defined on the convex hull of the points p_1, p_2, p_3, p_4.

Proposition 26.14. *The unique projective transformation mapping (p_1, p_2, p_3, p_4) to (q_1, q_2, q_3, q_4) (all points in the affine plane H of equation $z = 1$) is defined on the convex hull of the points p_1, p_2, p_3, p_4 iff the scalars in each of the pairs (α_1, λ_1), (α_2, λ_2) and (α_3, λ_3), have the same sign.*

Proof. With respect to the basis $\mathcal{P}$, the equation of the plane H is

$$x + y + z = 1,$$

so the image of $p = (x, y, 1 - x - y)$ under our linear map is

$$\begin{pmatrix} \frac{\lambda_1}{\alpha_1} x_1 & \frac{\lambda_2}{\alpha_2} x_2 & \frac{\lambda_3}{\alpha_3} x_3 \\ \frac{\lambda_1}{\alpha_1} y_1 & \frac{\lambda_2}{\alpha_2} y_2 & \frac{\lambda_3}{\alpha_3} y_3 \\ \frac{\lambda_1}{\alpha_1} & \frac{\lambda_2}{\alpha_2} & \frac{\lambda_3}{\alpha_3} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 - x - y \end{pmatrix}.$$

The above point is a point at infinity iff

$$\left(\frac{\lambda_1}{\alpha_1} - \frac{\lambda_3}{\alpha_3}\right) x + \left(\frac{\lambda_2}{\alpha_2} - \frac{\lambda_3}{\alpha_3}\right) y + \frac{\lambda_3}{\alpha_3} = 0. \tag{$*$}$$

The unique projective transformation mapping (p_1, p_2, p_3, p_4) to (q_1, q_2, q_3, q_4) is defined on the convex hull of the points p_1, p_2, p_3, p_4 iff all four points p_1, p_2, p_3, p_4 are strictly contained in one of the two open half spaces determined by the line of equation $(*)$, which means that the affine form in $(*)$ must have the same sign on these four points.

When we evaluate the affine form in $(*)$ on the four points p_1, p_2, p_3, p_4 using coordinates $(x, y, 1 - x - y)$, w.r.t. the basis $\mathcal{P} = (p_1, p_2, p_3)$,

1. for $p_1 = (1, 0, 0)$ we get λ_1/α_1,
2. for $p_2 = (0, 1, 0)$ we get λ_2/α_2,
3. for $p_3 = (0, 0, 1)$ we get λ_3/α_3,
4. and for $p_4 = (\alpha_1, \alpha_2, \alpha_3)$ we get

$$\begin{aligned} \left(\frac{\lambda_1}{\alpha_1} - \frac{\lambda_3}{\alpha_3}\right) \alpha_1 + \left(\frac{\lambda_2}{\alpha_2} - \frac{\lambda_3}{\alpha_3}\right) \alpha_2 + \frac{\lambda_3}{\alpha_3} &= \lambda_1 + \lambda_2 + \frac{\lambda_3}{\alpha_3}(1 - \alpha_1 - \alpha_2) \\ &= \lambda_1 + \lambda_2 + \lambda_3 = 1. \end{aligned}$$

The fourth case shows that the sign of the affine form in $(*)$ is positive, and thus $\lambda_1/\alpha_1, \lambda_2/\alpha_2, \lambda_3/\alpha_3 > 0$, which implies that the scalars in each of the pairs (α_1, λ_1), (α_2, λ_2) and (α_3, λ_3), must have the same sign. $\square$

The generalization to any dimension $n \geq 2$ is immediate: the scalars in each pair (α_i, λ_i) must have the same sign for $i = 1, \ldots, n+2$.

In dimension 2, since $\alpha_3 = 1 - \alpha_1 - \alpha_2$ and $\lambda_3 = 1 - \lambda_1 - \lambda_2$, there are four cases to consider:

(1) $\alpha_1, \lambda_1, \alpha_2, \lambda_2 < 0$. In this case, $\alpha_3, \lambda_3 > 1$ so α_3, λ_3 also have the same sign.

(2) $\alpha_1, \lambda_1 < 0$ and $\alpha_2, \lambda_2 > 0$. In this case, since $\alpha_3 = 1 - \alpha_1 - \alpha_2$ and $\lambda_3 = 1 - \lambda_1 - \lambda_2$, we must have either both $\alpha_1 + \alpha_2 < 1$ and $\lambda_1 + \lambda_2 < 1$, or both $\alpha_1 + \alpha_2 > 1$ and $\lambda_1 + \lambda_2 > 1$, in order for α_3 and λ_3 to have the same sign.

(3) $\alpha_1, \lambda_1 > 0$ and $\alpha_2, \lambda_2 < 0$. As in the previous case, since $\alpha_3 = 1 - \alpha_1 - \alpha_2$ and $\lambda_3 = 1 - \lambda_1 - \lambda_2$, we must have either both $\alpha_1 + \alpha_2 < 1$ and $\lambda_1 + \lambda_2 < 1$, or both $\alpha_1 + \alpha_2 > 1$ and $\lambda_1 + \lambda_2 > 1$, in order for α_3 and λ_3 to have the same sign.

(4) $\alpha_1, \lambda_1, \alpha_2, \lambda_2 > 0$. As in the previous case, since $\alpha_3 = 1 - \alpha_1 - \alpha_2$ and $\lambda_3 = 1 - \lambda_1 - \lambda_2$, we must have either both $\alpha_1 + \alpha_2 < 1$ and $\lambda_1 + \lambda_2 < 1$, or both $\alpha_1 + \alpha_2 > 1$ and $\lambda_1 + \lambda_2 > 1$, in order for α_3 and λ_3 to have the same sign.

Since $\alpha_3 = 1 - \alpha_1 - \alpha_2$ and $\lambda_3 = 1 - \lambda_1 - \lambda_2$, we can write

$$\begin{aligned} p_4 &= \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 = p_3 + \alpha_1(p_1 - p_3) + \alpha_2(p_2 - p_3) \\ q_4 &= \lambda_1 q_1 + \lambda_2 q_2 + \lambda_3 q_3 = q_3 + \lambda_1(q_1 - q_3) + \lambda_2(q_2 - q_3). \end{aligned}$$

In the affine frame $(p_3, (p_1 - p_3, p_2 - p_3))$, points have coordinates (α_1, α_2), and in the affine frame $(q_3, (q_1 - q_3, q_2 - q_3))$, points have coordinates (λ_1, λ_2). In the first affine frame, the line $\langle p_1, p_2 \rangle$ is given by the equation $\alpha_1 + \alpha_2 = 1$, and in the second affine frame, the line $\langle q_1, q_2 \rangle$ is given by the equation $\lambda_1 + \lambda_2 = 1$. The open half plane containing p_3 and bounded by the line $\langle p_1, p_2 \rangle$ corresponds to the points of coordinates (α_1, α_2) satisfying $\alpha_1 + \alpha_2 < 1$, and the other open half plane not containing p_3 corresponds to the points of coordinates (α_1, α_2) satisfying $\alpha_1 + \alpha_2 > 1$. Similarly, the open half plane containing q_3 and bounded by the line $\langle q_1, q_2 \rangle$ corresponds to the points of coordinates (λ_1, λ_2) satisfying $\lambda_1 + \lambda_2 < 1$, and the other open half plane not containing q_3 corresponds to the points of coordinates (λ_1, λ_2) satisfying $\lambda_1 + \lambda_2 > 1$.

Then, the above conditions have the following interpretation in terms of regions in the affine plane $z = 1$:

(1) When $\alpha_1 < 0$ and $\alpha_2 < 0$, the point p_4 lies in quadrant III (with respect to the affine frames $(p_3, (p_1 - p_3, p_2 - p_3))$). Under the mapping f, the point q_4 is also mapped to quadrant III (with respect to the affine frame $(q_3, (q_1 - q_3, q_2 - q_3))$); see Figure 26.14.

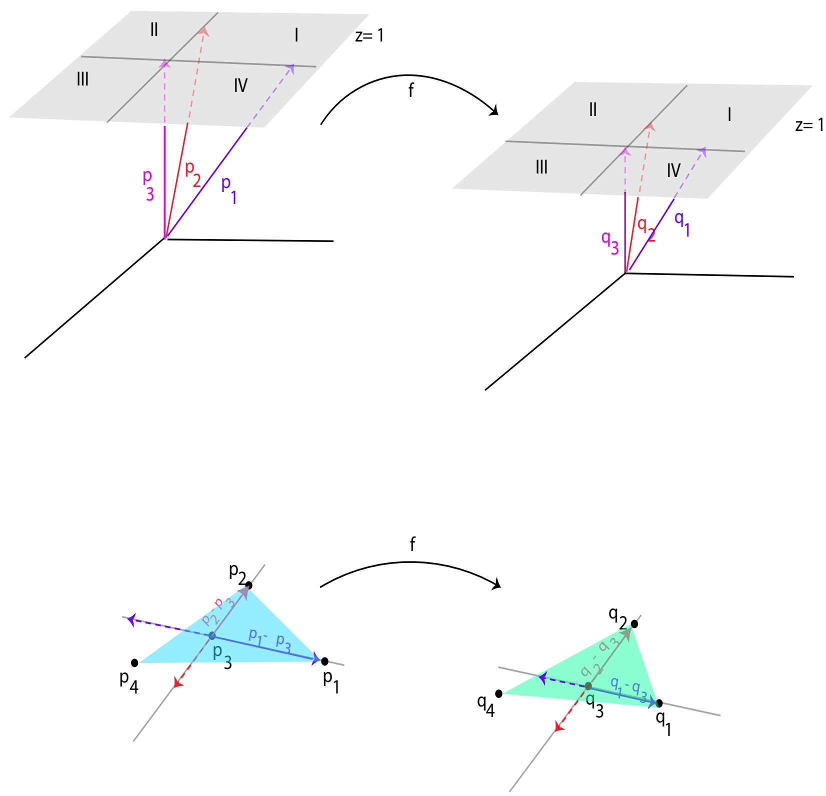


Figure 26.14: Case (1)

(2) When $\alpha_1, \lambda_1 < 0$ and $\alpha_2, \lambda_2 > 0$, the points p_4 and q_4 belongs to quadrant II (with respect to the affine frames $(p_3, (p_1 - p_3, p_2 - p_3))$ and $(q_3, (q_1 - q_3, q_2 - q_3)))$. Two possibilities occur. Either p_4 belong to the open half space containing p_3 and bounded by the line $\langle p_1, p_2 \rangle$ and q_4 belong to the open half space containing q_3 and bounded by the line $\langle q_1, q_2 \rangle$, or p_4 belong to the open half space not containing p_3 and bounded by the line $\langle p_1, p_2 \rangle$ and q_4 belong to the open half space not containing q_3 and bounded by the line $\langle q_1, q_2 \rangle$. The first possibility is illustrated by the top of Figure 26.15, while the second is illustrated by the bottom of Figure 26.15.

(3) When $\alpha_1, \lambda_1 > 0$ and $\alpha_2, \lambda_2 < 0$, the points p_4 and q_4 belongs to quadrant IV (with respect to the affine frames $(p_3, (p_1 - p_3, p_2 - p_3))$ and $(q_3, (q_1 - q_3, q_2 - q_3)))$. Two possibilities occur exactlty as in Case (2) depending on the position of p_4 with respect to the line $\langle p_1, p_2 \rangle$ and on the position of q_4 with respect to the line $\langle q_1, q_2 \rangle$. The first possibility is illustrated by the top of Figure 26.16, while the second is illustrated by the bottom of Figure 26.16.

(4) When $\alpha_1, \lambda_1, \alpha_2, \lambda > 0$ and $\alpha_2, \lambda_2 < 0$, the points p_4 and q_4 belongs to quadrant I

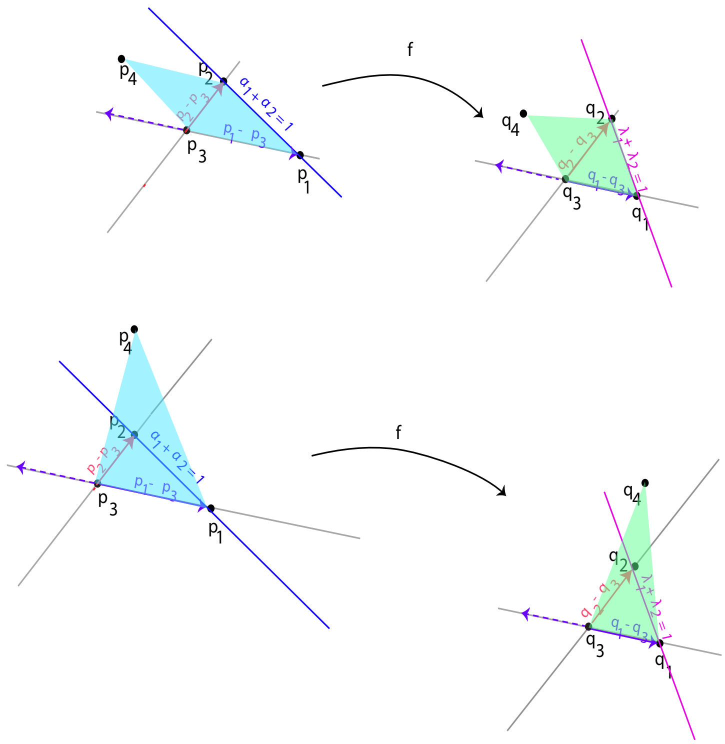


Figure 26.15: Case (2)

(with respect to the affine frames $(p_3, (p_1 - p_3, p_2 - p_3))$ and $(q_3, (q_1 - q_3, q_2 - q_3)))$. Two possibilities occur exactlty as in Cases (2) and (3) depending on the position of p_4 with respect to the line $\langle p_1, p_2 \rangle$ and on the position of q_4 with respect to the line $\langle q_1, q_2 \rangle$. The first possibility is illustrated by the top of Figure 26.17, while the second is illustrated by the bottom of Figure 26.17.

Thus, if both (p_1, p_2, p_3, p_4) and (q_1, q_2, q_3, q_4) satisfy the conditions listed above, there is no point at infinity inside of the convex hull of the quadrangle (p_1, p_2, p_3, p_4).

It remains to prove that the image of the convex hull of (p_1, p_2, p_3, p_4) is the convex hull of (q_1, q_2, q_3, q_4).

Proposition 26.15. *If both (p_1, p_2, p_3, p_4) and (q_1, q_2, q_3, q_4) satisfy the conditions of Proposition 26.14, then the image of the convex hull of (p_1, p_2, p_3, p_4) under the unique projective map mapping (p_1, p_2, p_3, p_4) to (q_1, q_2, q_3, q_4) is the convex hull of (q_1, q_2, q_3, q_4)*

Proof. It suffices to show that the restriction of our projective transformation maps a line segment to the convex hull of the images of the endpoints of this segment. Thus, the problem

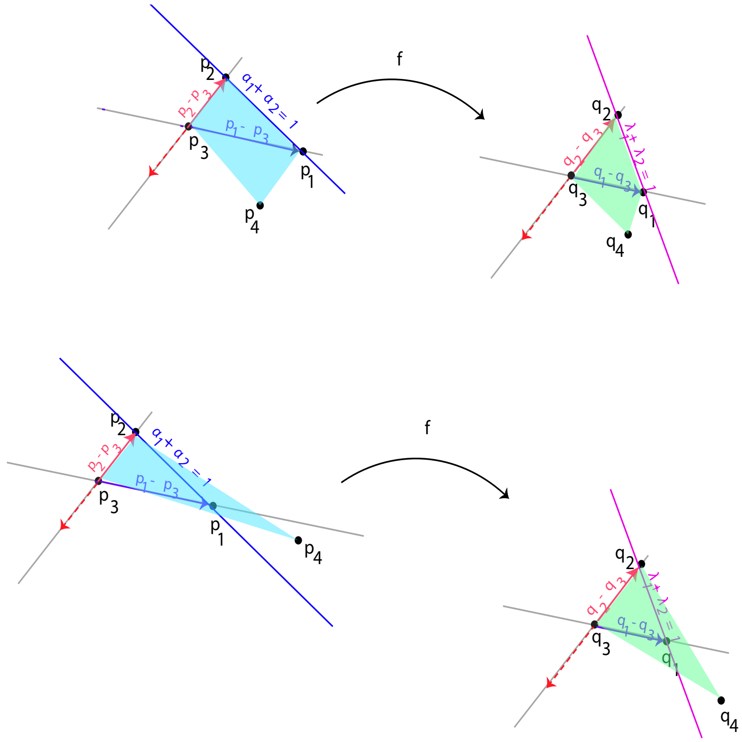


Figure 26.16: Case (3)

reduces to proving that if a projective transformation given by an invertible matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

does not have points at infinity on the line segment in $\mathbb{R}^2$ corresponding to the points of coordinates $(x, 1)$ with $0 \leq x \leq 1$, then the image of the line segment $[(0,1),(1,1)]$ is the line segment $[(b/d, 1), ((a+b)/(c+d), 1)]$ (or $[((a+b)/(c+d), 1), (b/d, 1)]$).

We have

$$\begin{aligned} \frac{ax+b}{cx+d} - \frac{b}{d} &= \frac{adx + bd - bcx - bd}{d(cx+d)} \\ &= \frac{(ad-bc)x}{d(cx+d)} \end{aligned}$$

and

$$\begin{aligned} \frac{ax+b}{cx+d} - \frac{a+b}{c+d} &= \frac{acx + bc + adx + bd - acx - ad - bcx - bd}{(c+d)(cx+d)} \\ &= \frac{(ad-bc)(x-1)}{(c+d)(cx+d)} \end{aligned}$$

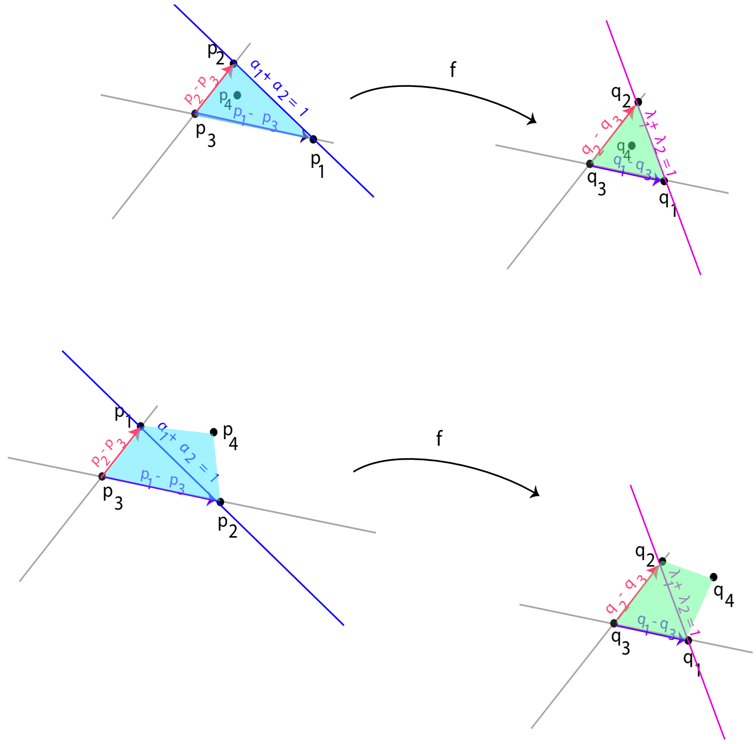


Figure 26.17: Case (4)

In order for our map to be defined for $0 \leq x \leq 1$, $cx + d$ must have a constant sign for $0 \leq x \leq 1$, which means that d and $c + d$ have the same sign. Then,

$$\frac{(ad - bc)x}{d(cx + d)}$$

and

$$\frac{(ad - bc)(x - 1)}{(c + d)(cx + d)}$$

have opposite signs when $0 < x < 1$, which means that the image of $[0, 1]$ is the interval $[b/d, (a + b)/(c + d)]$ (or $[(a + b)/(c + d), b/d]$). $\square$

We now consider the projective completion of an affine space. First, we introduce the notion of affine patch.

26.7 Affine Patches

Given an affine space E with associated vector space $\overrightarrow{E}$, we can form the vector space $\widehat{E}$, the homogenized version of E, and then, the projective space $\mathbf{P}(\widehat{E})$ induced by $\widehat{E}$. This projective space, also denoted by $\widetilde{E}$, has some very interesting properties. In fact, it satisfies a universal property, but before we can say what it is, we have to take a closer look at $\widetilde{E}$.

Since the vector space $\widehat{E}$ is the disjoint union of elements of the form $\langle a, \lambda \rangle$, where $a \in E$ and $\lambda \in K - \{0\}$, and elements of the form $u \in \overrightarrow{E}$, observe that if $\sim$ is the equivalence relation on $\widehat{E}$ used to define the projective space $\mathbf{P}(\widehat{E})$, then the equivalence class $[\langle a, \lambda \rangle]_{\sim}$ of a weighted point contains the special representative $a = \langle a, 1 \rangle$, and the equivalence class $[u]_{\sim}$ of a nonzero vector $u \in \overrightarrow{E}$ is just a point of the projective space $\mathbf{P}(\overrightarrow{E})$. Thus, there is a bijection

$$\mathbf{P}(\widehat{E}) \longleftrightarrow E \cup \mathbf{P}(\overrightarrow{E})$$

between $\mathbf{P}(\widehat{E})$ and the disjoint union $E \cup \mathbf{P}(\overrightarrow{E})$, which allows us to view E as being embedded in $\mathbf{P}(\widehat{E})$. The points of $\mathbf{P}(\widehat{E})$ in $\mathbf{P}(\overrightarrow{E})$ will be called *points at infinity*, and the projective hyperplane $\mathbf{P}(\overrightarrow{E})$ is called the *hyperplane at infinity*. We will also denote the point $[u]_{\sim}$ of $\mathbf{P}(\overrightarrow{E})$ (where $u \neq 0$) by u_{∞}.

Thus, we can think of $\widetilde{E} = \mathbf{P}(\widehat{E})$ as the projective completion of the affine space E obtained by adding points at infinity forming the hyperplane $\mathbf{P}(\overrightarrow{E})$. As we commented in Section 26.2 when we presented the hyperplane model of $\mathbf{P}(E)$, the notion of point at infinity is really an affine notion. But even if a vector space E doesn't arise from the completion of an affine space, there is an affine structure on the complement of any hyperplane $\mathbf{P}(H)$ in the projective space $\mathbf{P}(E)$. In the case of $\widetilde{E}$, the complement E of the projective hyperplane $\mathbf{P}(\overrightarrow{E})$ is indeed an affine space. This is a general property that is needed in order to figure out the universal property of $\widetilde{E}$.

Proposition 26.16. *Given a vector space E and a hyperplane H in E, the complement $E_H = \mathbf{P}(E) - \mathbf{P}(H)$ of the projective hyperplane $\mathbf{P}(H)$ in the projective space $\mathbf{P}(E)$ can be given an affine structure such that the associated vector space of E_H is H. The affine structure on E_H depends only on H, and under this affine structure, E_H is isomorphic to an affine hyperplane in E.*

Proof. Since H is a hyperplane in E, there is some $w \in E - H$ such that $E = Kw \oplus H$. Thus, every vector u in $E - H$ can be written in a unique way as $\lambda w + h$, where $\lambda \neq 0$ and $h \in H$. As a consequence, for every point $[u]$ in E_H, the equivalence class $[u]$ contains a representative of the form $w + \lambda^{-1} h$, with $\lambda \neq 0$. Then we see that the map $\varphi \colon (w + H) \to E_H$, defined such that

$$\varphi(w + h) = [w + h],$$

is a bijection. In order to define an affine structure on E_H, we define $+ \colon E_H \times H \to E_H$ as follows: For every point $[w + h_1] \in E_H$ and every $h_2 \in H$, we let

$$[w + h_1] + h_2 = [w + h_1 + h_2].$$

The axioms of an affine space are immediately verified. Now, $w + H$ is an affine hyperplane is E, and under the affine structure just given to E_H, the map $\varphi \colon (w + H) \to E_H$ is an affine map that is bijective. Thus, E_H is isomorphic to the affine hyperplane $w + H$. If we had chosen a different vector $w' \in E - H$ such that $E = Kw' \oplus H$, then E_H would be isomorphic to the affine hyperplane $w' + H$ parallel to $w + H$. But these two hyperplanes are clearly isomorphic by translation, and thus the affine structure on E_H depends only on H. $\square$

An affine space of the form E_H is called an *affine patch* on $\mathbf{P}(E)$. Proposition 26.16 allows us to view a projective space $\mathbf{P}(E)$ as the result of gluing some affine spaces together, at least when E is of finite dimension. For example, when E is of dimension 2, a hyperplane in E is just a line, and the complement of a point in the projective line $\mathbf{P}(E)$ can be viewed as an affine line. Thus, we can view $\mathbf{P}(E)$ as being covered by two affine lines glued together as illustrated by When $K = \mathbb{R}$, this shows that topologically, the projective line $\mathbb{RP}^1$ is equivalent to a circle. See Figure 26.18. When E is of dimension 3, a hyperplane in E is

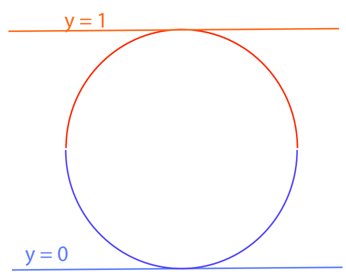


Figure 26.18: The covering of $\mathbb{RP}^1$ by the affine lines $y = 0$ and $y = 1$.

just a plane, and the complement of a projective line in the projective plane $\mathbf{P}(E)$ can be viewed as an affine plane. Thus, we can view $\mathbf{P}(E)$ as being covered by three affine planes glued together as illustrated by Figure 26.19.

However, even when $K = \mathbb{R}$, it is much more difficult to come up with a geometric embedding of the projective plane $\mathbb{RP}^2$ in $\mathbb{A}^3$, and in fact, this is impossible! Nevertheless, there are some fascinating immersions of the projective space $\mathbb{RP}^2$ as $3D$ surfaces with self-intersection, one of which is known as the Boy surface. We urge our readers to consult the remarkable book by Hilbert and Cohn-Vossen [92] for drawings of the Boy surface, and more. One should also consult Fischer's books [61, 60], where many beautiful models of surfaces are displayed, and the commentaries in Chapter 6 of [60] regarding models of $\mathbb{RP}^2$. More generally, when E is of dimension $n+1$, the projective space $\mathbf{P}(E)$ is covered by $n+1$ affine patches (hyperplanes) glued together. This idea is very fruitful, since it allows the treatment of projective spaces as manifolds, and it is essential in algebraic geometry.

We can now go back to the projective completion $\widetilde{E}$ of an affine space E.

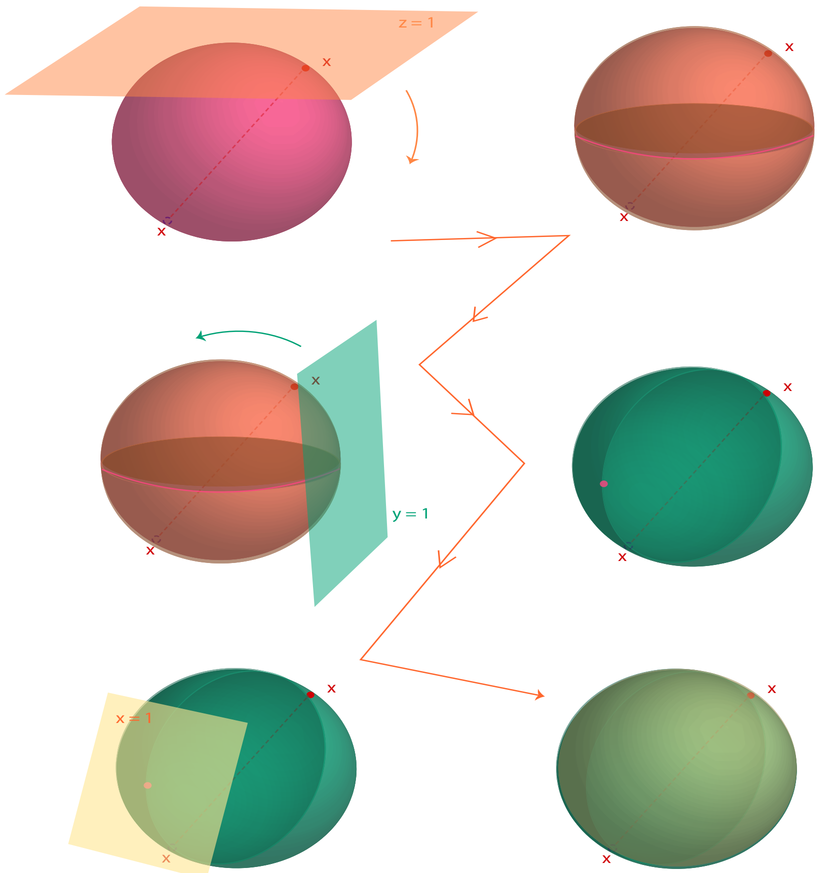


Figure 26.19: The covering of $\mathbb{RP}^2$ by the affine planes $z = 1$, $x = 1$, and $y = 1$. The plane $z = 1$ covers everything but the circle $x^2 + y^2 = 1$ in the xy-plane. The plane $y = 1$ covers that circle modulo the point $(1, 0, 0)$, which is then covered by the plane $x = 1$.

26.8 Projective Completion of an Affine Space

We begin by spelling out the universal property characterizing the projective completion of an affine space $(E, \overrightarrow{E})$. Then, we prove that $\langle \widetilde{E}, \mathbf{P}(\overrightarrow{E}), i\rangle$ where $\widetilde{E} = \mathbf{P}(\widehat{E})$ is the projective space obtained associated with the vector space $\widehat{E}$ obtained from E by the hat construction from Chapter 25 is indeed a projective completion of $(E, \overrightarrow{E})$.

Definition 26.7. Given any affine space E with associated vector space $\overrightarrow{E}$, a *projective completion of the affine space E with hyperplane at infinity* $\mathbf{P}(\mathcal{H})$ is a triple $\langle \mathbf{P}(\mathcal{E}), \mathbf{P}(\mathcal{H}), i\rangle$, where $\mathcal{E}$ is a vector space, $\mathcal{H}$ is a hyperplane in $\mathcal{E}$, $i\colon E \to \mathbf{P}(\mathcal{E})$ is an injective map such that $i(E) = \mathcal{E}_{\mathcal{H}}$ and i is affine (where $\mathcal{E}_{\mathcal{H}} = \mathbf{P}(\mathcal{E}) - \mathbf{P}(\mathcal{H})$ is an affine patch), and for every projective space $\mathbf{P}(F)$ (where F is some vector space), every hyperplane H in F, and every map $f\colon E \to \mathbf{P}(F)$ such that $f(E) \subseteq F_H$ and f is affine (where $F_H = \mathbf{P}(F) - \mathbf{P}(H)$ is an affine patch), there is a unique projective map $\widetilde{f}\colon \mathbf{P}(\mathcal{E}) \to \mathbf{P}(F)$ such that

$$f = \widetilde{f} \circ i \quad \text{and} \quad \mathbf{P}\big(\overrightarrow{f}\big) = \widetilde{f} \circ \mathbf{P}(\overrightarrow{i})$$

(where $\overrightarrow{i}\colon \overrightarrow{E} \to \mathcal{H}$ and $\overrightarrow{f}\colon \overrightarrow{E} \to H$ are the linear maps associated with the affine maps $i\colon E \to \mathbf{P}(\mathcal{E})$ and $f\colon E \to \mathbf{P}(F)$), as in the following diagram:

$$\begin{tikzcd}
E \arrow[r, "i"] \arrow[dr, "f"'] & \mathcal{E}_{\mathcal{H}} \subseteq \mathbf{P}(\mathcal{E}) \supseteq \mathbf{P}(\mathcal{H}) \arrow[d, "\widetilde{f}"] & \mathbf{P}\big(\overrightarrow{E}\big) \arrow[l, "\mathbf{P}(\overrightarrow{i})"'] \arrow[dl, "\mathbf{P}\big(\overrightarrow{f}\big)"] \\
& F_H \subseteq \mathbf{P}(F) \supseteq \mathbf{P}(H) &
\end{tikzcd}$$

The points of $\mathbf{P}(\mathcal{E})$ in $\mathbf{P}(\mathcal{H})$ are called *points at infinity*, and the projective hyperplane $\mathbf{P}(\mathcal{H})$ is called the *hyperplane at infinity*. We will also denote the point $[u]_\sim$ of $\mathbf{P}(\mathcal{H})$ (where $u \neq 0$) by u_∞. As usual, objects defined by a universal property are unique up to isomorphism. We leave the proof as an exercise.

The importance of the notion of projective completion stems from the fact that every affine map $f\colon E \to F$ extends *in a unique way* to a projective map $\widetilde{f}\colon \mathbf{P}(\mathcal{E}) \to \mathbf{P}(\mathcal{F})$, where $\langle \mathbf{P}(\mathcal{E}), \mathbf{P}(\mathcal{H}_E), i_E\rangle$ is a projective completion of E and $\langle \mathbf{P}(\mathcal{F}), \mathbf{P}(\mathcal{H}_F), i_F\rangle$ is a projective completion of F, provided that the restriction of $\widetilde{f}$ to $\mathbf{P}\big(\overrightarrow{E}\big)$ agrees with $\mathbf{P}\big(\overrightarrow{f}\big)$, as illustrated in the following commutative diagram:

$$\begin{tikzcd}
E \arrow[r, "f"] \arrow[d, "i_E"'] & F \arrow[d, "i_F"] \\
\mathbf{P}(\mathcal{E}) \arrow[r, "\widetilde{f}"'] & \mathbf{P}(\mathcal{F}).
\end{tikzcd}$$

We will now show that $\big\langle \widetilde{E}, \mathbf{P}\big(\overrightarrow{E}\big), i\big\rangle$ is the projective completion of E, where $i\colon E \to \widetilde{E}$ is the injection of E into $\widetilde{E} = E \cup \mathbf{P}\big(\overrightarrow{E}\big)$. For example, if $E = \mathbb{A}^1_K$ is an affine line, its projective completion $\widetilde{\mathbb{A}^1_K}$ is isomorphic to the projective line $\mathbf{P}(K^2)$, and they both can be identified with $\mathbb{A}^1_K \cup \{\infty\}$, the result of adding a point at infinity (∞) to $\mathbb{A}^1_K$. In general, the projective completion $\widetilde{\mathbb{A}^m_K}$ of the affine space $\mathbb{A}^m_K$ is isomorphic to $\mathbf{P}(K^{m+1})$. Thus, $\widetilde{\mathbb{A}^m}$ is isomorphic to $\mathbb{RP}^m$, and $\widetilde{\mathbb{A}^m_{\mathbb{C}}}$ is isomorphic to $\mathbb{CP}^m$.

First, let us observe that if E is a vector space and H is a hyperplane in E, then the homogenization $\widehat{E_H}$ of the affine patch E_H (the complement of the projective hyperplane $\mathbf{P}(H)$ in $\mathbf{P}(E)$) is isomorphic to E. The proof is rather simple and uses the fact that there is an affine bijection between E_H and the affine hyperplane $w + H$ in E, where $w \in E - H$ is any fixed vector. Choosing w as an origin in E_H, we know that $\widehat{E_H} = H \widehat{+} Kw$, and since $E = H \oplus Kw$, it is obvious how to define a linear bijection between $\widehat{E_H} = H \widehat{+} Kw$ and $E = H \oplus Kw$. As a consequence the projective spaces $\widetilde{E_H}$ and $\mathbf{P}(E)$ are isomorphic, i.e., there is a projectivity between them.

Proposition 26.17. *Given any affine space $\big(E, \overrightarrow{E}\big)$, for every projective space $\mathbf{P}(F)$ (where F is some vector space), every hyperplane H in F, and every map $f\colon E \to \mathbf{P}(F)$ such that $f(E) \subseteq F_H$ and f is affine (F_H being viewed as an affine patch), there is a unique projective map $\widetilde{f}\colon \widetilde{E} \to \mathbf{P}(F)$ such that*

$$f = \widetilde{f} \circ i \quad \textit{and} \quad \mathbf{P}\big(\overrightarrow{f}\big) = \widetilde{f} \circ \mathbf{P}(\overrightarrow{i}\,),$$

(where $\overrightarrow{i}\colon \overrightarrow{E} \to \overrightarrow{E}$ and $\overrightarrow{f}\colon \overrightarrow{E} \to H$ are the linear maps associated with the affine maps $i\colon E \to \widetilde{E}$ and $f\colon E \to \mathbf{P}(F)$), as in the following diagram:

$$\begin{tikzcd}
E \arrow[r, "i"] \arrow[dr, "f"'] & \mathcal{E}_{\mathcal{H}} \subseteq \mathbf{P}(\mathcal{E}) \supseteq \mathbf{P}(\mathcal{H}) \arrow[d, "\widetilde{f}"] & \mathbf{P}\big(\overrightarrow{E}\big) \arrow[l, "\mathbf{P}(\overrightarrow{i}\,)"'] \arrow[dl, "\mathbf{P}(\overrightarrow{f})"] \\
& F_H \subseteq \mathbf{P}(F) \supseteq \mathbf{P}(H) &
\end{tikzcd}$$

Proof. The existence of $\widetilde{f}$ is a consequence of Proposition 25.6, where we observe that $\widehat{F_H}$ is isomorphic to F. Just take the projective map $\mathbf{P}\big(\widehat{f}\big)\colon \widetilde{E} \to \mathbf{P}(F)$, where $\widehat{f}\colon \widehat{E} \to F$ is the unique linear map extending f. It remains to prove its uniqueness.

As explained in the proof of Proposition 26.16, the affine patch F_H is affinely isomorphic to some affine hyperplane of the form $w + H$ for some $w \in F - H$. If we pick any $a \in E$, since by hypothesis $f(a) \in F_H$, we may assume that $w \in F - H$ is chosen so that $f(a) = [w]$, and we have $F = Kw \oplus H$. Since $f\colon E \to F_H$ is affine, for any $a \in E$ and any $u \in \overrightarrow{E}$, we have

$$f(a + u) = f(a) + \overrightarrow{f}(u) = w + \overrightarrow{f}(u),$$

where $\overrightarrow{f}\colon \overrightarrow{E} \to H$ is a linear map, and where $f(a)$ is viewed as the vector w.

Assume that $\widetilde{f}\colon \widetilde{E} \to \mathbf{P}(F)$ exists with the desired property. Then there is some linear map $g\colon \widehat{E} \to F$ such that $\widetilde{f} = \mathbf{P}(g)$. Our goal is to prove that $g = \mu\widehat{f}$ for some nonzero $\mu \in K$. First, we prove that g vanishes on $\mathrm{Ker}\,\overrightarrow{f}$.

Since $f = \widetilde{f} \circ i$, we must have $f(a) = [w] = [g(a)]$, and thus $g(a) = \mu w$, for some $\mu \neq 0$. Also, for every $u \in \overrightarrow{E}$,

$$\begin{aligned}
f(a + u) &= [w] + \overrightarrow{f}(u) = \big[w + \overrightarrow{f}(u)\big] = [g(a + u)] \\
&= [g(a) + g(u)] = [\mu w + g(u)],
\end{aligned}$$

and thus we must have

$$\lambda(u)w + \lambda(u)\overrightarrow{f}(u) = \mu w + g(u), \tag{$*_1$}$$

for some $\lambda(u) \neq 0$.

If $\mathrm{Ker}\,\overrightarrow{f} = \overrightarrow{E}$, the linear map $\overrightarrow{f}$ is the null map, and since we are requiring that the restriction of $\widetilde{f}$ to $\mathbf{P}(\overrightarrow{E})$ be equal to $\mathbf{P}(\overrightarrow{f})$, the linear map g must also be the null map on $\overrightarrow{E}$. Thus, $\widetilde{f}$ is unique, and the restriction of $\widetilde{f}$ to $\mathbf{P}(\overrightarrow{E})$ is the partial map undefined everywhere.

If $\overrightarrow{E} - \mathrm{Ker}\,\overrightarrow{f} \neq \emptyset$, by taking a basis of $\mathrm{Im}\,\overrightarrow{f}$ and some inverse image of this basis, we obtain a basis B of a subspace $\overrightarrow{G}$ of $\overrightarrow{E}$ such that $\overrightarrow{E} = \mathrm{Ker}\,\overrightarrow{f} \oplus \overrightarrow{G}$. Since $\overrightarrow{E} = \mathrm{Ker}\,\overrightarrow{f} \oplus \overrightarrow{G}$ where $\dim(\overrightarrow{G}) \geq 1$, for any $x \in \mathrm{Ker}\,\overrightarrow{f}$ and any nonnull vector $y \in \overrightarrow{G}$, we have

$$\begin{aligned}
\lambda(x)w &= \mu w + g(x), \\
\lambda(y)w + \lambda(y)\overrightarrow{f}(y) &= \mu w + g(y),
\end{aligned}$$

and

$$\lambda(x+y)w + \lambda(x+y)\overrightarrow{f}(x+y) = \mu w + g(x+y),$$

which by linearity yields

$$(\lambda(x+y) - \lambda(x) - \lambda(y) + \mu)w + (\lambda(x+y) - \lambda(y))\overrightarrow{f}(y) = 0.$$

Since $F = Kw \oplus H$ and $\overrightarrow{f} \colon \overrightarrow{E} \to H$, we must have $\lambda(x+y) = \lambda(y)$ and $\lambda(x) = \mu$. Then the equation

$$\lambda(x)w = \mu w + g(x)$$

yields $\mu w = \mu w + g(x)$, shows that g vanishes on $\mathrm{Ker}\,\overrightarrow{f}$.

If $\dim(\overrightarrow{G}) = 1$ then by $(*_1)$, for any $y \in \overrightarrow{G}$ we have

$$\lambda(y)w + \lambda(y)\overrightarrow{f}(y) = \mu w + g(y),$$

and for any $\nu \neq 0$ we have

$$\lambda(\nu y)w + \lambda(\nu y)\overrightarrow{f}(\nu y) = \mu w + g(\nu y),$$

which by linearity yields

$$(\lambda(\nu y) - \nu\lambda(y) - \mu + \nu\mu)w + (\nu\lambda(\nu y) - \nu\lambda(y))\overrightarrow{f}(y) = 0.$$

Since $F = Kw \oplus H$, $\overrightarrow{f} \colon \overrightarrow{E} \to H$, and $\nu \neq 0$, we must have $\lambda(\nu y) = \lambda(y)$. Then we must also have $(\lambda(y) - \mu)(1 - \nu) = 0$.

If $K = \{0, 1\}$, since the only nonzero scalar is 1, it is immediate that $g(y) = \overrightarrow{f}(y)$, and we are done. Otherwise, for $\nu \neq 0, 1$, we get $\lambda(y) = \mu$ for all $y \in \overrightarrow{G}$. Then equation

$$\lambda(y)w + \lambda(y)\overrightarrow{f}(y) = \mu w + g(y)$$

yields $g = \mu\overrightarrow{f}$ on G, and since g vanishes on $\mathrm{Ker}\,\overrightarrow{f}$ we get $g = \mu\overrightarrow{f}$ on $\overrightarrow{E}$ and the restriction of $\widetilde{f} = \mathbf{P}(g)$ to $\mathbf{P}(\overrightarrow{E})$ is equal to $\mathbf{P}(\overrightarrow{f})$. But now, by Proposition 25.6 and since $\widehat{F_H}$ is isomorphic to F, the linear map $\widehat{f}$ is completely determined by

$$\widehat{f}(u \mathbin{\widehat{+}} \lambda a) = \lambda f(a) + \overrightarrow{f}(u) = \lambda w + \overrightarrow{f}(u),$$

and g is completely determined by

$$g(u \mathbin{\widehat{+}} \lambda a) = \lambda g(a) + g(u) = \lambda\mu w + \mu\overrightarrow{f}(u).$$

Thus, we have $g = \mu\widehat{f}$.

Otherwise, if $\dim(\overrightarrow{G}) \geq 2$, then for any two distinct basis vectors u and v in B,

$$\begin{aligned}
\lambda(u)w + \lambda(u)\overrightarrow{f}(u) &= \mu w + g(u),\\
\lambda(v)w + \lambda(v)\overrightarrow{f}(v) &= \mu w + g(v),
\end{aligned}$$

and

$$\lambda(u+v)w + \lambda(u+v)\overrightarrow{f}(u+v) = \mu w + g(u+v),$$

and by linearity, we get

$$(\lambda(u+v) - \lambda(u) - \lambda(v) + \mu)w + (\lambda(u+v) - \lambda(u))\overrightarrow{f}(u) + (\lambda(u+v) - \lambda(v))\overrightarrow{f}(v) = 0.$$

Since $F = Kw \oplus H$, $\overrightarrow{f}\colon \overrightarrow{E} \to H$, and $\overrightarrow{f}(u)$ and $\overrightarrow{f}(v)$ are linearly independent (because $\overrightarrow{f}$ in injective on $\overrightarrow{G}$), we must have

$$\lambda(u+v) = \lambda(u) = \lambda(v) = \mu,$$

which implies that $g = \mu\overrightarrow{f}$ on $\overrightarrow{E}$, and the restriction of $\widetilde{f} = \mathbf{P}(g)$ to $\mathbf{P}(\overrightarrow{E})$ is equal to $\mathbf{P}(\overrightarrow{f})$. As in the previous case, g is completely determined by

$$g(u \mathbin{\widehat{+}} \lambda a) = \lambda g(a) + g(u) = \lambda\mu w + \mu\overrightarrow{f}(u).$$

Again, we have $g = \mu\widehat{f}$, and thus $\widetilde{f}$ is unique. $\square$

The requirement that the restriction of $\widetilde{f} = \mathbf{P}(g)$ to $\mathbf{P}(\overrightarrow{E})$ be equal to $\mathbf{P}(\overrightarrow{f})$ is necessary for the uniqueness of $\widetilde{f}$. The problem comes up when f is a constant map. Indeed, if f is the constant map defined such that $f(a) = [w]$ for some fixed vector $w \in F$, it can be shown that any linear map $g\colon \widehat{E} \to F$ defined such that $g(a) = \mu w$ and $g(u) = \varphi(u)w$ for all $u \in \overrightarrow{E}$, for some $\mu \neq 0$, and some linear form $\varphi\colon \overrightarrow{E} \to F$ satisfies $f = \mathbf{P}(g) \circ i$.

Proposition 26.17 shows that $\langle \widetilde{E}, \mathbf{P}(\overrightarrow{E}), i \rangle$ is the projective completion of the affine space E.

The projective completion $\widetilde{E}$ of an affine space E is a very handy place in which to do geometry in, mainly because the following facts can be easily established.

There is a bijection between affine subspaces of E and projective subspaces of $\widetilde{E}$ not contained in $\mathbf{P}(\overrightarrow{E})$. Two affine subspaces of E are parallel iff the corresponding projective subspaces of $\widetilde{E}$ have the same intersection with the hyperplane at infinity $\mathbf{P}(\overrightarrow{E})$. There is also a bijection between affine maps from E to F and projective maps from $\widetilde{E}$ to $\widetilde{F}$ mapping the hyperplane at infinity $\mathbf{P}(\overrightarrow{E})$ into the hyperplane at infinity $\mathbf{P}(\overrightarrow{F})$. In the projective plane, two distinct lines intersect in a single point (possibly at infinity, when the lines are parallel). In the projective space, two distinct planes intersect in a single line (possibly at infinity, when the planes are parallel). In the projective space, a plane and a line not contained in that plane intersect in a single point (possibly at infinity, when the plane and the line are parallel).

26.9 Making Good Use of Hyperplanes at Infinity

Given a vector space E and a hyperplane H in E, we have already observed that the projective spaces $\widetilde{E_H}$ and $\mathbf{P}(E)$ are isomorphic. Thus, $\mathbf{P}(H)$ can be viewed as the hyperplane at infinity in $\mathbf{P}(E)$, and the considerations applying to the projective completion of an affine space apply to the affine patch E_H on $\mathbf{P}(E)$. This fact yields a powerful and elegant method for proving theorems in projective geometry. The general schema is to choose some projective hyperplane $\mathbf{P}(H)$ in $\mathbf{P}(E)$, view it as the "hyperplane at infinity," then prove an affine version of the desired result in the affine patch E_H (the complement of $\mathbf{P}(H)$ in $\mathbf{P}(E)$, which has an affine structure), and then transfer this result back to the projective space $\mathbf{P}(E)$. This technique is often called "sending objects to infinity." We refer the reader to geometry textbooks for a comprehensive development of these ideas (for example, Berger [11, 12], Samuel [142], Sidler [161], Tisseron [175], or Pedoe [136]), but we cannot resist presenting the projective versions of the theorems of Pappus and Desargues. Indeed, the method of sending points to infinity provides some strikingly elegant proofs. We begin with Pappus's theorem, illustrated in Figure 26.20.

Proposition 26.18. *(Pappus) Given any projective plane* $\mathbf{P}(E)$ *and any two distinct lines* D *and* D'*, for any distinct points* a, b, c, a', b', c'*, with* a, b, c *on* D *and* a', b', c' *on* D'*, if*

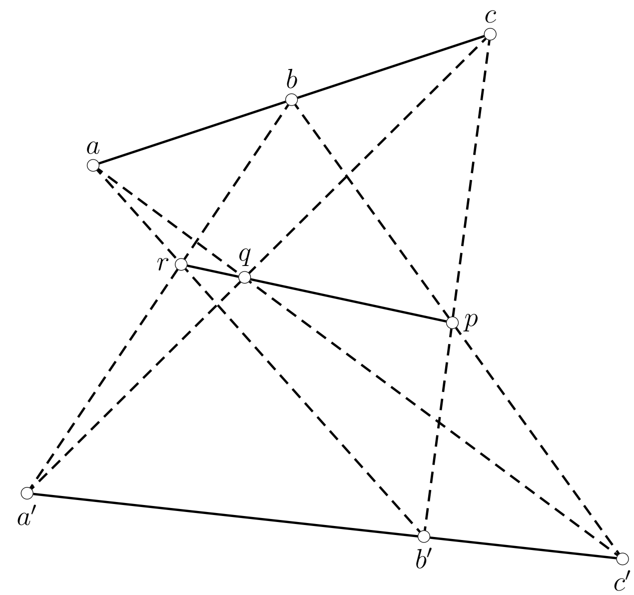


Figure 26.20: Pappus's theorem (projective version).

a, b, c, a', b', c' are distinct from the intersection of D and D', then the intersection points $p = \langle b, c' \rangle \cap \langle b', c \rangle$, $q = \langle a, c' \rangle \cap \langle a', c \rangle$, and $r = \langle a, b' \rangle \cap \langle a', b \rangle$ are collinear.

Proof. First, since any two lines in a projective plane intersect in a single point, the points p, q, r are well defined. Choose $\Delta = \langle p, r \rangle$ as the line at infinity, and consider the affine plane $X = \mathbf{P}(E) - \Delta$. Since $\langle a, b' \rangle$ and $\langle a', b \rangle$ intersect at a point at infinity r on Δ, $\langle a, b' \rangle$ and $\langle a', b \rangle$ are parallel, and similarly $\langle b, c' \rangle$ and $\langle b', c \rangle$ are parallel. Thus, by the affine version of Pappus's theorem (Proposition 24.12), the lines $\langle a, c' \rangle$ and $\langle a', c \rangle$ are parallel, which means that their intersection q is on the line at infinity $\Delta = \langle p, r \rangle$, which means that p, q, r are collinear. $\square$

By working in the projective completion of an affine plane, we can obtain an improved version of Pappus's theorem for affine planes. The reader will have to figure out how to deal with the special cases where some of p, q, r go to infinity.

Now, we prove a projective version of Desargues's theorem slightly more general than that given in Proposition 26.7. It is interesting that the proof is radically different, depending on the dimension of the projective space $\mathbf{P}(E)$. This is not surprising. In axiomatic presentations of projective plane geometry, Desargues's theorem is independent of the other axioms. Desargues's theorem is illustrated in Figure 26.21.

Proposition 26.19. *(Desargues) Let $\mathbf{P}(E)$ be a projective space. Given two triangles (a, b, c) and (a', b', c'), where the points a, b, c, a', b', c' are pairwise distinct and the lines $A = \langle b, c \rangle$, $B = \langle a, c \rangle$, $C = \langle a, b \rangle$, $A' = \langle b', c' \rangle$, $B' = \langle a', c' \rangle$, $C' = \langle a', b' \rangle$ are pairwise distinct, if the* *lines $\langle a, a'\rangle$, $\langle b, b'\rangle$, and $\langle c, c'\rangle$ intersect in a common point d distinct from a, b, c, a', b', c', then the intersection points $p = \langle b, c\rangle \cap \langle b', c'\rangle$, $q = \langle a, c\rangle \cap \langle a', c'\rangle$, and $r = \langle a, b\rangle \cap \langle a', b'\rangle$ belong to a common line distinct from A, B, C, A', B', C'.*

Proof. First, it is immediately shown that the line $\langle p, q\rangle$ is distinct from the lines A, B, C, A', B', C'. Let us assume that $\mathbf{P}(E)$ has dimension $n \geq 3$. If the seven points d, a, b, c, a', b', c' generate a projective subspace of dimension 3, then by Proposition 26.1, the intersection of the two planes $\langle a, b, c\rangle$ and $\langle a', b', c'\rangle$ is a line, and thus p, q, r are collinear.

If $\mathbf{P}(E)$ has dimension $n = 2$ or the seven points d, a, b, c, a', b', c' generate a projective subspace of dimension 2, we use the following argument. In the projective plane X generated by the seven points d, a, b, c, a', b', c', choose the projective line $\Delta = \langle p, r\rangle$ as the line at infinity. Then in the affine plane $Y = X - \Delta$, the lines $\langle b, c\rangle$ and $\langle b', c'\rangle$ are parallel, and the lines $\langle a, b\rangle$ and $\langle a', b'\rangle$ are parallel, and the lines $\langle a, a'\rangle$, $\langle b, b'\rangle$, and $\langle c, c'\rangle$ are either parallel or concurrent. Then by the converse of the affine version of Desargues's theorem (Proposition 24.13), the lines $\langle a, c\rangle$ and $\langle a', c'\rangle$ are parallel, which means that their intersection q belongs to the line at infinity $\Delta = \langle p, r\rangle$, and thus that p, q, r are collinear. $\square$

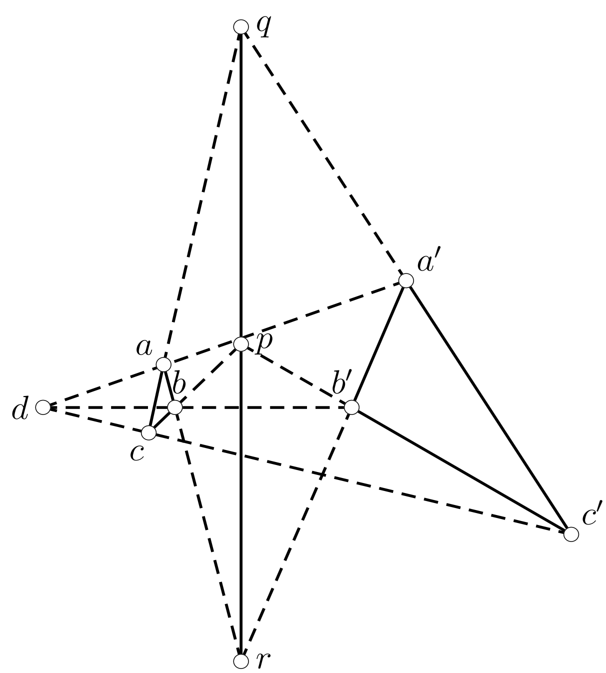


Figure 26.21: Desargues's theorem (projective version).

The converse of Desargues's theorem also holds. Using the projective completion of an affine space, it is easy to state an improved affine version of Desargues's theorem. The reader will have to figure out how to deal with the case where some of the points p, q, r go to infinity. It can also be shown that Pappus's theorem implies Desargues's theorem. Many results of projective or affine geometry can be obtained using the method of "sending points to infinity."

We now discuss briefly the notion of cross-ratio, since it is a major concept of projective geometry.

26.10 The Cross-Ratio

Recall that affine maps preserve the ratio of three collinear points. In general, projective maps do not preserve the ratio of three collinear points. However, bijective projective maps preserve the "ratio of ratios" of any four collinear points (three of which are distinct). Such ratios are called *cross-ratios* (in French, "birapport"). There are several ways of introducing cross-ratios, but since we already have Proposition 26.5 at our disposal, we can circumvent some of the tedious calculations needed if other approaches are chosen.

Given a field K, say $K = \mathbb{R}$, recall that the projective line $\mathbb{P}^1_K$ consists of all equivalence classes $[x, y]$ of pairs $(x, y) \in K^2$ such that $(x, y) \neq (0, 0)$, under the equivalence relation $\sim$ defined such that

$$(x_1, y_1) \sim (x_2, y_2) \quad \text{iff} \quad x_2 = \lambda x_1 \quad \text{and} \quad y_2 = \lambda y_1,$$

for some $\lambda \in K - \{0\}$. Letting $\infty = [1, 0]$, the projective line $\mathbb{P}^1_K$ is in bijection with $K \cup \{\infty\}$. Furthermore, letting $0 = [0, 1]$ and $1 = [1, 1]$, the triple $(\infty, 0, 1)$ forms a projective frame for $\mathbb{P}^1_K$. Using this projective frame and Proposition 26.5, we define the cross-ratio of four collinear points as follows.

Definition 26.8. Given a projective line $\Delta = \mathbf{P}(D)$ over a field K, for any sequence (a, b, c, d) of four points in Δ, where a, b, c are distinct (i.e., (a, b, c) is a projective frame), the *cross-ratio* $[a, b, c, d]$ is defined as the element $h(d) \in \mathbb{P}^1_K$, where $h\colon \Delta \to \mathbb{P}^1_K$ is the unique projectivity such that $h(a) = \infty$, $h(b) = 0$, and $h(c) = 1$ (which exists by Proposition 26.5, since (a, b, c) is a projective frame for Δ and $(\infty, 0, 1)$ is a projective frame for $\mathbb{P}^1_K$). For any projective space $\mathbf{P}(E)$ (of dimension ≥ 2) over a field K and any sequence (a, b, c, d) of four collinear points in $\mathbf{P}(E)$, where a, b, c are distinct, the cross-ratio $[a, b, c, d]$ is defined using the projective line Δ that the points a, b, c, d define. For any affine space E and any sequence (a, b, c, d) of four collinear points in E, where a, b, c are distinct, the cross-ratio $[a, b, c, d]$ is defined by considering E as embedded in $\widetilde{E}$.

It should be noted that the definition of the cross-ratio $[a, b, c, d]$ depends on the order of the points. Thus, there could be $24 = 4!$ different possible values depending on the permutation of $\{a, b, c, d\}$. In fact, there are at most 6 distinct values. Also, note that $[a, b, c, d] = \infty$ iff $d = a$, $[a, b, c, d] = 0$ iff $d = b$, and $[a, b, c, d] = 1$ iff $d = c$. Thus, $[a, b, c, d] \in K - \{0, 1\}$ iff $d \notin \{a, b, c\}$.

The following proposition is almost obvious, but very important. It shows that projectivities between projective lines are characterized by the preservation of the cross-ratio of any four points (three of which are distinct).

Proposition 26.20. *Given any two projective lines Δ and Δ', for any sequence (a, b, c, d) of points in Δ and any sequence (a', b', c', d') of points in Δ', if a, b, c are distinct and a', b', c' are distinct, there is a unique projectivity $f \colon \Delta \to \Delta'$ such that $f(a) = a'$, $f(b) = b'$, $f(c) = c'$, and $f(d) = d'$ iff $[a, b, c, d] = [a', b', c', d']$.*

Proof. First, assume that $f \colon \Delta \to \Delta'$ is a projectivity such that $f(a) = a'$, $f(b) = b'$, $f(c) = c'$, and $f(d) = d'$. Let $h \colon \Delta \to \mathbb{P}^1_K$ be the unique projectivity such that $h(a) = \infty$, $h(b) = 0$, and $h(c) = 1$, and let $h' \colon \Delta' \to \mathbb{P}^1_K$ be the unique projectivity such that $h'(a') = \infty$, $h'(b') = 0$, and $h'(c') = 1$. By definition, $[a, b, c, d] = h(d)$ and $[a', b', c', d'] = h'(d')$. However, $h' \circ f \colon \Delta \to \mathbb{P}^1_K$ is a projectivity such that $(h' \circ f)(a) = \infty$, $(h' \circ f)(b) = 0$, and $(h' \circ f)(c) = 1$, and by the uniqueness of h, we get $h = h' \circ f$. But then, $[a, b, c, d] = h(d) = h'(f(d)) = h'(d') = [a', b', c', d']$.

Conversely, assume that $[a, b, c, d] = [a', b', c', d']$. Since (a, b, c) and (a', b', c') are projective frames, by Proposition 26.5, there is a unique projectivity $g \colon \Delta \to \Delta'$ such that $g(a) = a'$, $g(b) = b'$, and $g(c) = c'$. Now, $h' \circ g \colon \Delta \to \mathbb{P}^1_K$ is a projectivity such that $(h' \circ g)(a) = \infty$, $(h' \circ g)(b) = 0$, and $(h' \circ g)(c) = 1$, and thus, $h = h' \circ g$. However, $h'(d') = [a', b', c', d'] = [a, b, c, d] = h(d) = h'(g(d))$, and since h' is injective, we get $d' = g(d)$. $\square$

As a corollary of Proposition 26.20, given any three distinct points a, b, c on a projective line Δ, for every $\lambda \in \mathbb{P}^1_K$ there is a unique point $d \in \Delta$ such that $[a, b, c, d] = \lambda$.

In order to compute explicitly the cross-ratio, we show the following easy proposition.

Proposition 26.21. *Given any projective line $\Delta = \mathbf{P}(D)$, for any three distinct points a, b, c in Δ, if $a = p(u)$, $b = p(v)$, and $c = p(u + v)$, where (u, v) is a basis of D, and for any $[\lambda, \mu]_\sim \in \mathbb{P}^1_K$ and any point $d \in \Delta$, we have*

$$d = p(\lambda u + \mu v) \quad \textit{iff} \quad [a, b, c, d] = [\lambda, \mu]_\sim.$$

Proof. If (e_1, e_2) is the basis of K^2 such that $e_1 = (1, 0)$ and $e_2 = (0, 1)$, it is obvious that $p(e_1) = \infty$, $p(e_2) = 0$, and $p(e_1 + e_2) = 1$. Let $f \colon D \to K^2$ be the bijective linear map such that $f(u) = e_1$ and $f(v) = e_2$. Then $f(u + v) = e_1 + e_2$, and thus f induces the unique projectivity $\mathbf{P}(f) \colon \mathbf{P}(D) \to \mathbb{P}^1_K$ such that $\mathbf{P}(f)(a) = \infty$, $\mathbf{P}(f)(b) = 0$, and $\mathbf{P}(f)(c) = 1$. Then

$$\mathbf{P}(f)(p(\lambda u + \mu v)) = [f(\lambda u + \mu v)]_\sim = [\lambda e_1 + \mu e_2]_\sim = [\lambda, \mu]_\sim,$$

that is,

$$d = p(\lambda u + \mu v) \quad \text{iff} \quad [a, b, c, d] = [\lambda, \mu]_\sim,$$

as claimed. $\square$

We can now compute the cross-ratio explicitly for any given basis (u, v) of D. Assume that a, b, c, d have homogeneous coordinates $[\lambda_1, \mu_1]$, $[\lambda_2, \mu_2]$, $[\lambda_3, \mu_3]$, and $[\lambda_4, \mu_4]$ over the projective frame induced by (u, v). Letting $w_i = \lambda_i u + \mu_i v$, we have $a = p(w_1)$, $b = p(w_2)$, $c = p(w_3)$, and $d = p(w_4)$. Since a and b are distinct, w_1 and w_2 are linearly independent, and we can write $w_3 = \alpha w_1 + \beta w_2$ and $w_4 = \gamma w_1 + \delta w_2$, which can also be written as

$$w_4 = \frac{\gamma}{\alpha}\,\alpha\, w_1 + \frac{\delta}{\beta}\,\beta\, w_2,$$

and by Proposition 26.21, $[a, b, c, d] = \big[\gamma/\alpha, \delta/\beta\big]$. However, since w_1 and w_2 are linearly independent, it is possible to solve for $\alpha, \beta, \gamma, \delta$ in terms of the homogeneous coordinates, obtaining expressions involving determinants:

$$\begin{aligned}
\alpha &= \frac{\det(w_3, w_2)}{\det(w_1, w_2)}, \qquad & \beta = \frac{\det(w_1, w_3)}{\det(w_1, w_2)},\\
\gamma &= \frac{\det(w_4, w_2)}{\det(w_1, w_2)}, \qquad & \delta = \frac{\det(w_1, w_4)}{\det(w_1, w_2)},
\end{aligned}$$

and thus, assuming that $d \neq a$, we get

$$[a, b, c, d] = \frac{\begin{vmatrix} \lambda_3 & \lambda_1 \\ \mu_3 & \mu_1 \end{vmatrix}}{\begin{vmatrix} \lambda_3 & \lambda_2 \\ \mu_3 & \mu_2 \end{vmatrix}} \Bigg/ \frac{\begin{vmatrix} \lambda_4 & \lambda_1 \\ \mu_4 & \mu_1 \end{vmatrix}}{\begin{vmatrix} \lambda_4 & \lambda_2 \\ \mu_4 & \mu_2 \end{vmatrix}}.$$

When $d = a$, we have $[a, b, c, d] = \infty$. In particular, if Δ is the projective completion of an affine line D, then $\mu_i = 1$, and we get

$$[a, b, c, d] = \frac{\lambda_3 - \lambda_1}{\lambda_3 - \lambda_2} \Bigg/ \frac{\lambda_4 - \lambda_1}{\lambda_4 - \lambda_2} = \frac{\overrightarrow{ca}}{\overrightarrow{cb}} \Bigg/ \frac{\overrightarrow{da}}{\overrightarrow{db}}.$$

When $d = \infty$, we get

$$[a, b, c, \infty] = \frac{\overrightarrow{ca}}{\overrightarrow{cb}},$$

which is just the usual ratio (although we defined it earlier as $-\mathrm{ratio}(a, c, b)$).

We briefly mention some of the properties of the cross-ratio. For example, the cross-ratio $[a, b, c, d]$ is invariant if any two elements and the complementary two elements are transposed, and letting $0^{-1} = \infty$ and $\infty^{-1} = 0$, we have

$$[a, b, c, d] = [b, a, c, d]^{-1} = [a, b, d, c]^{-1}$$

and

$$[a, b, c, d] = 1 - [a, c, b, d].$$

Since the permutations of $\{a,b,c,d\}$ are generated by the above transpositions, the cross-ratio takes at most six values. Letting $\lambda = [a,b,c,d]$, if $\lambda \in \{\infty, 0, 1\}$, then any permutation of $\{a,b,c,d\}$ yields a cross-ratio in $\{\infty, 0, 1\}$, and if $\lambda \notin \{\infty, 0, 1\}$, then there are at most the six values

$$\lambda, \quad \frac{1}{\lambda}, \quad 1-\lambda, \quad 1-\frac{1}{\lambda}, \quad \frac{1}{1-\lambda}, \quad \frac{\lambda}{\lambda-1}.$$

It can be shown that the function

$$\lambda \mapsto 256\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(1-\lambda)^2}$$

takes a constant value on the six values listed above.

We also define when four points form a harmonic division. For this, we need to assume that K is not of characteristic 2.

Definition 26.9. Given a projective line Δ, we say that a sequence of four collinear points (a,b,c,d) in Δ (where a,b,c are distinct) forms a *harmonic division* if $[a,b,c,d] = -1$. When $[a,b,c,d] = -1$, we also say that c and d are *harmonic conjugates* of a and b.

If a,b,c are distinct collinear points in some affine space, from

$$[a,b,c,\infty] = \frac{\overrightarrow{ca}}{\overrightarrow{cb}},$$

we note that c is the midpoint of (a,b) iff $[a,b,c,\infty] = -1$, that is, if (a,b,c,∞) forms a harmonic division. Figure 26.22 shows a harmonic division (a,b,c,d) on the real line, where the coordinates of (a,b,c,d) are $(-2,2,1,4)$.

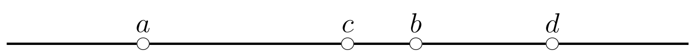


Figure 26.22: Four points forming a harmonic division.

If $\Delta = \mathbb{P}^1_K$ and a,b,c,d are all distinct from ∞, then we see immediately from the formula

$$[a,b,c,d] = \frac{c-a}{c-b}\bigg/\frac{d-a}{d-b}$$

that $[a,b,c,d] = -1$ iff

$$2(ab+cd) = (a+b)(c+d).$$

We also check immediately that $[a,b,c,\infty] = -1$ iff

$$a+b = 2c.$$

There is a nice geometric interpretation of harmonic divisions in terms of quadrangles (or complete quadrilaterals). Consider the quadrangle (projective frame) (a, b, c, d) in a projective plane, and let a' be the intersection of $\langle d, a\rangle$ and $\langle b, c\rangle$, b' be the intersection of $\langle d, b\rangle$ and $\langle a, c\rangle$, and c' be the intersection of $\langle d, c\rangle$ and $\langle a, b\rangle$. If we let g be the intersection of $\langle a, b\rangle$ and $\langle a', b'\rangle$, then it is an interesting exercise to show that (a, b, g, c') is a harmonic division. One way to prove this is to pick (a, c, b, d) as a projective frame and to compute the coordinates of a', b', c', and g. Then because $\langle a, c\rangle$ is the line at infinity, $[a, b, g, c'] = [\infty, b, g, c']$, which is computed using the above formula. Another way is to send some well chosen points to infinity; see Berger [11] (Chapter 6, Section 6.4).

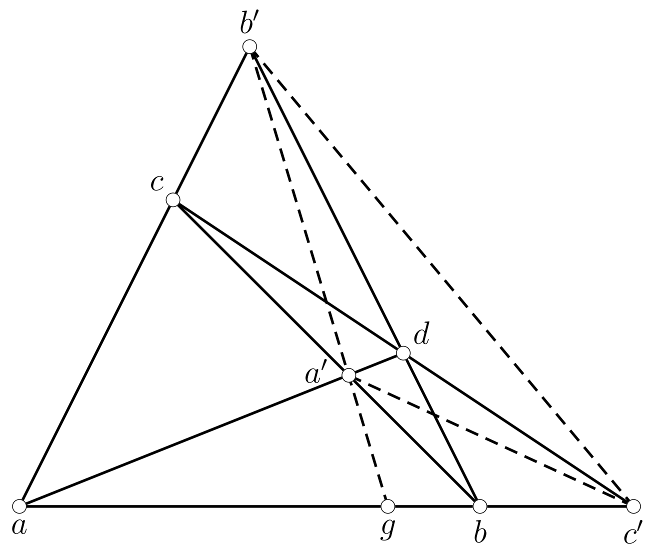


Figure 26.23: A quadrangle, and harmonic divisions.

In fact, it can be shown that the following quadruples of lines induce harmonic divisions: $(\langle c, a\rangle, \langle b', a'\rangle, \langle d, b\rangle, \langle b', c'\rangle)$ on $\langle a, b\rangle$, $(\langle b, a\rangle, \langle c', a'\rangle, \langle d, c\rangle, \langle c', b'\rangle)$ on $\langle a, c\rangle$, and $(\langle b, c\rangle, \langle a', c'\rangle, \langle a, d\rangle, \langle a', b'\rangle)$ on $\langle c, d\rangle$; see Figure 26.23. For more on harmonic divisions, the interested reader should consult any text on projective geometry (for example, Berger [11, 12], Samuel [142], Sidler [161], Tisseron [175], or Pedoe [136]).

26.11 Fixed Points of Homographies and Homologies; Homographies of $\mathbb{RP}^1$ and $\mathbb{RP}^2$

Let $\mathbb{P}(E)$ be a projective space where E is a vector space over some field K, and let $h\colon \mathbb{P}(E) \to \mathbb{P}(E)$ be homography (or projectivity) of $\mathbb{P}(E)$ where h is given by the linear isomorphism $f\colon E \to E$ so that $h = \mathbb{P}(f)$. Observe that if $u \in E$ is an eigenvector of f for some eigenvalue

$\lambda \in K$, then

$$h([u]) = [f(u)] = [\lambda u] = [u]$$

since $\lambda \neq 0$ because f is an isomorphism, which means that the point $[u] \in \mathbf{P}(E)$ is a fixed pointh of h. In other words, *eigenvectors of f induce fixed points of $h = \mathbb{P}(f)$.*

Consequently, it makes sense to try to classify homographies in terms of their fixed points. Of course this depends on the field K. If K is algebraically closed, for instance $K = \mathbb{C}$, then all the eigenvalues of f belong to K, and we can use the Jordan form of a matrix representing f. If $K = \mathbb{R}$, which is of particular interest to us, then we can use the real Jordan form, and we can obtain a compete classification for $E = \mathbb{R}^2$ and $E = \mathbb{R}^3$. We will also see that special kinds of homographies that leave every point of some projective hyperplane $\mathbf{P}(H)$ fixed, called *homologies*, play a special role.

We begin with the classification of the homographies of the real projective line $\mathbb{RP}^1$. Since a homography h of $\mathbb{RP}^1$ is represented by a real invertible 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

and since A either 0, 1, or 2, real eigenvalues, the homography h has 0, 1, or 2 fixed points.

Definition 26.10. A homography of the real projective line $\mathbb{RP}^1$ not equal to the identity is *elliptic* if is has no fixed point, *parabolic* if it has a single fixed point, or *hyperbolic* if it has two fixed points.

(1) *Elliptic homographies.* In this case, $(a+d)^2 - 4(ad - bc) < 0$, so A has two distinct complex conjugate eigenvalues $\alpha \pm i\beta$, and in $\mathbb{C}^2$, they correspond to two complex eigenvectors $w_1 = u + iv$ and $w_2 = u - iv$, with $u, v \in \mathbb{R}^2$. Since

$$f(w_1) = (\alpha - i\beta)w_1$$

we obtain

$$f(u) + if(v) = \alpha u + \beta v + i(-\beta u + \alpha v),$$

which shows that in the basis (u, v), the homography h is represented by the matrix

$$\Gamma = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}.$$

If we let $\theta \in (0, 2\pi)$ be the angle given by

$$\cos\theta = \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}$$
$$\sin\theta = \frac{\beta}{\sqrt{\alpha^2 + \beta^2}}$$

and write

$$\rho = \sqrt{\alpha^2 + \beta^2},$$

then

$$\Gamma = \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

which corresponds to a similarity. Observe that h is an involution, that is, $h^2 = \mathrm{id}$ iff $\theta = \pi/2$.

(2) *Parabolic homographies*. In this case, we must have $(a+d)^2 - 4(ad - bc) = 0$. The matrix A is not diagonalizable and it has a Jordan form of the form

$$\Gamma = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

In the affine line $y = 1$, a parabolic homography behaves like the translation by $1/\lambda$.

(3) *Hyperbolic homographies*. In this case, $(a+d)^2 - 4(ad - bc) > 0$, so A has two distinct nonzero reals eigenvalues λ and μ, and in a basis of eigenvectors it is represented by the diagonal matrix

$$\Gamma = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}.$$

If P and Q are the distinct fixed points of the the homography h, it is not hard to show that for every $M \in \mathbb{RP}^1$ such that $M \neq P, Q$, we have

$$[P, Q, M, h(M)] = k$$

where $k = \lambda/\mu$. For example, see Sidler [161] (Chapter 3, Proposition 3.3.1), and Berger [11] (Lemma 6.6.3). It can also be shown that h is an involution ($h^2 = \mathrm{id}$) with two distinct fixed points P and Q iff $a + d = 0$ iff $k = -1$ in the above equation; see Sidler [161] (Chapter 3, Proposition 3.3.2), and Samuel [142] (Section 2.4).

We now classify the homographies of $\mathbb{RP}^2$. Since the characteristic polynomial of a 3×3 real matrix A has degree 3 and since every real polynomial of degree 3 has at least one real zero, A has some real eigenvalue. Since $\mathbb{C}$ is algebraically closed, every complex polynomial of degree 3 has three zeros (counted with multiplicity), in which case, all three eigenvalues of a 3×3 complex matrix A belong to $\mathbb{C}$. Thus we have the following useful fact.

Proposition 26.22. *Every homography of the real projective plane $\mathbb{RP}^2$ or of the complex projective plane $\mathbb{CP}^2$ has at least one fixed point.*

Here is the classification of the homographies of $\mathbb{RP}^2$ based on the real Jordan form of a 3×3 matrix. Most details are left as exercises. We denote by Γ the 3×3 matrix representing the real Jordan form of the matrix of the linear map representing the homography h.

(I) Three real eigenvalues α, β, γ. The matrix Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix},$$

with $\alpha, \beta, \gamma \in \mathbb{R}$ nonzero and all distinct. As illustrated in Figure 26.24, the homography h has three fixed points P, Q, R, forming a triangle. The sides (lines) of this triangle are invariant under h. The restriction of h to each of these sides is hyperbolic.

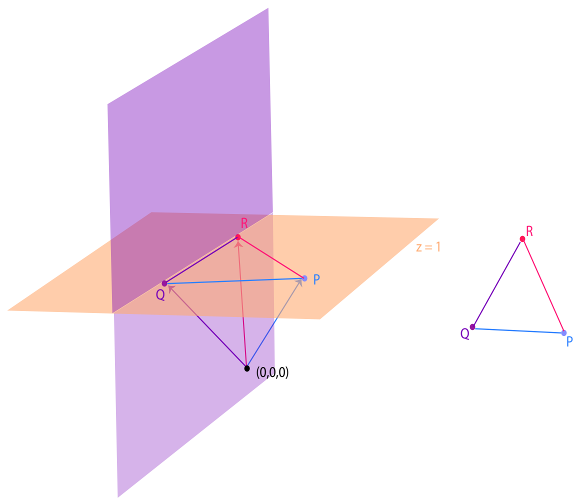


Figure 26.24: Case (I): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed points P, Q, R. The purple (linear) hyperplane maps to itself in a manner which is not the identity.

(II) One real eigenvalue α and two complex conjugate eigenvalues. Then Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & -\gamma \\ 0 & \gamma & \beta \end{pmatrix},$$

with $\alpha, \gamma \in \mathbb{R}$ nonzero. The homography h, which is illustrated in Figure 26.25, has one fixed point P, and a line Δ invariant under h and not containing P. The restriction of h to Δ is elliptic.

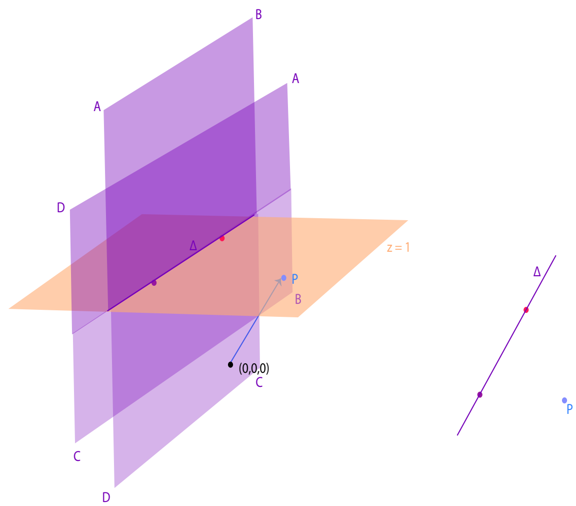


Figure 26.25: Case (II): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed point P and invariant line Δ. The purple (linear) hyperplane maps to itself under a rotation and rescaling.

(III) Two real eigenvalues α, β. The matrix Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \beta \end{pmatrix},$$

with $\alpha, \beta \in \mathbb{R}$ nonzero and distinct. The homography h, as illustrated in Figure 26.26, has one fixed point P, and a line Δ invariant under h and not containing P. The restriction of h to Δ is the identity. Every line through P is invariant under h and the restriction of h to this line is hyperbolic.

(IV) One real eigenvalue α. The matrix Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{pmatrix},$$

with $\alpha \in \mathbb{R}$ nonzero. As illustrated by Figure 26.27, the homography h has one fixed point P, and a line Δ invariant under h containing P. The restriction of h to Δ is the identity. Every line through P is invariant under h and the restriction of h to this line is parabolic.

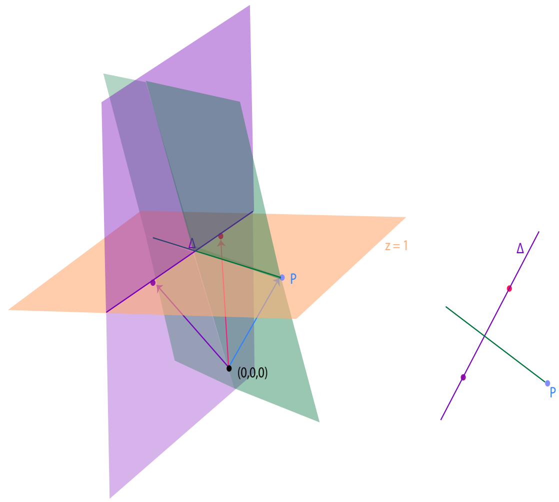


Figure 26.26: Case (III): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed point P and invariant line Δ. The purple (linear) hyperplane maps to itself under rescaling; as such the restriction of the homography to Δ is the identity. The green (linear) hyperplane also is invariant under the homography, but the invariance is not given by the identity map.

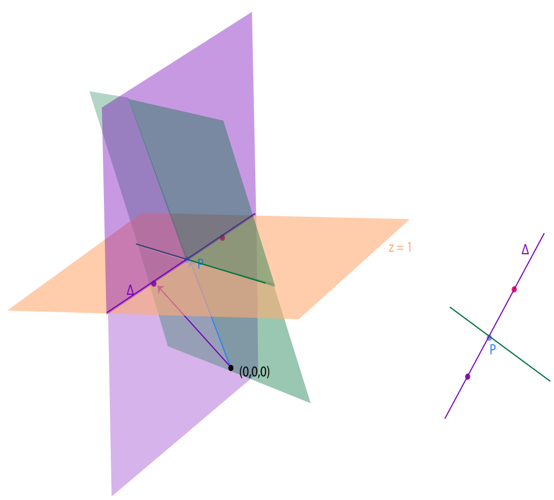


Figure 26.27: Case (IV): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed point P and invariant line Δ containing P. The purple (linear) hyperplane maps to itself under rescaling; as such the restriction of the homography to Δ is the identity. The green (linear) hyperplane also is invariant under the homography, but the invariance is not given by the identity map.

(V) Two real eigenvalues α, β. The matrix Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 1 \\ 0 & 0 & \beta \end{pmatrix},$$

with $\alpha, \beta \in \mathbb{R}$ nonzero and distinct. The homography h, which is illustrated in Figure 26.28, has two fixed points P and Q. The line $\langle P, Q\rangle$ is invariant under h, and there is is another line Δ through Q invariant under h. The restriction of h to Δ is parabolic, and the restriction of h to $\langle P, Q\rangle$ is hyperbolic.

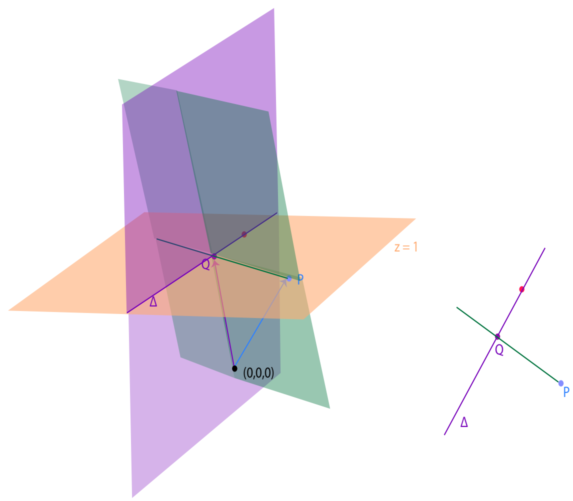


Figure 26.28: Case (V): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed points P, Q and invariant line Δ. Both the purple and green (linear) hyperplanes are invariant under the homography, but the invariance is not given by the identity map.

(VI) One real eigenvalue α. The matrix Γ has the form

$$\Gamma = \begin{pmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{pmatrix},$$

with $\alpha \in \mathbb{R}$ nonzero. The homography h, which is illustrated in Figure 26.29, has one fixed point P, and a line Δ invariant under h containing P. The restriction of h to Δ is parabolic.

For the classification of the homographies of $\mathbb{CP}^2$, Case (II) becomes Case (I).

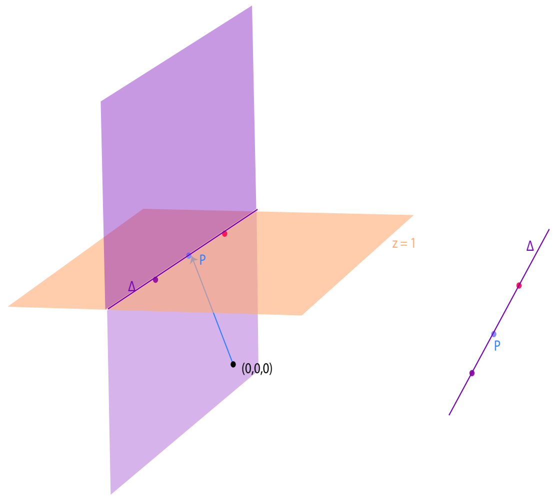


Figure 26.29: Case (VI): The left figure is the hyperplane representation of $\mathbb{RP}^2$ and a homography with fixed point P and invariant line Δ. The purple (linear) hyperplane maps to itself in a manner which is not the identity.

Observe that in Cases (III) and (IV), the homography h has a line Δ of fixed points, as well as a fixed point P. In Case (III), $P \notin \Delta$, and in Case (IV), $P \in \Delta$. This kind of homography is called a *homology*. The point P is called the *center* and the line Δ is called the *axis* (or *base*). Some authors only use the term homology when $P \notin \Delta$, and when $P \in \Delta$, they use the term *elation*. When $P \in \Delta$, other authors use the term *projective transvection*, which we prefer. The center is usually denoted by O (instead of P).

One of the nice features of homologies (and projective transvections) is that there is a nice geometric construction of the image $h(M)$ of a point M in terms of the center O, the axis Δ, and any pair (A, A') where $A' = h(A)$, $A \neq O$, and $A \notin \Delta$.

This construction is possible because for any point $M \neq O$, the line $\langle M, h(M)\rangle$ passes through O. This can be proved using Desargues' Theorem; for example, see Silder [161] (Chapter 4, Section 4.2). We will prove this property for a generalization of homologies to any projective space $\mathbb{P}(E)$, where E is a vector space of any finite dimension.

For the construction, first assume that $M \neq O$ is not on the line $\langle A, A'\rangle$. In this case, the line $\langle A, M\rangle$ intersects Δ in some point I. Since $I \in \Delta$, it is fixed by h, so the image of the line $\langle A, I\rangle$ is the line $\langle A', I\rangle$, and since M is on the line $\langle A, I\rangle$, its image $M' = h(M)$ is on the line $\langle A', I\rangle$. But $M' = h(M)$ is also on the line $\langle O, M\rangle$, which implies that $M' = h(M)$ is the intersection point of the lines $\langle A', I\rangle$ and $\langle O, M\rangle$; see Figure 26.30.

If $M \neq O$ is on the line $\langle A, A'\rangle$, then we use the construction of the image B' of some point $B \neq O$ and not on $\langle A, A'\rangle$ as before, and then repeat the construction by finding the intersection J of $\langle M, B\rangle$ and Δ, and then $M' = h(M)$ is the intersection point of $\langle B', J\rangle$ and $\langle A, A'\rangle$; see Figure 26.31.

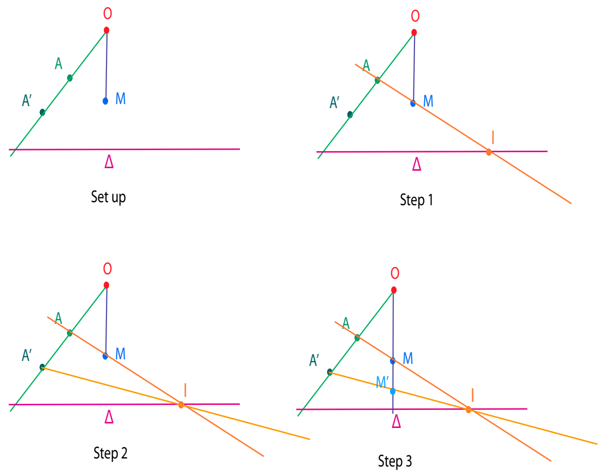


Figure 26.30: The three step process for determining the homology point $h(M) = M'$ when M is not on the line $\langle A, A' \rangle$. Step 1 finds the intersection between the extension of $\langle A, M \rangle$ and Δ. Step 2 forms the line $\langle A', I \rangle$. Step 3 extends $\langle O, M' \rangle$ and determines its intersection with $\langle A', I \rangle$. The intersection point is M'.

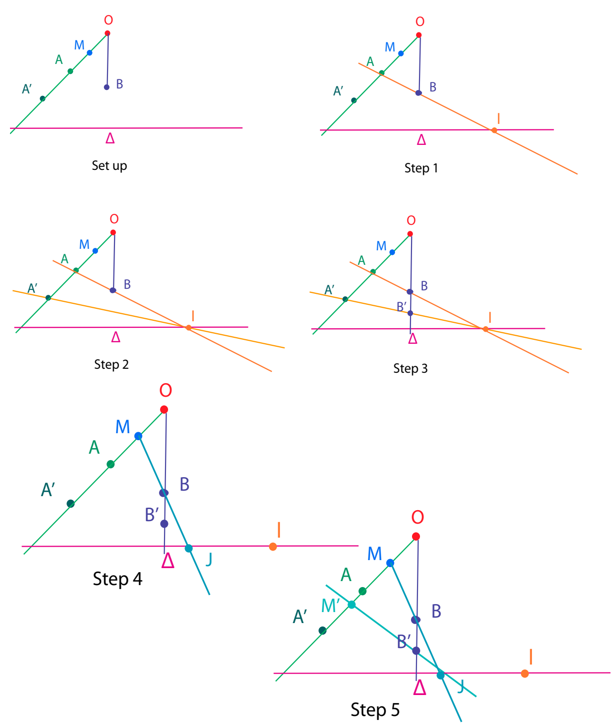


Figure 26.31: The five step process for determining the homology point $h(M) = M'$ when M is on the line $\langle A, A' \rangle$. Steps 1 through 3 determine the line $\langle B, B' \rangle$. Step 4 finds the intersection between $\langle M, B \rangle$ and Δ, namely J. Step 5 forms the line $\langle J, B' \rangle$ and intersects it with $\langle A, A' \rangle$. The intersection point is M'.

The above construction also works if $O \in \Delta$; see Figures 26.32 and 26.33.

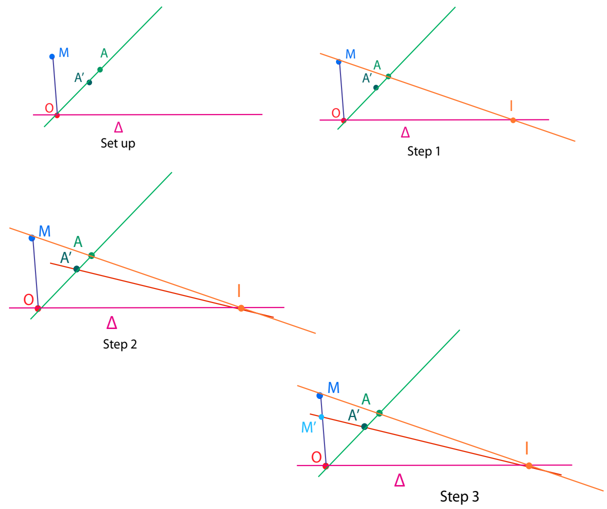


Figure 26.32: The three step process for determining the elation point $h(M) = M'$ when M is not on the line $\langle A, A' \rangle$. Step 1 finds the intersection between the extension of $\langle A, M \rangle$ and Δ. Step 2 forms the line $\langle A', I \rangle$. Step 3 extends $\langle A'I \rangle$ and determines its intersection with $\langle O, M \rangle$. The intersection point is M'.

Another useful property of homologies (here, $O \notin \Delta$) is that for any line d passing through the center O, if I is the intersection point of the line d and Δ, then for any $M \in d$ distinct from O and not on Δ and its image M', the cross-ratio $[O, I, M, M']$ is independent of d. If $[O, I, M, M'] = -1$ for all $M \neq O$, we say that h is a *harmonic homology*. It can be shown that a homography h is a harmonic homology iff h is an involution ($h^2 = \text{id}$); see Silder [161] (Chapter 4, Section 4.4). It can also be shown that any homography of $\mathbb{RP}^2$ can be expressed as the composition of two homologies; see Silder [161] (Chapter 4, Section 4.5).

We now consider the generalization of the notion of homology (and projective transvection) to any projective space $\mathbb{P}(E)$, where E is a vector space of any finite dimension over a field K. We need to review a few concepts from Section 8.15.

Let E be a vector space and let H be a hyperplane in E. Recall from Definition 8.8 that for any nonzero vector $u \in E$ such that $u \notin H$, and any scalar $\alpha \neq 0, 1$, a linear map $f \colon E \to E$ such that $f(x) = x$ for all $x \in H$ and $f(x) = \alpha x$ for every $x \in D = Ku$ is called a *dilatation of hyperplane H, direction D, and scale factor α*. See Figure 26.34.

From Definition 8.9, for any nonzero nonlinear form $\varphi \in E^*$ defining H (which means that $H = \text{Ker}\,(\varphi)$) and any nonzero vector $u \in H$, the linear map $\tau_{\varphi,u}$ given by

$$\tau_{\varphi,u}(x) = x + \varphi(x)u, \quad \varphi(u) = 0,$$

for all $x \in E$ is called a *transvection of hyperplane H and direction u*. See Figure 26.35.

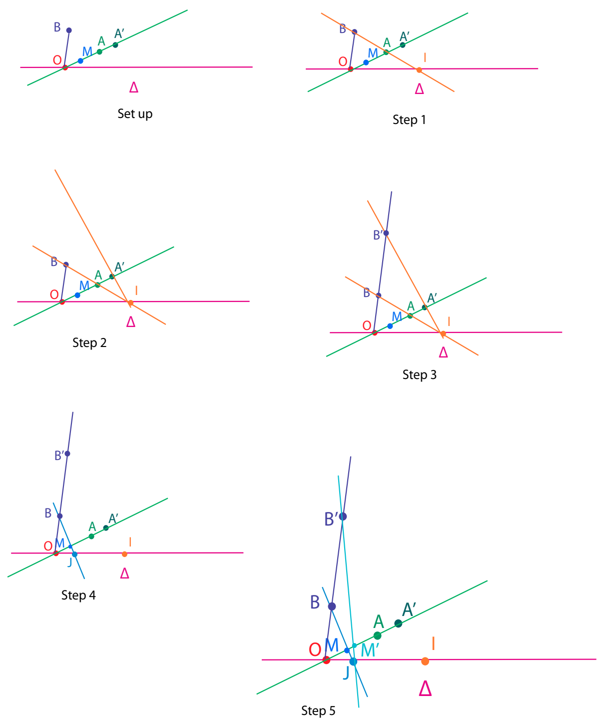


Figure 26.33: The five step process for determining elation point $h(M) = M'$ when M is on the line $\langle A, A' \rangle$. Steps 1 through 3 determine the line $\langle B, B' \rangle$. Step 4 finds the intersection between $\langle M, B \rangle$ and Δ, namely J. Step 5 forms the line $\langle J, B' \rangle$ and intersects it with $\langle A, A' \rangle$. The intersection point is M'.

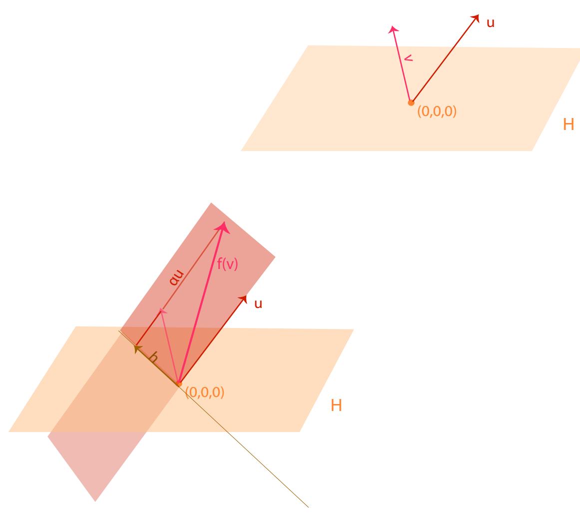


Figure 26.34: A dilation f of the xy-plane in direction $u = (1, 1, 1)$. Every vector v not in the xy-plane determines a rose-colored plane through u, and the image $f(v)$ is an element of this rose hyperplane since it is stretched in the u direction.

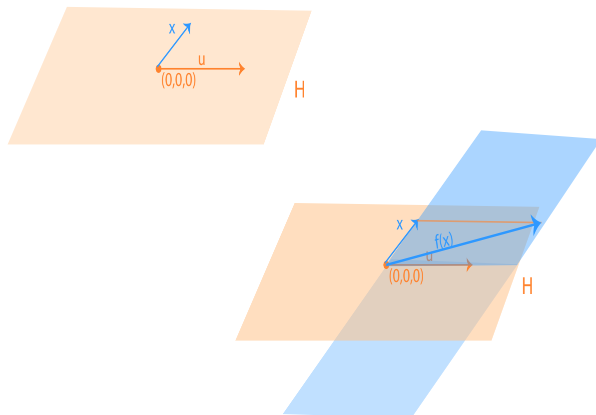


Figure 26.35: A transvection $\tau_{\varphi,u}$ of the xy-plane in direction $u = (0, 1, 0)$, where $\varphi(x, y, z) = z$. Every vector x not in the xy-plane determines a light-blue plane through x and u. The image $f(x)$ stays in the light-blue hyperplane since it is "stretched" in the u direction by a factor of $\varphi(x, y, z)$.

Proposition 26.23, which we repeat here for the convenience of the reader, characterizes the linear isomorphisms $f \neq \mathrm{id}$ that leave every point in the hyperplane H fixed.

Proposition 26.23. *Let $f : E \to E$ be a bijective linear map of a finite-dimensional vector space E and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If $\det(f) = 1$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H. In either case, the vector u is uniquely defined up to a nonzero scalar.*

Proof. Only the last part was not proved in Proposition 8.23, Since f is bijective and the identity on H, the linear map $f - \mathrm{id}$ has kernel exactly H. Since H is a hyperplane in E, the image of $f - \mathrm{id}$ has dimension 1, and since u belong to this image, it is uniquely defined up to a nonzero scalar. $\square$

The proof of Proposition 8.23 shows that if $\dim(E) = n + 1$ and if f is a dilatation of hyperplane H, direction $D = Ku$, and scale α, then 1 is an eigenvalue of f with multiplicity n and $\alpha \neq 0, 1$ is an eigenvalue of f with multiplicity 1; the vector u is an eigenvector for α, and f is diagonalizable. If f is a transvection of hyperplane H and direction u, then 1 is the only eigenvalue of f, and it has multiplicity n; the vector u is an eigenvector for 1, and f is not diagonalizable.

A homology is the projective version of the type of maps involved in Proposition 26.23.

Definition 26.11. For any vector space E and any hyperplane H in E, a homography $h : \mathbb{P}(E) \to \mathbb{P}(E)$ is a *homology of axis (or base)* $\mathbb{P}(H)$ if $h(P) = P$ for all $P \in \mathbb{P}(H)$. In other words, the restriction of h to $\mathbb{P}(H)$ is the identity. More explicitly, if $h = \mathbb{P}(f)$ for some linear isomorphism $f : E \to E$, we have $\mathbb{P}(f)(P) = P$ for all points $P = [u] \in \mathbb{P}(H)$.

Using Proposition 26.23 we obtain the following characterization of homologies. Write $\dim(E) = n + 1$.

Proposition 26.24. *If $h\colon \mathbb{P}(E) \to \mathbb{P}(E)$ is a homology of axis $\mathbb{P}(H)$ and if $h \neq \mathrm{id}$, then for any linear isomorphism $f\colon E \to E$ such that $h = \mathbb{P}(f)$, the following properties hold:*

(1) Either f is a dilatation of hyperplane H and of direction u for some nonzero $u \in E - H$ uniquely defined up to a scalar;

(2) Or f is a transvection of hyperplane H and direction u for some nonzero $u \in H$ uniquely defined up to a scalar.

In both cases, $O = [u] \in \mathbb{P}(E)$ is a fixed point of h, and h has no other fixed points besides O and points in $\mathbb{P}(H)$. In Case (1), $O \notin \mathbb{P}(H)$, whereas in Case (2), $O \in \mathbb{P}(H)$. Furthermore, for any point $M \in \mathbb{P}(E)$, if $M \neq O$ and if $M \notin \mathbb{P}(H)$, then the line $\langle M, h(M)\rangle$ passes through O. If $\dim(E) \geq 3$, the point O is the only point satisfying the above property.

Proof. Since the restriction of $h = \mathbb{P}(f)$ to $\mathbb{P}(H)$ is the identity, and since $\mathbb{P}(f) = \mathbb{P}(\mathrm{id}_H)$, by Proposition 26.4 we have $f = \lambda \mathrm{id}_H$ on H for some nonzero scalar $\lambda \in K$. Then $g = \lambda^{-1} f$ is the identity on H, so by Proposition 26.23 we obtain (1) and (2).

In Case (1), we have $g(u) = \alpha u$, so $\mathbb{P}(g)([u]) = \mathbb{P}(f)([u]) = [u]$. In Case (2), $g(u) = u$, so again $\mathbb{P}(g)([u]) = \mathbb{P}(f)([u]) = [u]$. Therefore, $O = [u]$ is a fixed point of $\mathbb{P}(f)$. In Case (1), the eigenvalues of f are 1 with multiplicity n and α with multiplicity 1. If $Q = [v] \neq O$ was a fixed point of h not in $\mathbb{P}(H)$, then v would be an eigenvector corresponding to a nonzero eigenvalue λ of f with $\lambda \neq 1, \alpha$, and then f would have $n + 2$ eigenvalues (counted with multiplicty), which is absurd. In Case (2), the only eigenvalue of f is 1, with multiplicity n, so f not diagonalizable, and as above, a vector v such that $Q = [v]$ is a fixed point of h not in $\mathbb{P}(H)$ would be an eigenvector corresponding to a nonzero eigenvalue $\lambda \neq 1$ of f, so f would be diagonalizable, a contradiction.

Since in Case (1), for any $x \neq u$ and $x \notin H$ we have $x = \lambda u + h$ for some unique $h \in H$ and some unique $\lambda \neq 0$, so

$$g(x) = g(\lambda u) + g(h) = \lambda\alpha u + h = \lambda u + h + (\lambda\alpha - \lambda)u = x + \lambda(\alpha - 1)u,$$

which shows that $O, [x]$ and $\mathbb{P}(g)([x]) = \mathbb{P}(f)([x])$ are collinear. In Case (2), for any $x \neq u$ and $x \notin H$ we have

$$g(x) = x + \varphi(x)u,$$

which also shows that $O, [x]$ and $\mathbb{P}(g)([x]) = \mathbb{P}(f)([x])$ are collinear. The last property is left as an exercise (see Vienne [185], Chapter 4, Proposition 7). $\square$

Proposition 26.24 suggests the following definition.

Definition 26.12. Let $h\colon \mathbb{P}(E) \to \mathbb{P}(E)$ be a homology of axis $\mathbb{P}(H)$ with $h \neq \mathrm{id}$, where $h = \mathbb{P}(f)$ for some linear isomorphism $f\colon E \to E$. The fixed point $O = [u]$ associated with the vector u involved in the definition of f, which is unique up to a scalar, is called the *center* of h. If $O \in \mathbb{P}(H)$, then h is called a *projective transvection* (or *elation*).

The same geometric construction that we used in the case of the projective plane shows that a homology is determined by its center O, its axis $\mathbb{P}(H)$, and a pair of points A and $A' = h(A)$, with $A \neq O$ and $A \notin \mathbb{P}(H)$. As a kind of converse, we have the following proposition which is easily shown; see Vienne [185] (Chapter IV, Proposition 8).

Proposition 26.25. *Let $\mathbb{P}(H)$ be a hyperplane of $\mathbb{P}(E)$ and let $O \in \mathbb{P}(E)$ be a point. For any pair of distinct points (A, A') such that O, A, A' are collinear and $A, A' \notin \mathbb{P}(H) \cup \{O\}$, there is a unique homology $h\colon \mathbb{P}(E) \to \mathbb{P}(E)$ of centrer O and axis $\mathbb{P}(H)$ such that $h(A) = A'$.*

Remark: From the proof of Proposition 8.23, since every dilatation can be represented by a matrix of the form
$$\begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$
we see that by choosing the hyperplane at infinity to be $x_1 = 0$, on the affine hyperplane $x_1 = 1$, a homology becomes a central magnification by α^{-1}. Similarly, since every transvection can be represented by a matrix of the form
$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$
we see that by choosing the hyperplane at infinity to be $x_1 = 0$, on the affine hyperplane $x_1 = 1$, an elation becomes a translation.

Theorem 8.26 immediately yields the following result showing that the group of homographies $\mathbf{PGL}(E)$ is generated by the homologies.

Theorem 26.26. *Let E be any finite-dimensional vector space over a field K of characteristic not equal to 2. Then, the group of homographies $\mathbf{PGL}(E)$ is generated by the homologies.*

When $E = \mathbb{R}^3$, we saw earlier that the involutions of $\mathbb{RP}^2$ have a nice structure. In particular, if an involution has two fixed points, then it is a harmonic homology.

If $\dim(E) \geq 4$, it is harder to characterize the involutions of $\mathbb{P}(E)$, but it is possible. The case where the linear isomorphism $f\colon E \to E$ defining the involutive homography $h = \mathbb{P}(f)$

has no eigenvalue in the field K is quite different from the case where f has some eigenvalue in K. In the first case, h has no fixed point. It turns out that this implies that $\dim(E)$ is even and there is a simple description of the matrices representing an involution. If h has some fixed point, then f is an involution of E, so it has the eigenvalues $+1$ and -1, and E is the direct sum of the corresponding eigenspaces E_1 and E_{-1}. Then h can be described in terms of $\mathbb{P}(E_1)$ and $\mathbb{P}(E_{-1})$. For details, we refer the reader to Vienne [185] (Chapter IV, Propositions 11 and 12).

26.12 Duality in Projective Geometry

We now consider duality in projective geometry. Given a vector space E of finite dimension $n+1$, recall that its *dual space* E^* is the vector space of all linear forms $f\colon E \to K$ and that E^* is isomorphic to E. We also have a canonical isomorphism between E and its bidual E^{**}, which allows us to identify E and E^{**}.

Let $\mathcal{H}(E)$ denote the set of hyperplanes in $\mathbf{P}(E)$. In Section 26.3 we observed that the map

$$p(f) \mapsto \mathbf{P}(\operatorname{Ker} f)$$

is a bijection between $\mathbf{P}(E^*)$ and $\mathcal{H}(E)$, in which the equivalence class $p(f) = \{\lambda f \mid \lambda \neq 0\}$ of a nonnull linear form $f \in E^*$ is mapped to the hyperplane $\mathbf{P}(\operatorname{Ker} f)$. Using the above bijection between $\mathbf{P}(E^*)$ and $\mathcal{H}(E)$, a projective subspace $\mathbf{P}(U)$ of $\mathbf{P}(E^*)$ (where U is a subspace of E^*) can be identified with a subset of $\mathcal{H}(E)$, namely the family

$$\{\mathbf{P}(H) \mid H = \operatorname{Ker} f,\ f \in U - \{0\}\}$$

consisting of the projective hyperplanes in $\mathcal{H}(E)$ corresponding to nonnull linear forms in U. Such subsets of $\mathcal{H}(E)$ are called *linear systems (of hyperplanes)*.

The bijection between $\mathbf{P}(E^*)$ and $\mathcal{H}(E)$ allows us to view $\mathcal{H}(E)$ as a projective space, and linear systems as projective subspaces of $\mathcal{H}(E)$. In the projective space $\mathcal{H}(E)$, a point is a hyperplane in $\mathbf{P}(E)$! The duality between subspaces of E and subspaces of E^* (reviewed below) and the fact that there is a bijection between $\mathbf{P}(E^*)$ and $\mathcal{H}(E)$ yields a powerful duality between the set of projective subspaces of $\mathbf{P}(E)$ and the set of linear systems in $\mathcal{H}(E)$ (or equivalently, the set of projective subspaces of $\mathbf{P}(E^*)$).

The idea of duality in projective geometry goes back to Gergonne and Poncelet, in the early nineteenth century. However, Poncelet had a more restricted type of duality in mind (polarity with respect to a conic or a quadric), whereas Gergonne had the more general idea of the duality between points and lines (or points and planes). This more general duality arises from a specific pairing between E and E^* (a nonsingular bilinear form). Here we consider the pairing $\langle -, - \rangle\colon E^* \times E \to K$, defined such that

$$\langle f, v \rangle = f(v),$$

for all $f \in E^*$ and all $v \in E$. Recall that given a subset V of E (respectively a subset U of E^*), the *orthogonal* V^0 *of* V is the subspace of E^* defined such that

$$V^0 = \{f \in E^* \mid \langle f, v\rangle = 0, \text{ for every } v \in V\},$$

and that the *orthogonal* U^0 *of* U is the subspace of E defined such that

$$U^0 = \{v \in E \mid \langle f, v\rangle = 0, \text{ for every } f \in U\} = \bigcap_{f\in U} \operatorname{Ker} f.$$

Then, by Theorem 11.4 (since E and E^* have the same finite dimension $n+1$), $U = U^{00}$, $V = V^{00}$, and the maps

$$V \mapsto V^0 \quad \text{and} \quad U \mapsto U^0$$

are inverse bijections, where V is a subspace of E, and U is a subspace of E^*.

These maps set up a *duality* between subspaces of E and subspaces of E^*. Furthermore, we know that U has dimension k iff U^0 has dimension $n+1-k$, and similarly for V and V^0.

Since a linear system $P = \mathbf{P}(U)$ of hyperplanes in $\mathcal{H}(E)$ corresponds to a subspace U of E^*, and since

$$U^0 = \bigcap_{f\in U} \operatorname{Ker} f$$

is the intersection of all the hyperplanes defined by nonnull linear forms in U, we can view a linear system $P = \mathbf{P}(U) = \mathbf{P}(U^{00})$ in $\mathcal{H}(E)$ as the family of hyperplanes in $\mathbf{P}(E)$ containing $\mathbf{P}(U^0)$.

In view of the identification of $\mathbf{P}(E^*)$ with the set $\mathcal{H}(E)$ of hyperplanes in $\mathbf{P}(E)$, by passing to projective spaces, the above bijection between the set of subspaces of E and the set of subspaces of E^* yields a bijection between the set of projective subspaces of $\mathbf{P}(E)$ and the set of linear systems in $\mathcal{H}(E)$ (or equivalently, the set of projective subspaces of $\mathbf{P}(E^*)$) called *duality*. Recall that a point of $\mathcal{H}(E)$ is a hyperplane in $\mathbf{P}(E)$.

More specifically, assuming that E has dimension $n+1$, so that $\mathbf{P}(E)$ has dimension n, if $Q = \mathbf{P}(V)$ is any projective subspace of $\mathbf{P}(E)$ (where V is any subspace of E) and if $P = \mathbf{P}(U)$ is any linear system in $\mathcal{H}(E)$ (where U is any subspace of E^*), we get a subspace Q^0 of $\mathcal{H}(E)$ defined by

$$Q^0 = \{\mathbf{P}(H) \mid Q \subseteq \mathbf{P}(H),\ \mathbf{P}(H) \text{ a hyperplane in } \mathcal{H}(E)\},$$

and a subspace P^0 of $\mathbf{P}(E)$ defined by

$$P^0 = \bigcap\{\mathbf{P}(H) \mid \mathbf{P}(H) \in P,\ \mathbf{P}(H) \text{ a hyperplane in } \mathcal{H}(E)\}.$$

We have $P = P^{00}$ and $Q = Q^{00}$. Since Q^0 is determined by $\mathbf{P}(V^0)$, if $Q = \mathbf{P}(V)$ has dimension k (i.e., if V has dimension $k+1$), then Q^0 has dimension $n-k-1$ (since V has dimension $k+1$ and $\dim(E) = n+1$, then V^0 has dimension $n+1-(k+1) = n-k$). Thus,

$$\dim(Q) + \dim(Q^0) = n - 1,$$

and similarly, $\dim(P) + \dim(P^0) = n - 1$.

A linear system $P = \mathbf{P}(U)$ of hyperplanes in $\mathcal{H}(E)$ is called a *pencil of hyperplanes* if it corresponds to a projective line in $\mathbf{P}(E^*)$, which means that U is a subspace of dimension 2 of E^*. From $\dim(P) + \dim(P^0) = n - 1$, a pencil of hyperplanes P is the family of hyperplanes in $\mathcal{H}(E)$ containing some projective subspace $\mathbf{P}(V)$ of dimension $n - 2$ (where $\mathbf{P}(V)$ is a projective subspace of $\mathbf{P}(E)$, and $\mathbf{P}(E)$ has dimension n). When $n = 2$, a pencil of hyperplanes in $\mathcal{H}(E)$, also called a *pencil of lines*, is the family of lines passing through a given point. When $n = 3$, a pencil of hyperplanes in $\mathcal{H}(E)$, also called a *pencil of planes*, is the family of planes passing through a given line.

When $n = 2$, the above duality takes a rather simple form. In this case (of a projective plane $\mathbf{P}(E)$), the duality is a bijection between points in $\mathbf{P}(E)$ and lines in $\mathbf{P}(E^*)$, represented by pencils of lines in $\mathcal{H}(E)$, with the following properties:

- A point a in $\mathbf{P}(E)$ maps to the line D_a in $\mathbf{P}(E^*)$ represented by the pencil of lines in $\mathcal{H}(E)$ containing a, also denoted by a^*. See Figure 26.36.

- A line D in $\mathbf{P}(E)$ maps to the point p_D in $\mathbf{P}(E^*)$ represented by the line D in $\mathcal{H}(E)$. See Figure 26.37.

- Two points a, b in $\mathbf{P}(E)$ map to lines D_a, D_b in $\mathbf{P}(E^*)$ represented by pencils of lines through a and b, and the intersection of D_a and D_b is the point $p_{\langle a,b\rangle}$ in $\mathbf{P}(E^*)$ corresponding to the line $\langle a, b\rangle$ belonging to both pencils. The point $p_{\langle a,b\rangle}$ is the image of the line $\langle a, b\rangle$ *via* duality. See Figure 26.38

- A line D in $\mathbf{P}(E)$ containing two points a, b maps to the intersection p_D of the lines D_a and D_b in $\mathbf{P}(E^*)$ which are the images of a and b under duality. This is because a, b map to lines D_a, D_b in $\mathbf{P}(E^*)$ represented by pencils of lines through a and b, and the intersection of D_a and D_b is the point p_D in $\mathbf{P}(E^*)$ corresponding to the line $D = \langle a, b\rangle$ belonging to both pencils. The point p_D is the image of the line $D = \langle a, b\rangle$ under duality. Once again, see Figure 26.38.

- If $a \in D$, where a is a point and D is a line in $\mathbf{P}(E)$, then $p_D \in D_a$ in $\mathbf{P}(E^*)$. This is because under duality, a is mapped to the line D_a in $\mathbf{P}(E^*)$ represented by the pencil of lines containing a, and D is mapped to the point $p_D \in \mathbf{P}(E^*)$ represented by the line D through a in this pencil, so $p_D \in D_a$.

The reader will discover that the dual of Desargues's theorem is its converse. This is a nice way of getting the converse for free! We will not spoil the reader's fun and let him discover the dual of Pappus's theorem.

In general, when $n \geq 2$, the above duality is a bijection between points in $\mathbf{P}(E)$ and hyperplanes in $\mathbf{P}(E^*)$, which are represented by linear systems of dimension $n - 1$ in $\mathcal{H}(E)$, with the following properties:

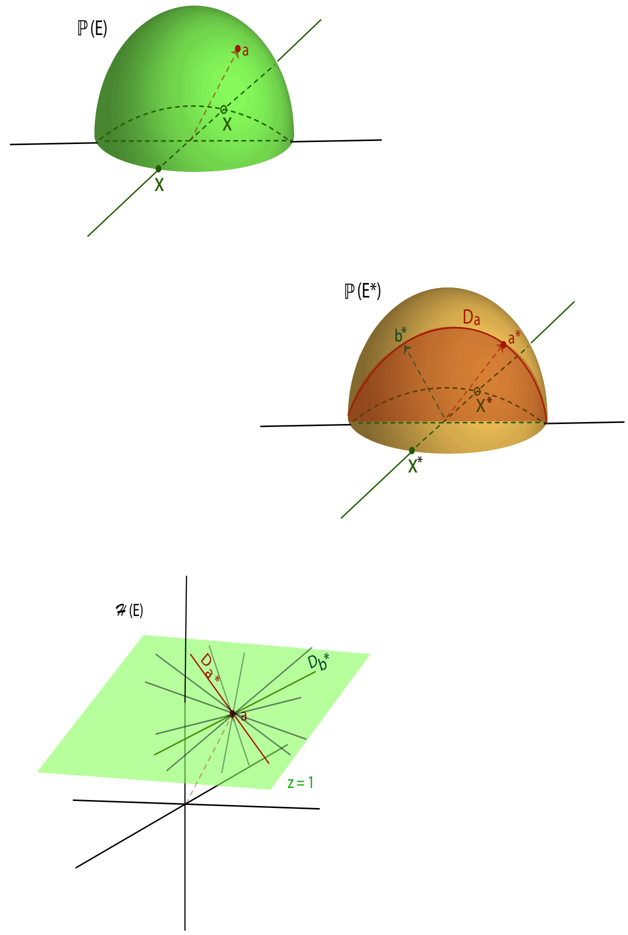


Figure 26.36: The duality between a point in $\mathbf{P}(E)$ and a line in $\mathbf{P}(E^*)$. The line in $\mathbf{P}(E^*)$ is also represented by the pencil of lines through a in $\mathcal{H}(E)$.

- A point a in $\mathbf{P}(E)$ maps to the hyperplane H_a in $\mathbf{P}(E^*)$ (the linear system of hyperplanes in $\mathcal{H}(E)$ containing a, also denoted by a^*).

- A hyperplane H in $\mathbf{P}(E)$ maps to the point p_H in $\mathbf{P}(E^*)$ (represented by the hyperplane H in $\mathcal{H}(E)$).

To conclude our quick tour of projective geometry, we establish a connection between the cross-ratio of hyperplanes in a pencil of hyperplanes with the cross-ratio of the intersection points of any line not contained in any hyperplane in this pencil with four hyperplanes in this pencil.

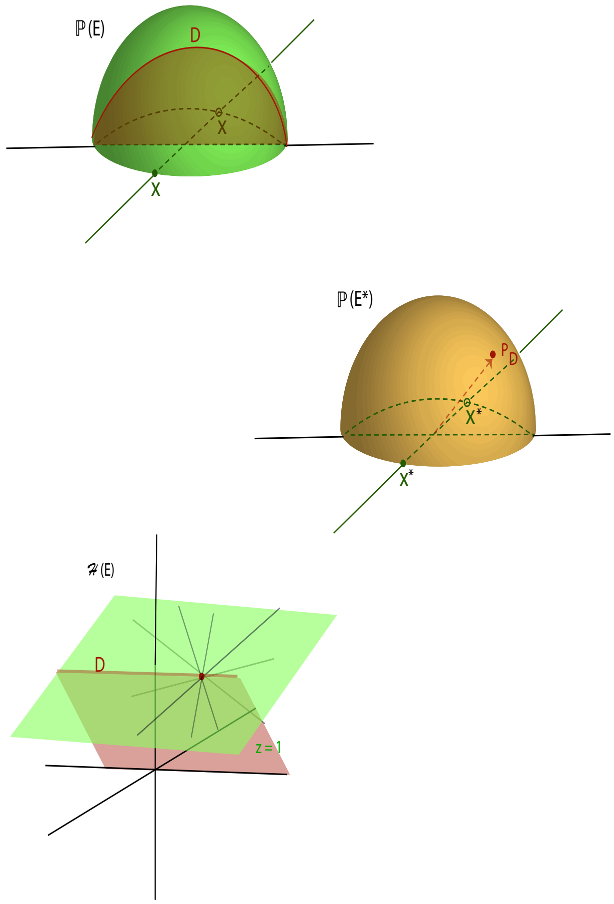


Figure 26.37: The duality between a line in $\mathbf{P}(E)$ and point in $\mathbf{P}(E^*)$. The point in $\mathbf{P}(E^*)$ is also represented by Line D in $\mathcal{H}(E)$.

26.13 Cross-Ratios of Hyperplanes

Given a pencil $P = \mathbf{P}(U)$ of hyperplanes in $\mathcal{H}(E)$, for any sequence $(H_1,\ H_2,\ H_3,\ H_4)$ of hyperplanes in this pencil, if H_1, H_2, H_3 are distinct, we define the cross-ratio $[H_1, H_2, H_3, H_4]$ as the cross-ratio of the hyperplanes H_i considered as points on the projective line P in $\mathbf{P}(E^*)$. In particular, in a projective plane $\mathbf{P}(E)$, given any four concurrent lines D_1, D_2, D_3, D_4, where D_1, D_2, D_3 are distinct, for any two distinct lines Δ and Δ' not passing through the common intersection c of the lines D_i, letting $d_i = \Delta \cap D_i$, and $d'_i = \Delta' \cap D_i$, note that the projection of center c from Δ to Δ' maps each d_i to d'_i.

Since such a projection is a projectivity, and since projectivities between lines preserve cross-ratios, we have

$$[d_1, d_2, d_3, d_4] = [d'_1, d'_2, d'_3, d'_4],$$

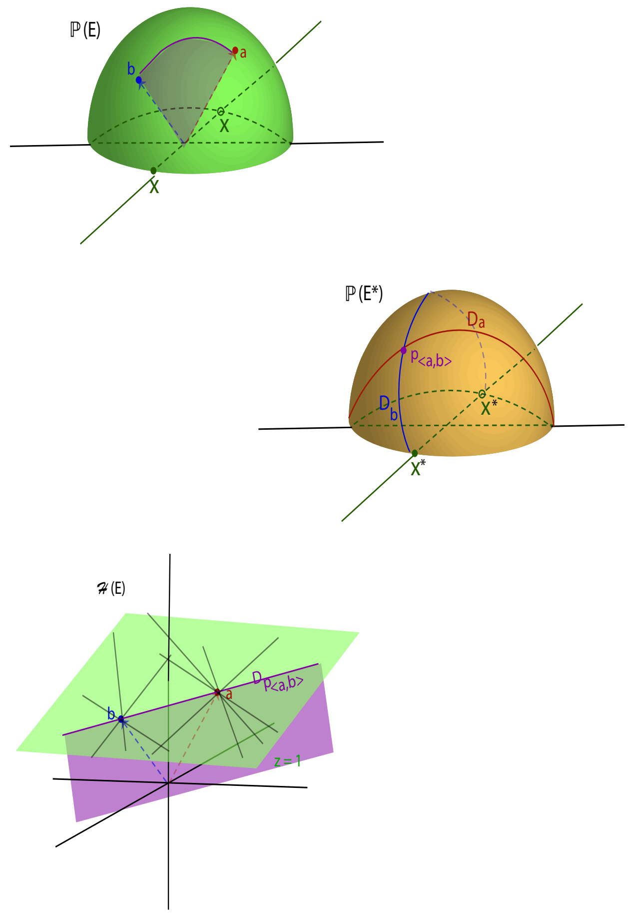


Figure 26.38: The duality between a line through two points in $\mathbf{P}(E)$ and a point incident to two lines in $\mathbf{P}(E^*)$.

which means that the cross-ratio of the d_i is independent of the line Δ (see Figure 26.39).

In fact, this cross-ratio is equal to $[D_1, D_2, D_3, D_4]$, as shown in the next proposition.

Proposition 26.27. *Let $P = \mathbf{P}(U)$ be a pencil of hyperplanes in $\mathcal{H}(E)$, and let $\Delta = \mathbf{P}(D)$ be any projective line such that $\Delta \notin H$ for all $H \in P$. The map $h\colon P \to \Delta$ defined such that $h(H) = H \cap \Delta$ for every hyperplane $H \in P$ is a projectivity. Furthermore, for any sequence (H_1, H_2, H_3, H_4) of hyperplanes in the pencil P, if H_1, H_2, H_3 are distinct and $d_i = \Delta \cap H_i$, then $[d_1, d_2, d_3, d_4] = [H_1, H_2, H_3, H_4]$.*

Proof. First, the map $h\colon P \to \Delta$ is well–defined, since in a projective space, every line $\Delta = \mathbf{P}(D)$ not contained in a hyperplane intersects this hyperplane in exactly one point. Since $P = \mathbf{P}(U)$ is a pencil of hyperplanes in $\mathcal{H}(E)$, U has dimension 2, and let φ and ψ be two nonnull linear forms in E^* that constitute a basis of U, and let $F = \varphi^{-1}(0)$ and

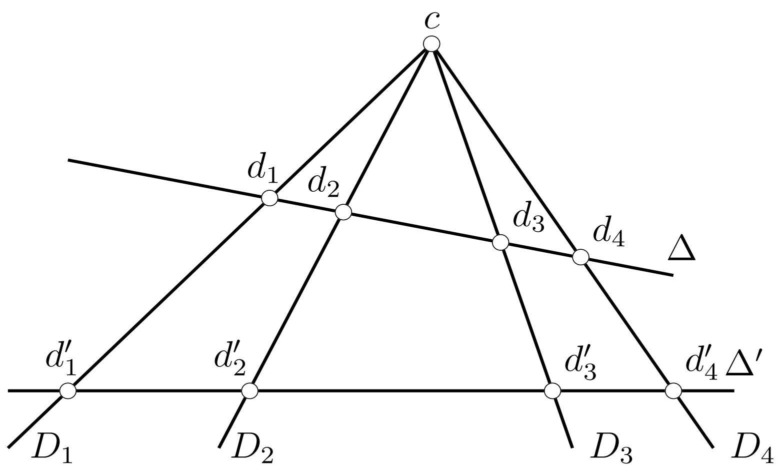


Figure 26.39: A pencil of lines and its cross-ratio with intersecting lines.

$G = \psi^{-1}(0)$. Let $a = \mathbf{P}(F) \cap \Delta$ and $b = \mathbf{P}(G) \cap \Delta$. There are some vectors $u, v \in D$ such that $a = p(u)$ and $b = p(v)$, and since φ and ψ are linearly independent, we have $a \neq b$, and we can choose φ and ψ such that $\varphi(v) = -1$ and $\psi(u) = 1$. Also, (u, v) is a basis of D. Then a point $p(\alpha u + \beta v)$ on Δ belongs to the hyperplane $H = p(\gamma\varphi + \delta\psi)$ of the pencil P iff

$$(\gamma\varphi + \delta\psi)(\alpha u + \beta v) = 0,$$

which, since $\varphi(u) = 0$, $\psi(v) = 0$, $\varphi(v) = -1$, and $\psi(u) = 1$, yields $\gamma\beta = \delta\alpha$, which is equivalent to $[\alpha, \beta] = [\gamma, \delta]$ in $\mathbf{P}(K^2)$. But then the map $h \colon P \to \Delta$ is a projectivity. Letting $d_i = \Delta \cap H_i$, since by Proposition 26.20 a projectivity of lines preserves the cross-ratio, we get $[d_1, d_2, d_3, d_4] = [H_1, H_2, H_3, H_4]$. $\square$

26.14 Complexification of a Real Projective Space

Notions such as orthogonality, angles, and distance between points are not projective concepts. In order to define such notions, one needs an inner product on the underlying vector space. We say that such notions belong to *Euclidean geometry*. At first glance, the fact that some important Euclidean concepts are not covered by projective geometry seems a major drawback of projective geometry. Fortunately, geometers of the nineteenth century (including Laguerre, Monge, Poncelet, Chasles, von Staudt, Cayley, and Klein) found an astute way of recovering certain Euclidean notions such as angles and orthogonality (also circles) by embedding real projective spaces into complex projective spaces. In the next two sections we will give a brief account of this method. More details can be found in Berger [11, 12], Pedoe [136], Samuel [142], Coxeter [43, 44], Sidler [161], Tisseron [175], Lehmann and Bkouche [115], and, of course, Volume II of Veblen and Young [184].

The first step is to embed a real vector space E into a complex vector space $E_{\mathbb{C}}$. A quick but somewhat bewildering way to do so is to define the complexification of E as the tensor product $\mathbb{C} \otimes E$. A more tangible way is to define the following structure.

Definition 26.13. Given a real vector space E, let $E_{\mathbb{C}}$ be the structure $E \times E$ under the addition operation

$$(u_1,\, u_2) + (v_1,\, v_2) = (u_1 + v_1,\, u_2 + v_2),$$

and let multiplication by a complex scalar $z = x + iy$ be defined such that

$$(x + iy) \cdot (u,\, v) = (xu - yv,\, yu + xv).$$

It is easily shown that the structure $E_{\mathbb{C}}$ is a complex vector space. It is also immediate that

$$(0,\, v) = i(v,\, 0),$$

and thus, identifying E with the subspace of $E_{\mathbb{C}}$ consisting of all vectors of the form $(u,\, 0)$, we can write

$$(u,\, v) = u + iv.$$

Given a vector $w = u + iv$, its *conjugate* $\overline{w}$ is the vector $\overline{w} = u - iv$. Then conjugation is a map from $E_{\mathbb{C}}$ to itself that is an involution. If $(e_1, \ldots, e_n)$ is any basis of E, then $((e_1, 0), \ldots, (e_n, 0))$ is a basis of $E_{\mathbb{C}}$. We call such a basis a *real basis*.

Given a linear map $f\colon E \to E$, the map f can be extended to a linear map $f_{\mathbb{C}}\colon E_{\mathbb{C}} \to E_{\mathbb{C}}$ defined such that

$$f_{\mathbb{C}}(u + iv) = f(u) + if(v).$$

We define the *complexification* of $\mathbf{P}(E)$ as $\mathbf{P}(E_{\mathbb{C}})$. If $\big(E, \overrightarrow{E}\big)$ is a real affine space, we define the *complexified projective completion of* $\big(E, \overrightarrow{E}\big)$ as $\mathbf{P}(\widehat{E}_{\mathbb{C}})$ and denote it by $\widetilde{E}_{\mathbb{C}}$. Then $\widetilde{E}$ is naturally embedded in $\widetilde{E}_{\mathbb{C}}$, and it is called the set of *real points* of $\widetilde{E}_{\mathbb{C}}$.

If E has dimension $n+1$ and $(e_1, \ldots, e_{n+1})$ is a basis of E, given any homogeneous polynomial $P(x_1, \ldots, x_{n+1})$ over $\mathbb{C}$ of total degree m, because P is homogeneous, it is immediately verified that

$$P(x_1, \ldots, x_{n+1}) = 0$$

iff

$$P(\lambda x_1, \ldots, \lambda x_{n+1}) = 0,$$

for any $\lambda \neq 0$. Thus, we can define the *hypersurface* $V(P)$ *of equation* $P(x_1, \ldots, x_{n+1}) = 0$ as the subset of $\widetilde{E}_{\mathbb{C}}$ consisting of all points of homogeneous coordinates $(x_1, \ldots, x_{n+1})$ such that $P(x_1, \ldots, x_{n+1}) = 0$. We say that the hypersurface $V(P)$ of equation $P(x_1, \ldots, x_{n+1}) = 0$ is *real* whenever $P(x_1, \ldots, x_{n+1}) = 0$ implies that $P(\overline{x}_1, \ldots, \overline{x}_{n+1}) = 0$.

Note that a real hypersurface may have points other than real points, or no real points at all. For example,

$$x^2 + y^2 - z^2 = 0$$

contains real and complex points such as $(1, i, 0)$ and $(1, -i, 0)$, and

$$x^2 + y^2 + z^2 = 0$$

contains only complex points. When $m = 2$ (where m is the total degree of P), a hypersurface is called a *quadric*, and when $m = 2$ and $n = 2$, a *conic*. When $m = 1$, a hypersurface is just a hyperplane.

Given any homogeneous polynomial $P(x_1, \ldots, x_{n+1})$ over $\mathbb{R}$ of total degree m, since $\mathbb{R} \subseteq \mathbb{C}$, P viewed as a homogeneous polynomial over $\mathbb{C}$ defines a hypersurface $V(P)_{\mathbb{C}}$ in $\widetilde{E}_{\mathbb{C}}$, and also a hypersurface $V(P)$ in $\mathbf{P}(E)$. It is clear that $V(P)$ is naturally embedded in $V(P)_{\mathbb{C}}$, and $V(P)_{\mathbb{C}}$ is called the *complexification of* $V(P)$.

We now show how certain real quadrics without real points can be used to define orthogonality and angles.

26.15 Similarity Structures on a Projective Space

We begin with a real Euclidean plane $\big(E, \overrightarrow{E}\big)$. We will show that the angle of two lines D_1 and D_2 can be expressed as a certain cross-ratio involving the lines D_1, D_2 and also two lines D_I and D_J joining the intersection point $D_1 \cap D_2$ of D_1 and D_2 to two complex points at infinity I and J called the *circular points*. However, there is a slight problem, which is that we haven't yet defined the angle of two lines! Recall that we define the (oriented) angle $\widehat{u_1 u_2}$ of two unit vectors u_1, u_2 as the equivalence class of pairs of unit vectors under the equivalence relation defined such that

$$\langle u_1, u_2 \rangle \equiv \langle u_3, u_4 \rangle$$

iff there is some rotation r such that $r(u_1) = u_3$ and $r(u_2) = u_4$. The set of (oriented) angles of vectors is a group isomorphic to the group $\mathbf{SO}(2)$ of plane rotations. If the Euclidean plane is oriented, the measure of the angle of two vectors is defined up to $2k\pi$ ($k \in \mathbb{Z}$). The angle of two vectors has a measure that is either θ or $2\pi - \theta$, where $\theta \in [0, 2\pi[$, depending on the orientation of the plane. The problem with lines is that they are not oriented: A line is defined by a point a and a vector u, but also by a and $-u$. Given any two lines D_1 and D_2, if r is a rotation of angle θ such that $r(D_1) = D_2$, note that the rotation $-r$ of angle $\theta + \pi$ also maps D_1 onto D_2. Thus, in order to define the (oriented) angle $\widehat{D_1 D_2}$ of two lines D_1, D_2, we define an equivalence relation on pairs of lines as follows:

$$\langle D_1, D_2 \rangle \equiv \langle D_3, D_4 \rangle$$

if there is some rotation r such that $r(D_1) = D_2$ and $r(D_3) = D_4$.

It can be verified that the set of (oriented) angles of lines is a group isomorphic to the quotient group $\mathbf{SO}(2)/\{\mathrm{id}, -\mathrm{id}\}$, also denoted by $\mathbf{PSO}(2)$. In order to define the measure of the angle of two lines, the Euclidean plane E must be oriented. The measure of the angle $\widehat{D_1 D_2}$ of two lines is defined up to $k\pi$ ($k \in \mathbb{Z}$). The angle of two lines has a measure that is either θ or $\pi - \theta$, where $\theta \in [0, \pi[$, depending on the orientation of the plane. We now go back to the circular points.

Let (a_0, a_1, a_2, a_3) be any projective frame for $\widetilde{E}_{\mathbb{C}}$ such that (a_0, a_1) arises from an orthonormal basis (u_1, u_2) of $\overrightarrow{E}$ and the line at infinity H corresponds to $z = 0$ (where (x, y, z) are the homogeneous coordinates of a point w.r.t. (a_0, a_1, a_2, a_3)). Consider the points belonging to the intersection of the real conic Σ of equation

$$x^2 + y^2 - z^2 = 0$$

with the line at infinity $z = 0$. For such points, $x^2 + y^2 = 0$ and $z = 0$, and since

$$x^2 + y^2 = (y - ix)(y + ix),$$

we get exactly two points I and J of homogeneous coordinates $(1, -i, 0)$ and $(1, i, 0)$. The points I and J are called the *circular points*, or the *absolute points*, of $\widetilde{E}_{\mathbb{C}}$. They are complex points at infinity. Any line containing either I or J is called an *isotropic line*.

What is remarkable about I and J is that they allow the definition of the angle of two lines in terms of a certain cross-ratio. Indeed, consider two distinct real lines D_1 and D_2 in E, and let D_I and D_J be the isotropic lines joining $D_1 \cap D_2$ to I and J. We will compute the cross-ratio $[D_1, D_2, D_I, D_J]$. For this, we simply have to compute the cross-ratio of the four points obtained by intersecting D_1, D_2, D_I, D_J with any line not passing through $D_1 \cap D_2$. By changing frame if necessary, so that $D_1 \cap D_2 = a_0$, we can assume that the equations of the lines D_1, D_2, D_I, D_J are of the form

$$\begin{aligned} y &= m_1 x, \\ y &= m_2 x, \\ y &= -ix, \\ y &= ix, \end{aligned}$$

leaving the cases $m_1 = \infty$ and $m_2 = \infty$ as a simple exercise. If we choose $z = 0$ as the intersecting line, we need to compute the cross-ratio of the points $(D_1)_\infty = (1, m_1, 0)$, $(D_2)_\infty = (1, m_2, 0)$, $I = (1, -i, 0)$, and $J = (1, i, 0)$, and we get

$$[D_1, D_2, D_I, D_J] = [(D_1)_\infty, (D_2)_\infty, I, J] = \frac{(-i - m_1)}{(i - m_1)} \frac{(i - m_2)}{(-i - m_2)},$$

that is,

$$[D_1, D_2, D_I, D_J] = \frac{m_1 m_2 + 1 + i(m_2 - m_1)}{m_1 m_2 + 1 - i(m_2 - m_1)}.$$

However, since m_1 and m_2 are the slopes of the lines D_1 and D_2, it is well known that if θ is the (oriented) angle between D_1 and D_2, then

$$\tan\theta = \frac{m_2 - m_1}{m_1 m_2 + 1}.$$

Thus, we have

$$[D_1, D_2, D_I, D_J] = \frac{m_1 m_2 + 1 + i(m_2 - m_1)}{m_1 m_2 + 1 - i(m_2 - m_1)} = \frac{1 + i\tan\theta}{1 - i\tan\theta},$$

that is,

$$[D_1, D_2, D_I, D_J] = \cos 2\theta + i\sin 2\theta = e^{i2\theta}.$$

One can check that the formula still holds when $m_1 = \infty$ or $m_2 = \infty$, and also when $D_1 = D_2$. The formula

$$[D_1, D_2, D_I, D_J] = e^{i2\theta}$$

is known as *Laguerre's formula*.

If U denotes the group $\{e^{i\theta} \mid -\pi \leq \theta \leq \pi\}$ of complex numbers of modulus 1, recall that the map $\Lambda\colon \mathbb{R} \to U$ defined such that

$$\Lambda(t) = e^{it}$$

is a group homomorphism such that $\Lambda^{-1}(1) = 2k\pi$, where $k \in \mathbb{Z}$. The restriction

$$\Lambda\colon\ ]-\pi,\, \pi[\, \to (U - \{-1\})$$

of Λ to $]-\pi,\, \pi[$ is a bijection, and its inverse will be denoted by

$$\log_U\colon (U - \{-1\}) \to\,]-\pi,\, \pi[\,.$$

For stating Proposition 26.28 more conveniently, we extend $\log_U$ to U by letting $\log_U(-1) = \pi$, even though the resulting function is not continuous at -1!. Then we can write

$$\theta = \frac{1}{2}\log_U([D_1, D_2, D_I, D_J]).$$

If the orientation of the plane E is reversed, θ becomes $\pi - \theta$, and since

$$e^{i2(\pi-\theta)} = e^{2i\pi - i2\theta} = e^{-i2\theta},$$

$\log_U(e^{i2(\pi-\theta)}) = -\log_U(e^{i2\theta})$, and

$$\theta = -\frac{1}{2}\log_U([D_1, D_2, D_I, D_J]).$$

In all cases, we have

$$\theta = \frac{1}{2}|\log_U([D_1, D_2, D_I, D_J])|,$$

a formula due to Cayley. We summarize the above in the following proposition.

Proposition 26.28. *Given any two lines D_1, D_2 in a real Euclidean plane $(E, \overrightarrow{E})$, letting D_I and D_J be the isotropic lines in $\widetilde{E}_{\mathbb{C}}$ joining the intersection point $D_1 \cap D_2$ of D_1 and D_2 to the circular points I and J, if θ is the angle of the two lines D_1, D_2, we have*

$$[D_1, D_2, D_I, D_J] = e^{i2\theta},$$

known as Laguerre's formula, and independently of the orientation of the plane, we have

$$\theta = \frac{1}{2}|\log_U([D_1, D_2, D_I, D_J])|,$$

known as Cayley's formula.

In particular, note that $\theta = \pi/2$ iff $[D_1, D_2, D_I, D_J] = -1$, that is, if (D_1, D_2, D_I, D_J) forms a harmonic division. Thus, two lines D_1 and D_2 are orthogonal iff they form a harmonic division with D_I and D_J.

The above considerations show that it is not necessary to assume that $(E, \overrightarrow{E})$ is a real Euclidean plane to define the angle of two lines and orthogonality. Instead, it is enough to assume that two complex conjugate points I, J on the line H at infinity are given. We say that $\langle I, J\rangle$ provides a *similarity structure* on $\widetilde{E}_{\mathbb{C}}$. Note in passing that a circle can be defined as a conic in $\widetilde{E}_{\mathbb{C}}$ that contains the circular points I, J. Indeed, the equation of a conic is of the form

$$ax^2 + by^2 + cxy + dxz + eyz + fz^2 = 0.$$

If this conic contains the circular points $I = (1, -i, 0)$ and $J = (1, i, 0)$, we get the two equations

$$\begin{aligned} a - b - ic &= 0, \\ a - b + ic &= 0, \end{aligned}$$

from which we get $2ic = 0$ and $a = b$, that is, $c = 0$ and $a = b$. The resulting equation

$$ax^2 + ay^2 + dxz + eyz + fz^2 = 0$$

is indeed that of a circle.

Instead of using the function $\log_U \colon (U - \{-1\}) \to]-\pi, \pi[$ as logarithm, one may use the complex logarithm function $\log\colon \mathbb{C}^* \to B$, where $\mathbb{C}^* = \mathbb{C} - \{0\}$ and

$$B = \{x + iy \mid x, y \in \mathbb{R},\ -\pi < y \leq \pi\}.$$

Indeed, the restriction of the complex exponential function $z \mapsto e^z$ to B is bijective, and thus, log is well-defined on C^* (note that log is a homeomorphism from $\mathbb{C} - \{x \mid x \in \mathbb{R},\ x \leq 0\}$ onto $\{x + iy \mid x, y \in \mathbb{R},\ -\pi < y < \pi\}$, the interior of B). Then Cayley's formula reads as

$$\theta = \frac{1}{2i}\log([D_1, D_2, D_I, D_J]),$$

with a $\pm$ in front when the plane is nonoriented. Observe that this formula allows the definition of the angle of two complex lines (possibly a complex number) and the notion of orthogonality of complex lines. In this case, note that the isotropic lines are orthogonal to themselves!

The definition of orthogonality of two lines D_1, D_2 in terms of (D_1, D_2, D_I, D_J) forming a harmonic division can be used to give elegant proofs of various results. Cayley's formula can even be used in computer vision to explain modeling and calibrating cameras! (see Faugeras [59]). As an illustration, consider a triangle (a, b, c), and recall that the line a' passing through a and orthogonal to (b, c) is called the *altitude of* a, and similarly for b and c. It is well known that the altitudes a', b', c' intersect in a common point called the *orthocenter* of the triangle (a, b, c). This can be shown in a number of ways using the circular points. Indeed, letting bc_∞, ab_∞, ac_∞, a'_∞, b'_∞, and c'_∞ denote the points at infinity of the lines $\langle b, c\rangle, \langle a, b\rangle$, $\langle a, c\rangle$, a', b', and c', we have

$$[bc_\infty, a'_\infty, I, J] = -1, \quad [ab_\infty, c'_\infty, I, J] = -1, \quad [ac_\infty, b'_\infty, I, J] = -1,$$

and it is easy to show that there is an involution σ of the line at infinity such that

$$\begin{aligned}
\sigma(I) &= J, \\
\sigma(J) &= I, \\
\sigma(bc_\infty) &= a'_\infty, \\
\sigma(ab_\infty) &= c'_\infty, \\
\sigma(ac_\infty) &= b'_\infty.
\end{aligned}$$

Then, it can be shown that the lines a', b', c' are concurrent. For more details and other results, notably on the conics, see Sidler [161], Berger [12], and Samuel [142].

The generalization of what we just did to real Euclidean spaces $\left(E, \overrightarrow{E}\right)$ of dimension n is simple. Let $(a_0, \ldots, a_{n+1})$ be any projective frame for $\widetilde{E}_{\mathbb{C}}$ such that $(a_0, \ldots, a_{n-1})$ arises from an orthonormal basis $(u_1, \ldots, u_n)$ of $\overrightarrow{E}$ and the hyperplane at infinity H corresponds to $x_{n+1} = 0$ (where $(x_1, \ldots, x_{n+1})$ are the homogeneous coordinates of a point with respect to $(a_0, \ldots, a_{n+1})$). Consider the points belonging to the intersection of the real quadric Σ of equation

$$x_1^2 + \cdots + x_n^2 - x_{n+1}^2 = 0$$

with the hyperplane at infinity $x_{n+1} = 0$. For such points,

$$x_1^2 + \cdots + x_n^2 = 0 \quad \text{and} \quad x_{n+1} = 0.$$

Such points belong to a quadric called the *absolute quadric* of $\widetilde{E}_{\mathbb{C}}$, and denoted by Ω. Any line containing any point on the absolute quadric is called an *isotropic line*. Then, given any two coplanar lines D_1 and D_2 in E, these lines intersect the hyperplane at infinity H in two points $(D_1)_\infty$ and $(D_2)_\infty$, and the line Δ joining $(D_1)_\infty$ and $(D_2)_\infty$ intersects the absolute quadric Ω in two conjugate points I_Δ and J_Δ (also called circular points). It can be shown that the angle θ between D_1 and D_2 is defined by Laguerre's formula:

$$[(D_1)_\infty, (D_2)_\infty, I_\Delta, J_\Delta] = [D_1, D_2, D_{I_\Delta}, D_{J_\Delta}] = e^{i2\theta},$$

where D_{I_Δ} and D_{J_Δ} are the lines joining the intersection $D_1 \cap D_2$ of D_1 and D_2 to the circular points I_Δ and J_Δ.

As in the case of a plane, the above considerations show that it is not necessary to assume that $\left(E, \overrightarrow{E}\right)$ is a real Euclidean space to define the angle of two lines and orthogonality. Instead, it is enough to assume that a nondegenerate real quadric Ω in the hyperplane at infinity H and without real points is given. In particular, when $n = 3$, the absolute quadric Ω is a nondegenerate real conic consisting of complex points at infinity. We say that Ω provides a *similarity structure* on $\widetilde{E}_{\mathbb{C}}$.

It is also possible to show that the real projectivities of $\widetilde{E}_{\mathbb{C}}$ that leave both the hyperplane H at infinity and the absolute quadric Ω (globally) invariant form a group which is none other than the group of affine similarities; see Lehmann and Bkouche [115] (Chapter 10, page 321), and Berger [11] (Chapter 8, Proposition 8.8.6.4).

Definition 26.14. Let $(E, \overrightarrow{E}, \langle -, - \rangle)$ be a Euclidean affine space of finite dimension. An *affine similarity* of $(E, \overrightarrow{E})$ is an invertible affine map $f \in \mathbf{GA}(E)$ such that if $\overrightarrow{f}$ is the linear map associated with f, then there is some positive real $\rho > 0$ satisfying the condition $\left\| \overrightarrow{f}(u) \right\| = \rho \left\| u \right\|$ for all $u \in \overrightarrow{E}$. The number ρ is called the *ratio* of the affine similarity f.

If $f \in \mathbf{GA}(E)$ is an affine similarity of ratio ρ, let $\overrightarrow{g} = \rho^{-1} \overrightarrow{f}$. Since $\rho > 0$, we have

$$\left\| \overrightarrow{g}(u) \right\| = \left\| \rho^{-1} \overrightarrow{f}(u) \right\| = \rho^{-1} \left\| \overrightarrow{f}(u) \right\| = \rho^{-1} \rho \left\| u \right\| = \left\| u \right\|$$

for all $u \in \overrightarrow{E}$, and by Proposition 12.12, the map $\overrightarrow{g} = \rho^{-1} \overrightarrow{f}$ is an isometry; that is, $\overrightarrow{g} \in \mathbf{O}(E)$.

Consequently, every affine similarity f of E can be written as the composition of an isometry (a member of $\mathbf{O}(E)$), a central dilatation, and a translation. For example, when $n = 2$, a similarity is a transformation of the form

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & -\epsilon b \\ b & \epsilon a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c \\ c' \end{pmatrix},$$

with $\epsilon = \pm 1$ and $a, b, c, c' \in \mathbb{R}$. We have the following result showing that the affine similarities of the plane can be viewed as special kinds of projectivities of $\mathbb{CP}^2$.

Proposition 26.29. *If a projectivity h of $\mathbb{CP}^2$ leaves the set of circular points $\{I, J\}$ fixed and maps the affine space $\mathbb{R}^2$ into itself (where $\mathbb{R}^2$ is viewed as the subspace of all points $(x, y, 1)$ with $x, y \in \mathbb{R}$), then h is an affine similarity.*

Proof. The fact that h leaves the set of circular points $\{I, J\}$ fixed means that either $h(I) = I$ and $h(J) = J$ or $h(I) = J$ and $h(J) = I$. If we define I' and J' by

$$I' = (1, -\epsilon i, 0) \quad \text{and} \quad J' = (1, \epsilon i, 0)$$

where $\epsilon = \pm 1$, then the fact that h leaves the set of circular points $\{I, J\}$ fixed is equivalent to

$$h(I) = I' \quad \text{and} \quad h(J) = J'.$$

Assume that h is represented by the invertible matrix

$$A = \begin{pmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{pmatrix}.$$

Then $h(I) = I'$ and $h(J) = J'$ means that there is some nonzero scalars $\lambda, \mu \in \mathbb{C}$ such

$$\begin{pmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{pmatrix} \begin{pmatrix} 1 \\ -i \\ 0 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ -\epsilon i \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{pmatrix} \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} = \mu \begin{pmatrix} 1 \\ \epsilon i \\ 0 \end{pmatrix}.$$

We obtain the following equations:

$$\begin{aligned} \lambda &= a - ia' & \qquad \mu &= a + ia' \\ -\lambda\epsilon i &= b - ib' & \mu\epsilon i &= b + ib' \\ 0 &= c + ic' & 0 &= c - ic'. \end{aligned}$$

By adding the two equations on the first row we obtain

$$\lambda + \mu = 2a,$$

by subtracting the first equation from the second on the second row we obtain

$$(\lambda + \mu)\epsilon i = 2ib',$$

so we get

$$b' = \epsilon a.$$

By subtracting the first equation from the second on the first row we obtain

$$\mu - \lambda = 2ia',$$

and by adding the equations on the second row we obtain

$$(\mu - \lambda)\epsilon i = 2b,$$

and since $\epsilon = \pm 1$, we have $\epsilon^2 = 1$, so we get

$$a' = -\epsilon b.$$

By adding and subtracting the equations on the third row we obtain

$$c = c' = 0.$$

Since A is invertible, $c'' \neq 0$, and since A is determined up to a nonzero scalar we may assume that $c'' = 1$, and we conclude that

$$A = \begin{pmatrix} a & -\epsilon b & a'' \\ b & \epsilon a & b'' \\ 0 & 0 & 1 \end{pmatrix}.$$

If h maps $\mathbb{R}^2$ into itself, then

$$\begin{pmatrix} a & -\epsilon b & a'' \\ b & \epsilon a & b'' \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

must be real for all $x, y \in \mathbb{R}$, which implies that $a, b, a'', b'' \in \mathbb{R}$. $\square$

The following proposition from Berger [11] (Chapter 8, Proposition 8.8.5.1) gives a convenient characterization of the affine similarities.

Proposition 26.30. *Let $(E, \overrightarrow{E}, \langle -, - \rangle)$ be a Euclidean affine space of finite dimension $n \geq 2$. An affine map $f \in \mathbf{GA}(E)$ is an affine similarity iff $\overrightarrow{f}$ preserves orthogonality; that is, for any two vectors $u, v \in \overrightarrow{E}$, if $\langle u, v \rangle = 0$, then $\langle \overrightarrow{f}(u), \overrightarrow{f}(v) \rangle = 0$.*

Proof. Assume that $f \in \mathbf{GA}(E)$ is an affine map such that for any two vectors $u, v \in \overrightarrow{E}$, if $\langle u, v \rangle = 0$, then $\langle \overrightarrow{f}(u), \overrightarrow{f}(v) \rangle = 0$. Fix any nonzero $u \in \overrightarrow{E}$ and consider the linear form φ_u given by

$$\varphi_u(v) = \langle \overrightarrow{f}(u), \overrightarrow{f}(v) \rangle, \quad v \in \overrightarrow{E}.$$

Since $\overrightarrow{f}$ is invertible, $\varphi_u(u) \neq 0$. For any $v \in \overrightarrow{E}$ such that $\langle u, v \rangle = 0$, we have

$$\varphi_u(v) = \langle \overrightarrow{f}(u), \overrightarrow{f}(v) \rangle = 0,$$

thus φ_u is a nonzero linear form vanishing on the hyperplane H orthogonal to u, which is the kernel of the linear form $v \mapsto \langle u, v \rangle$. Therefore, there is some nonzero scalar $\rho(u) \in \mathbb{R}$ such that

$$\varphi_u(v) = \rho(u) \langle u, v \rangle \quad \text{for all } v \in \overrightarrow{E}.$$

Evaluating φ_u at u, we see that $\rho(u) > 0$. If we can show that $\rho(u)$ is a constant $\rho > 0$ independent of u, we will have shown that

$$\langle \overrightarrow{f}(u), \overrightarrow{f}(v) \rangle = \rho \langle u, v \rangle \quad \text{for all } u, v \in \overrightarrow{E},$$

and we will be done.

Since $\dim(E) \geq 2$, pick v to be any nonzero vector in $\overrightarrow{E}$ such that u and v are linearly independent, and let us evaluate $\langle \overrightarrow{f}(u+v), \overrightarrow{f}(w)\rangle$ for any $w \in \overrightarrow{E}$. We have

$$\begin{aligned}\langle \overrightarrow{f}(u+v), \overrightarrow{f}(w)\rangle &= \varphi_{u+v}(w)\\ &= \rho(u+v)\langle u+v, w\rangle\\ &= \rho(u+v)\langle u, w\rangle + \rho(u+v)\langle v, w\rangle\end{aligned}$$

and

$$\begin{aligned}\langle \overrightarrow{f}(u+v), \overrightarrow{f}(w)\rangle &= \langle \overrightarrow{f}(u) + \overrightarrow{f}(v), \overrightarrow{f}(w)\rangle\\ &= \langle \overrightarrow{f}(u), \overrightarrow{f}(w)\rangle + \langle \overrightarrow{f}(v), \overrightarrow{f}(w)\rangle\\ &= \rho(u)\langle u, w\rangle + \rho(v)\langle v, w\rangle,\end{aligned}$$

so we get

$$\langle (\rho(u+v) - \rho(u))u + (\rho(u+v) - \rho(v))v, w\rangle = 0 \quad \text{for all } w \in \overrightarrow{E},$$

which implies that

$$(\rho(u+v) - \rho(u))u + (\rho(u+v) - \rho(v))v = 0.$$

Since u and v are linearly independent, we must have

$$\rho(u+v) = \rho(u) = \rho(v).$$

This proves that $\rho(u)$ is a constant ρ independent of u, as claimed.

The converse is trivial. $\square$

Remark: Let $f \in \mathbf{GA}(E)$ be an affine similarity of ratio ρ. If either $\rho \neq 1$ or $\rho = 1$ and $\overrightarrow{f} \in \mathbf{O}(E)$ does not admit the eigenvalue 1, then f has a unique fixed point.

Indeed, we have $\overrightarrow{f} = \rho \overrightarrow{g}$ for some $\rho > 0$ and some linear isometry $\overrightarrow{g} \in \mathbf{O}(E)$, so for any origin $a \in E$, the point $a + u$ is a fixed point of f iff

$$f(a+u) = a + u$$

iff

$$f(a) + \overrightarrow{f}(u) = a + u$$

iff

$$\rho \overrightarrow{g}(u) = \overrightarrow{f(a)a} + u$$

iff

$$(\overrightarrow{g} - \rho^{-1}\mathrm{id})(u) = \rho^{-1}\overrightarrow{f(a)a}.$$

The linear map $\overrightarrow{g} - \rho^{-1}\mathrm{id}$ is singular iff ρ^{-1} is an eigenvalue or $\overrightarrow{g}$, and since $\overrightarrow{g} \in \mathbf{O}(E)$ its eigenvalues have modulus 1, so if $\rho \neq 1$ or if $\rho = 1$ is not an eigenvalue of $\overrightarrow{g}$, then $\overrightarrow{g} - \rho^{-1}\mathrm{id}$ is invertible, and then there is a unique $u \in \overrightarrow{E}$ such that

$$(\overrightarrow{g} - \rho^{-1}\mathrm{id})(u) = \rho^{-1}\overrightarrow{f(a)a}.$$

For more details on the use of absolute quadrics to obtain some very sophisticated results, the reader should consult Berger [11, 12], Pedoe [136], Samuel [142], Coxeter [43], Sidler [161], Tisseron [175], Lehmann and Bkouche [115], and, of course, Volume II of Veblen and Young [184], which also explains how some non-Euclidean geometries are obtained by chosing the absolute quadric in an appropriate fashion (after Cayley and Klein).

26.16 Some Applications of Projective Geometry

Projective geometry is definitely a jewel of pure mathematics and one of the major mathematical achievements of the nineteenth century. It turns out to be a prerequisite for algebraic geometry, but to our surprise (and pleasure), it also turns out to have applications in engineering. In this short section we summarize some of these applications.

We first discuss applications of projective geometry to camera calibration, a crucial problem in computer vision. Our brief presentation follows quite closely Trucco and Verri [178] (Chapter 2 and Chapter 6). One should also consult Faugeras [59], or Jain, Katsuri, and Schunck [100].

The *pinhole* (or *perspective*) model of a camera is a typical example from computer vision that can be explained very simply in terms of projective transformations. A pinhole camera consists of a point $\mathbf{O}$ called the *center* or *focus of projection*, and a plane π (not containing $\mathbf{O}$) called the *image plane*. The distance f from the image plane π to the center $\mathbf{O}$ is called the *focal length*. The line through $\mathbf{O}$ and perpendicular to π is called the *optical axis*, and the point $\mathbf{o}$, intersection of the optical axis with the image plane is called the *principal point* or *image center*. The way the camera works is that a point P in 3D space is projected onto the image plane (the film) to a point p via the central projection of center $\mathbf{O}$.

It is assumed that an orthonormal frame $\mathcal{F}_c$ is attached to the camera, with its origin at $\mathbf{O}$ and its z-axis parallel to the optical axis. Such a frame is called the *camera reference frame*. With respect to the camera reference frame, it is very easy to write the equations relating the coordinates (x, y) (omitting $z = f$) of the image p (in the image plane π) of a point P of coordinates (X, Y, Z):

$$x = f\,\frac{X}{Z}, \quad y = f\,\frac{Y}{Z}.$$

Typically, points in 3D space are defined by their coordinates not with respect to the camera reference frame, but with respect to another frame $\mathcal{F}_w$, called the *world reference frame*.

However, for most computer vision algorithms, it is necessary to know the coordinates of a point in 3D space with respect to the camera reference frame. Thus, it is necessary to know the position and orientation of the camera with respect to the frame $\mathcal{F}_w$. The position and orientation of the camera are given by some affine transformation $(R, \mathbf{T})$ mapping the frame $\mathcal{F}_w$ to the frame $\mathcal{F}_c$, where R is a rotation matrix and $\mathbf{T}$ is a translation vector. Furthermore, the coordinates of an image point are typically known in terms of *pixel coordinates*, and it is also necessary to transform the coordinates of an image point with respect to the camera reference frame to pixel coordinates. In summary, it is necessary to know the transformation that maps a point P in world coordinates (w.r.t. $\mathcal{F}_w$) to pixel coordinates.

This transformation of world coordinates to pixel coordinates turns out to be a projective transformation that depends on the extrinsic and the intrinsic parameters of the camera. The *extrinsic parameters* of a camera are the location and orientation of the camera with respect to the world reference frame $\mathcal{F}_w$. It is given by an affine map (in fact, a rigid motion, see Chapter 13, Section 27.2). The *intrinsic parameters* of a camera are the parameters needed to link the pixel coordinates of an image point to the corresponding coordinates in the camera reference frame. If $\mathbf{P_w} = (X_w, Y_w, Z_w)$ and $\mathbf{P_c} = (X_c, Y_c, Z_c)$ are the coordinates of the 3D point P with respect to the frames $\mathcal{F}_w$ and $\mathcal{F}_c$, respectively, we can write

$$\mathbf{P_c} = R(\mathbf{P_w} - \mathbf{T}).$$

Neglecting distorsions possibly introduced by the optics, the correspondence between the coordinates (x, y) of the image point with respect to $\mathcal{F}_c$ and the pixel coordinates $(x_{\mathrm{im}}, y_{\mathrm{im}})$ is given by

$$\begin{aligned} x &= -(x_{\mathrm{im}} - o_x)s_x, \\ y &= -(y_{\mathrm{im}} - o_y)s_y, \end{aligned}$$

where (o_x, o_y) are the pixel coordinates the principal point $\mathbf{o}$ and s_x, s_y are scaling parameters.

After some simple calculations, the upshot of all this is that the transformation between the homogeneous coordinates $(X_w, Y_w, Z_w, 1)$ of a 3D point and its homogeneous pixel coordinates (x_1, x_2, x_3) is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = M \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}$$

where the matrix M, known as the *projection matrix*, is a 3×4 matrix depending on R, $\mathbf{T}$, o_x, o_y, f (the focal length), and s_x, s_y (for the derivation of this equation, see Trucco and Verri [178], Chapter 2).

The problem of estimating the extrinsic and the instrinsic parameters of a camera is known as the *camera calibration* problem. It is an important problem in computer vision.

Now, using the equations

$$\begin{aligned} x &= -(x_{\mathrm{im}} - o_x)s_x, \\ y &= -(y_{\mathrm{im}} - o_y)s_y, \end{aligned}$$

we get

$$\begin{aligned} x_{\mathrm{im}} &= -\frac{f}{s_x}\frac{X_c}{Z_c} + o_x, \\ y_{\mathrm{im}} &= -\frac{f}{s_y}\frac{Y_c}{Z_c} + o_y, \end{aligned}$$

relating the coordinates w.r.t. the camera reference frame to the pixel coordinates. This suggests using the parameters $f_x = f/s_x$ and $f_y = f/s_y$ instead of the parameters f, s_x, s_y. In fact, all we need are the parameters $f_x = f/s_x$ and $\alpha = s_y/s_x$, called the *aspect ratio*. Without loss of generality, it can also be assumed that (o_x, o_y) are known. Then we have a total of eight parameters.

One way of solving the calibration problem is to try estimating f_x, α, the rotation matrix R, and the translation vector $\mathbf{T}$ from N image points (x_i, y_i), projections of N suitably chosen world points (X_i, Y_i, Z_i), using the system of equations obtained from the projection matrix. It turns out that if $N \geq 7$ and the points are not coplanar, the rank of the system is 7, and the system has a nontrivial solution (up to a scalar) that can be found using SVD methods (see Chapter 22, Trucco and Verri [178], or Jain, Katsuri, and Schunck [100]).

Another method consists in estimating the whole projection matrix M, which depends on 11 parameters, and then extracting extrinsic and intrinsic parameters. Again, SVD methods are used (see Trucco and Verri [178], and Faugeras [59]).

Cayley's formula can also be used to solve the calibration cameras, as explained in Faugeras [59]. Other problems in computer vision can be reduced to problems in projective geometry (see Faugeras [59]).

In computer graphics, it is also necessary to convert the 3D world coordinates of a point to a two-dimensional representation on a *view plane*. This is achieved by a so-called *viewing system* using a projective transformation. For details on viewing systems see Watt [189] or Foley, van Dam, Feiner, and Hughes [63].

Projective spaces are also the right framework to deal with rational curves and rational surfaces. Indeed, in the projective framework it is easy to deal with vanishing denominators and with "infinite" values of the parameter(s).

It is much less obvious that projective geometry has applications to efficient communication, error-correcting codes, and cryptography, as very nicely explained by Beutelspacher and Rosenbaum [22]. We sketch these applications very briefly, referring our readers to [22] for details. We begin with efficient communication. Suppose that eight students would like to exchange information to do their homework economically. The idea is that each student solves part of the exercises and copies the rest from the others (which we do not recommend, of course!). It is assumed that each student solves his part of the homework at home, and that the solutions are communicated by phone. The problem is to minimize the number of phone calls. An obvious but expensive method is for each student to call each of the other seven students. A much better method is to imagine that the eight students are the vertices of a cube, say with coordinates from $\{0, 1\}^3$. There are three types of edges:

1. Those parallel to the z-axis, called *type* 1;
2. Those parallel to the y-axis, called *type* 2;
3. Those parallel to the x-axis, called *type* 3.

The communication can proceed in three rounds as follows: All nodes connected by type 1 edges exchange solutions; all nodes connected by type 2 edges exchange solutions; and finally all nodes connected by type 3 edges exchange solutions.

It is easy to see that everybody has all the answers at the end of the three rounds. Furthermore, each student is involved only in three calls (making a call or receiving it), and the total number of calls is twelve.

In the general case, N nodes would like to exchange information in such a way that eventually every node has all the information. A good way to to this is to construct certain finite projective spaces, as explained in Beutelspacher and Rosenbaum [22]. We pick q to be an integer (for instance, a prime number) such that there is a finite projective space of any dimension over the finite field of order q. Then, we pick d such that

$$q^{d-1} < N \leq q^d.$$

Since q is prime, there is a projective space $\mathbf{P}(K^{d+1})$ of dimension d over the finite field K of order q, and letting $\mathcal{H}$ be the hyperplane at infinity in $\mathbf{P}(K^{d+1})$, we pick a frame $P_1, \ldots, P_d$ in $\mathcal{H}$. It turns out that the affine space $\mathcal{A} = \mathbf{P}(K^{d+1}) - \mathcal{H}$ has q^d points. Then the communication nodes can be identified with points in the affine space $\mathcal{A}$. Assuming for simplicity that $N = q^d$, the algorithm proceeds in d rounds. During round i, each node $Q \in \mathcal{A}$ sends the information it has received to all nodes in $\mathcal{A}$ on the line QP_i.

It can be shown that at the end of the d rounds, each node has the total information, and that the total number of transactions is at most

$$(q-1)\log_q(N)N.$$

Other applications of projective spaces to communication systems with switches are described in Chapter 2, Section 8, of Beutelspacher and Rosenbaum [22]. Applications to error-correcting codes are described in Chapter 5 of the same book. Introducing even the most elementary notions of coding theory would take too much space. Let us simply say that the existence of certain types of good codes called *linear* $[n, n-r]$*-codes with minimum* *distance* d is equivalent to the existence of certain sets of points called $(n, d-1)$-sets in the finite projective space $\mathbf{P}(\{0,1\}^r)$. For the sake of completeness, a set of n points in a projective space is an (n, s)-*set* if s is the largest integer such that every subset of s points is projectively independent. For example, an $(n, 3)$-set is a set of n points no three of which are collinear, but at least four of them are coplanar.

Other applications of projective geometry to cryptography are given in Chapter 6 of Beutelspacher and Rosenbaum [22].

Part III

The Geometry of Bilinear Forms

Chapter 27

The Cartan–Dieudonné Theorem

In this chapter the structure of the orthogonal group is studied in more depth. In particular, we prove that every isometry in $\mathbf{O}(n)$ is the composition of at most n reflections about hyperplanes (for $n \geq 2$, see Theorem 27.1). This important result is a special case of the "Cartan–Dieudonné theorem" (Cartan [33], Dieudonné [50]). We also prove that every rotation in $\mathbf{SO}(n)$ is the composition of at most n flips (for $n \geq 3$).

Affine isometries are defined, and their fixed points are investigated. First, we characterize the set of fixed points of an affine map. Then we show that the Cartan–Dieudonné theorem can be generalized to affine isometries: Every rigid motion in $\mathbf{Is}(n)$ is the composition of at most n affine reflections if it has a fixed point, or else of at most $n+2$ affine reflections. We prove that every rigid motion in $\mathbf{SE}(n)$ is the composition of at most n affine flips (for $n \geq 3$).

27.1 The Cartan–Dieudonné Theorem for Linear Isometries

The fact that the group $\mathbf{O}(n)$ of linear isometries is generated by the reflections is a special case of a theorem known as the Cartan–Dieudonné theorem. Elie Cartan proved a version of this theorem early in the twentieth century. A proof can be found in his book on spinors [33], which appeared in 1937 (Chapter I, Section 10, pages 10–12). Cartan's version applies to nondegenerate quadratic forms over $\mathbb{R}$ or $\mathbb{C}$. The theorem was generalized to quadratic forms over arbitrary fields by Dieudonné [50]. One should also consult Emil Artin's book [6], which contains an in-depth study of the orthogonal group and another proof of the Cartan–Dieudonné theorem.

Theorem 27.1. *Let E be a Euclidean space of dimension $n \geq 1$. Every isometry $f \in \mathbf{O}(E)$ that is not the identity is the composition of at most n reflections. When $n \geq 2$, the identity is the composition of any reflection with itself.*

Proof. We proceed by induction on n. When $n = 1$, every isometry $f \in \mathbf{O}(E)$ is either the identity or $-\mathrm{id}$, but $-\mathrm{id}$ is a reflection about $H = \{0\}$. When $n \geq 2$, we have $\mathrm{id} = s \circ s$ for every reflection s. Let us now consider the case where $n \geq 2$ and f is not the identity. There are two subcases.

Case 1. The map f admits 1 as an eigenvalue, i.e., there is some nonnull vector w such that $f(w) = w$. In this case, let H be the hyperplane orthogonal to w, so that $E = H \oplus \mathbb{R}w$. We claim that $f(H) \subseteq H$. Indeed, if

$$v \cdot w = 0$$

for any $v \in H$, since f is an isometry, we get

$$f(v) \cdot f(w) = v \cdot w = 0,$$

and since $f(w) = w$, we get

$$f(v) \cdot w = f(v) \cdot f(w) = 0,$$

and thus $f(v) \in H$. Furthermore, since f is not the identity, f is not the identity of H. Since H has dimension $n-1$, by the induction hypothesis applied to H, there are at most $k \leq n-1$ reflections $s_1, \ldots, s_k$ about some hyperplanes $H_1, \ldots, H_k$ in H, such that the restriction of f to H is the composition $s_k \circ \cdots \circ s_1$. Each s_i can be extended to a reflection in E as follows: If $H = H_i \oplus L_i$ (where $L_i = H_i^{\perp}$, the orthogonal complement of H_i in H), $L = \mathbb{R}w$, and $F_i = H_i \oplus L$, since H and L are orthogonal, F_i is indeed a hyperplane, $E = F_i \oplus L_i = H_i \oplus L \oplus L_i$, and for every $u = h + \lambda w \in H \oplus L = E$, since

$$s_i(h) = p_{H_i}(h) - p_{L_i}(h),$$

we can define s_i on E such that

$$s_i(h + \lambda w) = p_{H_i}(h) + \lambda w - p_{L_i}(h),$$

and since $h \in H$, $w \in L$, $F_i = H_i \oplus L$, and $H = H_i \oplus L_i$, we have

$$s_i(h + \lambda w) = p_{F_i}(h + \lambda w) - p_{L_i}(h + \lambda w),$$

which defines a reflection about $F_i = H_i \oplus L$. Now, since f is the identity on $L = \mathbb{R}w$, it is immediately verified that $f = s_k \circ \cdots \circ s_1$, with $k \leq n-1$. See Figure 27.1.

Case 2. The map f does not admit 1 as an eigenvalue, i.e., $f(u) \neq u$ for all $u \neq 0$. Pick any $w \neq 0$ in E, and let H be the hyperplane orthogonal to $f(w) - w$. Since f is an isometry, we have $\|f(w)\| = \|w\|$, and by Lemma 13.2, we know that $s(w) = f(w)$, where s is the reflection about H, and we claim that $s \circ f$ leaves w invariant. Indeed, since $s^2 = \mathrm{id}$, we have

$$s(f(w)) = s(s(w)) = w.$$

See Figure 27.2.

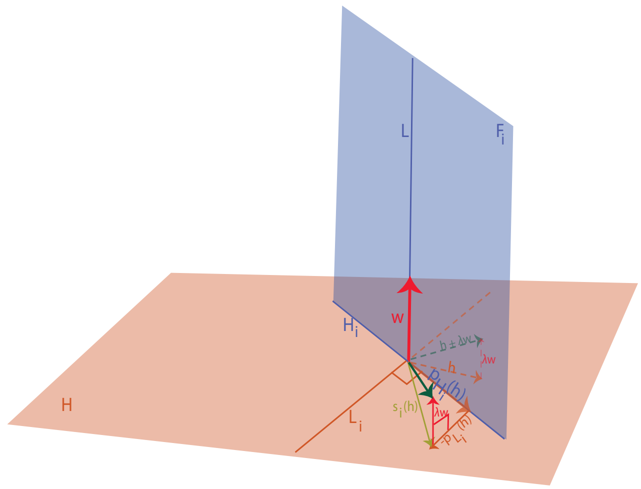


Figure 27.1: An illustration of how to extend the reflection s_i of Case 1 in Theorem 27.1 to E. The result of this extended reflection is the bold green vector.

Since $s^2 = \mathrm{id}$, we cannot have $s \circ f = \mathrm{id}$, since this would imply that $f = s$, where s is the identity on H, contradicting the fact that f is not the identity on any vector. Thus, we are back to Case 1. Thus, there are $k \leq n-1$ hyperplane reflections such that $s \circ f = s_k \circ \cdots \circ s_1$, from which we get

$$f = s \circ s_k \circ \cdots \circ s_1,$$

with at most $k + 1 \leq n$ reflections. □

Remarks:

(1) A slightly different proof can be given. Either f is the identity, or there is some nonnull vector u such that $f(u) \neq u$. In the second case, proceed as in the second part of the proof, to get back to the case where f admits 1 as an eigenvalue.

(2) Theorem 27.1 still holds if the inner product on E is replaced by a nondegenerate symmetric bilinear form φ, but the proof is a lot harder; see Section 29.9.

(3) The proof of Theorem 27.1 shows more than stated. If 1 is an eigenvalue of f, for any eigenvector w associated with 1 (i.e., $f(w) = w$, $w \neq 0$), then f is the composition of $k \leq n - 1$ reflections about hyperplanes F_i such that $F_i = H_i \oplus L$, where L is the line $\mathbb{R}w$ and the H_i are subspaces of dimension $n - 2$ all orthogonal to L (the H_i are hyperplanes in H). This situation is illustrated in Figure 27.3.

If 1 is not an eigenvalue of f, then f is the composition of $k \leq n$ reflections about hyperplanes $H, F_1, \ldots, F_{k-1}$, such that $F_i = H_i \oplus L$, where L is a line intersecting H, and the H_i are subspaces of dimension $n-2$ all orthogonal to L (the H_i are hyperplanes in $L^{\perp}$). This situation is illustrated in Figure 27.4.

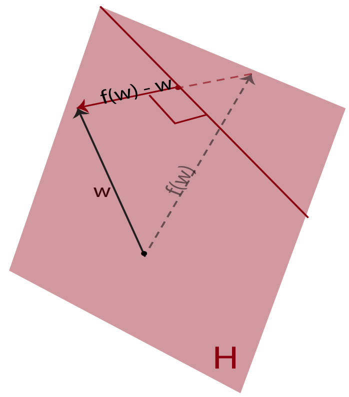


Figure 27.2: The construction of the hyperplane H for Case 2 of Theorem 27.1.

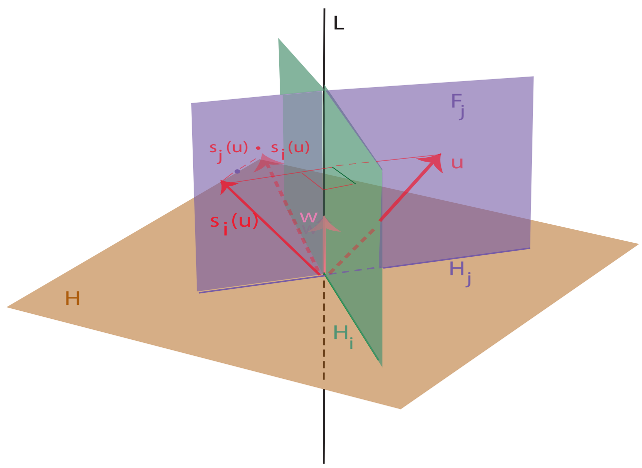


Figure 27.3: An isometry f as a composition of reflections, when 1 is an eigenvalue of f.

(4) It is natural to ask what is the minimal number of hyperplane reflections needed to obtain an isometry f. This has to do with the dimension of the eigenspace $\mathrm{Ker}\,(f - \mathrm{id})$ associated with the eigenvalue 1. We will prove later that every isometry is the composition of k hyperplane reflections, where

$$k = n - \dim(\mathrm{Ker}\,(f - \mathrm{id})),$$

and that this number is minimal (where $n = \dim(E)$).

When $n = 2$, a reflection is a reflection about a line, and Theorem 27.1 shows that every isometry in $\mathbf{O}(2)$ is either a reflection about a line or a rotation, and that every rotation is the product of two reflections about some lines. In general, since $\det(s) = -1$ for a reflection s, when $n \geq 3$ is odd, every rotation is the product of an even number less than or equal

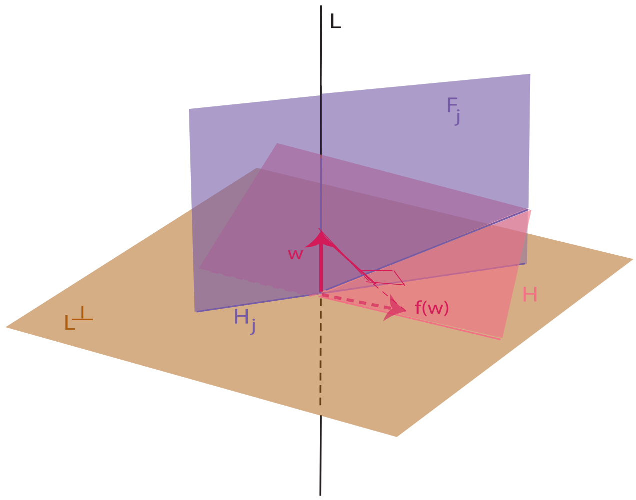


Figure 27.4: An isometry f as a composition of reflections when 1 is not an eigenvalue of f. Note that the pink plane H is perpendicular to $f(w) - w$.

to $n-1$ of reflections, and when n is even, every improper orthogonal transformation is the product of an odd number less than or equal to $n-1$ of reflections.

In particular, for $n = 3$, every rotation is the product of two reflections about planes. When n is odd, we can say more about improper isometries. Indeed, when n is odd, every improper isometry admits the eigenvalue -1. This is because if E is a Euclidean space of finite dimension and $f\colon E \to E$ is an isometry, because $\|f(u)\| = \|u\|$ for every $u \in E$, if λ is any eigenvalue of f and u is an eigenvector associated with λ, then

$$\|f(u)\| = \|\lambda u\| = |\lambda|\|u\| = \|u\|,$$

which implies $|\lambda| = 1$, since $u \neq 0$. Thus, the real eigenvalues of an isometry are either $+1$ or -1. However, it is well known that polynomials of odd degree always have some real root. As a consequence, the characteristic polynomial $\det(f - \lambda \mathrm{id})$ of f has some real root, which is either $+1$ or -1. Since f is an improper isometry, $\det(f) = -1$, and since $\det(f)$ is the product of the eigenvalues, the real roots cannot all be $+1$, and thus -1 is an eigenvalue of f. Going back to the proof of Theorem 27.1, since -1 is an eigenvalue of f, there is some nonnull eigenvector w such that $f(w) = -w$. Using the second part of the proof, we see that the hyperplane H orthogonal to $f(w) - w = -2w$ is in fact orthogonal to w, and thus f is the product of $k \leq n$ reflections about hyperplanes $H, F_1, \ldots, F_{k-1}$ such that $F_i = H_i \oplus L$, where L is a line orthogonal to H, and the H_i are hyperplanes in $H = L^\perp$ orthogonal to L. However, k must be odd, and so $k-1$ is even, and thus the composition of the reflections about $F_1, \ldots, F_{k-1}$ is a rotation. Thus, when n is odd, an improper isometry is the composition of a reflection about a hyperplane H with a rotation consisting of reflections about hyperplanes $F_1, \ldots, F_{k-1}$ containing a line, L, orthogonal to

H. In particular, when $n = 3$, every improper orthogonal transformation is the product of a rotation with a reflection about a plane orthogonal to the axis of rotation.

Using Theorem 27.1, we can also give a rather simple proof of the classical fact that in a Euclidean space of odd dimension, every rotation leaves some nonnull vector invariant, and thus a line invariant.

If λ is an eigenvalue of f, then the following lemma shows that the orthogonal complement $E_\lambda(f)^\perp$ of the eigenspace associated with λ is closed under f.

Proposition 27.2. *Let E be a Euclidean space of finite dimension n, and let $f\colon E \to E$ be an isometry. For any subspace F of E, if $f(F) = F$, then $f(F^\perp) \subseteq F^\perp$ and $E = F \oplus F^\perp$.*

Proof. We just have to prove that if $w \in E$ is orthogonal to every $u \in F$, then $f(w)$ is also orthogonal to every $u \in F$. However, since $f(F) = F$, for every $v \in F$, there is some $u \in F$ such that $f(u) = v$, and we have

$$f(w) \cdot v = f(w) \cdot f(u) = w \cdot u,$$

since f is an isometry. Since we assumed that $w \in E$ is orthogonal to every $u \in F$, we have

$$w \cdot u = 0,$$

and thus

$$f(w) \cdot v = 0,$$

and this for every $v \in F$. Thus, $f(F^\perp) \subseteq F^\perp$. The fact that $E = F \oplus F^\perp$ follows from Lemma 12.11. $\square$

Lemma 27.2 is the starting point of the proof that every orthogonal matrix can be diagonalized over the field of complex numbers. Indeed, if λ is any eigenvalue of f, then $f(E_\lambda(f)) = E_\lambda(f)$, where $E_\lambda(f)$ is the eigenspace associated with λ, and thus the orthogonal $E_\lambda(f)^\perp$ is closed under f, and $E = E_\lambda(f) \oplus E_\lambda(f)^\perp$. The problem over $\mathbb{R}$ is that there may not be any real eigenvalues. However, when n is odd, the following lemma shows that every rotation admits 1 as an eigenvalue (and similarly, when n is even, every improper orthogonal transformation admits 1 as an eigenvalue).

Proposition 27.3. *Let E be a Euclidean space.*

(1) If E has odd dimension $n = 2m + 1$, then every rotation f admits 1 as an eigenvalue and the eigenspace F of all eigenvectors left invariant under f has an odd dimension $2p + 1$. Furthermore, there is an orthonormal basis of E, in which f is represented by a matrix of the form

$$\begin{pmatrix} R_{2(m-p)} & 0 \\ 0 & I_{2p+1} \end{pmatrix},$$

where $R_{2(m-p)}$ is a rotation matrix that does not have 1 as an eigenvalue.

(2) *If E has even dimension $n = 2m$, then every improper orthogonal transformation f admits* 1 *as an eigenvalue and the eigenspace F of all eigenvectors left invariant under f has an odd dimension $2p + 1$. Furthermore, there is an orthonormal basis of E, in which f is represented by a matrix of the form*

$$\begin{pmatrix} S_{2(m-p)-1} & 0 \\ 0 & I_{2p+1} \end{pmatrix},$$

where $S_{2(m-p)-1}$ is an improper orthogonal matrix that does not have 1 *as an eigenvalue.*

Proof. We prove only (1), the proof of (2) being similar. Since f is a rotation and $n = 2m+1$ is odd, by Theorem 27.1, f is the composition of an even number less than or equal to $2m$ of reflections. From Lemma 24.15, recall the Grassmann relation

$$\dim(M) + \dim(N) = \dim(M + N) + \dim(M \cap N),$$

where M and N are subspaces of E. Now, if M and N are hyperplanes, their dimension is $n - 1$, and thus $\dim(M \cap N) \geq n - 2$. Thus, if we intersect $k \leq n$ hyperplanes, we see that the dimension of their intersection is at least $n - k$. Since each of the reflections is the identity on the hyperplane defining it, and since there are at most $2m = n - 1$ reflections, their composition is the identity on a subspace of dimension at least 1. This proves that 1 is an eigenvalue of f. Let F be the eigenspace associated with 1, and assume that its dimension is q. Let $G = F^{\perp}$ be the orthogonal of F. By Lemma 27.2, G is stable under f, and $E = F \oplus G$. Using Lemma 12.10, we can find an orthonormal basis of E consisting of an orthonormal basis for G and orthonormal basis for F. In this basis, the matrix of f is of the form

$$\begin{pmatrix} R_{2m+1-q} & 0 \\ 0 & I_q \end{pmatrix}.$$

Thus, $\det(f) = \det(R)$, and R must be a rotation, since f is a rotation and $\det(f) = 1$. Now, if f left some vector $u \neq 0$ in G invariant, this vector would be an eigenvector for 1, and we would have $u \in F$, the eigenspace associated with 1, which contradicts $E = F \oplus G$. Thus, by the first part of the proof, the dimension of G must be even, since otherwise, the restriction of f to G would admit 1 as an eigenvalue. Consequently, q must be odd, and R does not admit 1 as an eigenvalue. Letting $q = 2p + 1$, the lemma is established. ☐

An example showing that Lemma 27.3 fails for n even is the following rotation matrix (when $n = 2$):

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The above matrix does not have real eigenvalues for $\theta \neq k\pi$.

It is easily shown that for $n = 2$, with respect to any chosen orthonormal basis (e_1, e_2), every rotation is represented by a matrix of form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

where $\theta \in [0, 2\pi[$, and that every improper orthogonal transformation is represented by a matrix of the form

$$S = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

In the first case, we call $\theta \in [0, 2\pi[$ the *measure* of the angle of rotation of R w.r.t. the orthonormal basis (e_1, e_2). In the second case, we have a reflection about a line, and it is easy to determine what this line is. It is also easy to see that S is the composition of a reflection about the x-axis with a rotation (of matrix R).

We refrained from calling θ "the angle of rotation," because there are some subtleties involved in defining rigorously the notion of angle of two vectors (or two lines). For example, note that with respect to the "opposite basis" (e_2, e_1), the measure θ must be changed to $2\pi - \theta$ (or $-\theta$ if we consider the quotient set $\mathbb{R}/2\pi$ of the real numbers modulo 2π).

It is easily shown that the group $\mathbf{SO}(2)$ of rotations in the plane is abelian. First, recall that every plane rotation is the product of two reflections (about lines), and that every isometry in $\mathbf{O}(2)$ is either a reflection or a rotation. To alleviate the notation, we will omit the composition operator $\circ$, and write rs instead of $r \circ s$. Now, if r is a rotation and s is a reflection, rs being in $\mathbf{O}(2)$ must be a reflection (since $\det(rs) = \det(r)\det(s) = -1$), and thus $(rs)^2 = \mathrm{id}$, since a reflection is an involution, which implies that

$$srs = r^{-1}.$$

Then, given two rotations r_1 and r_2, writing r_1 as $r_1 = s_2 s_1$ for two reflections s_1, s_2, we have

$$r_1 r_2 r_1^{-1} = s_2 s_1 r_2 (s_2 s_1)^{-1} = s_2 s_1 r_2 s_1^{-1} s_2^{-1} = s_2 s_1 r_2 s_1 s_2 = s_2 r_2^{-1} s_2 = r_2,$$

since $srs = r^{-1}$ for all reflections s and rotations r, and thus $r_1 r_2 = r_2 r_1$.

We can also perform the following calculation, using some elementary trigonometry:

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix} \begin{pmatrix} \cos\psi & \sin\psi \\ \sin\psi & -\cos\psi \end{pmatrix} = \begin{pmatrix} \cos(\varphi+\psi) & \sin(\varphi+\psi) \\ \sin(\varphi+\psi) & -\cos(\varphi+\psi) \end{pmatrix}.$$

The above also shows that the inverse of a rotation matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is obtained by changing θ to $-\theta$ (or $2\pi - \theta$). Incidentally, note that in writing a rotation r as the product of two reflections $r = s_2 s_1$, the first reflection s_1 can be chosen arbitrarily, since $s_1^2 = \mathrm{id}$, $r = (r s_1) s_1$, and $r s_1$ is a reflection.

For $n = 3$, the only two choices for p are $p = 1$, which corresponds to the identity, or $p = 0$, in which case f is a rotation leaving a line invariant. This line D is called the *axis of*

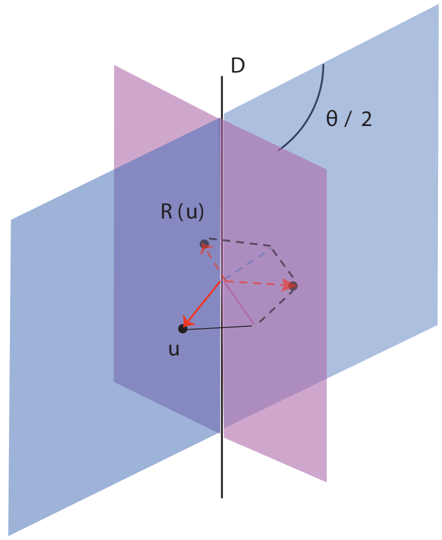


Figure 27.5: 3D rotation as the composition of two reflections.

rotation. The rotation R behaves like a two-dimensional rotation around the axis of rotation. Thus, the rotation R is the composition of two reflections about planes containing the axis of rotation D and forming an angle $\theta/2$. This is illustrated in Figure 27.5.

The measure of the angle of rotation θ can be determined through its cosine via the formula

$$\cos\theta = u \cdot R(u),$$

where u is any unit vector orthogonal to the direction of the axis of rotation. However, this does not determine $\theta \in [0, 2\pi[$ uniquely, since both θ and $2\pi - \theta$ are possible candidates. What is missing is an orientation of the plane (through the origin) orthogonal to the axis of rotation.

In the orthonormal basis of the lemma, a rotation is represented by a matrix of the form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Remark: For an arbitrary rotation matrix A, since $a_{1\,1} + a_{2\,2} + a_{3\,3}$ (the *trace* of A) is the sum of the eigenvalues of A, and since these eigenvalues are $\cos\theta + i\sin\theta$, $\cos\theta - i\sin\theta$, and 1, for some $\theta \in [0, 2\pi[$, we can compute $\cos\theta$ from

$$1 + 2\cos\theta = a_{1\,1} + a_{2\,2} + a_{3\,3}.$$

It is also possible to determine the axis of rotation (see the problems).

An improper transformation is either a reflection about a plane or the product of three reflections, or equivalently the product of a reflection about a plane with a rotation, and we noted in the discussion following Theorem 27.1 that the axis of rotation is orthogonal to the plane of the reflection. Thus, an improper transformation is represented by a matrix of the form

$$S = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

When $n \geq 3$, the group of rotations $\mathbf{SO}(n)$ is not only generated by hyperplane reflections, but also by flips (about subspaces of dimension $n-2$). We will also see, in Section 27.2, that every proper affine rigid motion can be expressed as the composition of at most n flips, which is perhaps even more surprising! The proof of these results uses the following key lemma.

Proposition 27.4. *Given any Euclidean space E of dimension $n \geq 3$, for any two reflections h_1 and h_2 about some hyperplanes H_1 and H_2, there exist two flips f_1 and f_2 such that $h_2 \circ h_1 = f_2 \circ f_1$.*

Proof. If $h_1 = h_2$, it is obvious that

$$h_1 \circ h_2 = h_1 \circ h_1 = \mathrm{id} = f_1 \circ f_1$$

for any flip f_1. If $h_1 \neq h_2$, then $H_1 \cap H_2 = F$, where $\dim(F) = n-2$ (by the Grassmann relation). We can pick an orthonormal basis $(e_1, \ldots, e_n)$ of E such that $(e_1, \ldots, e_{n-2})$ is an orthonormal basis of F. We can also extend $(e_1, \ldots, e_{n-2})$ to an orthonormal basis $(e_1, \ldots, e_{n-2}, u_1, v_1)$ of E, where $(e_1, \ldots, e_{n-2}, u_1)$ is an orthonormal basis of H_1, in which case

$$\begin{aligned} e_{n-1} &= \cos\theta_1\, u_1 + \sin\theta_1\, v_1, \\ e_n &= \sin\theta_1\, u_1 - \cos\theta_1\, v_1, \end{aligned}$$

for some $\theta_1 \in [0, 2\pi]$. See Figure 27.6

Since h_1 is the identity on H_1 and v_1 is orthogonal to H_1, it follows that $h_1(u_1) = u_1$, $h_1(v_1) = -v_1$, and we get

$$\begin{aligned} h_1(e_{n-1}) &= \cos\theta_1\, u_1 - \sin\theta_1\, v_1, \\ h_1(e_n) &= \sin\theta_1\, u_1 + \cos\theta_1\, v_1. \end{aligned}$$

After some simple calculations, we get

$$\begin{aligned} h_1(e_{n-1}) &= \cos 2\theta_1\, e_{n-1} + \sin 2\theta_1\, e_n, \\ h_1(e_n) &= \sin 2\theta_1\, e_{n-1} - \cos 2\theta_1\, e_n. \end{aligned}$$

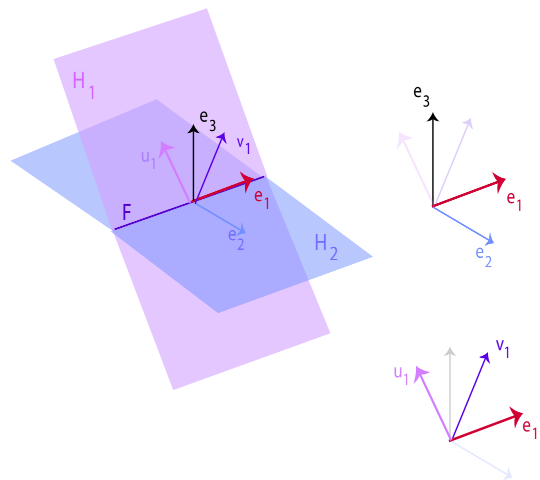


Figure 27.6: An illustration of the hyperplanes H_1, H_2, their intersection F, and the two orthonormal basis utilized in the proof of Proposition 27.4.

As a consequence, the matrix A_1 of h_1 over the basis $(e_1, \ldots, e_n)$ is of the form

$$A_1 = \begin{pmatrix} I_{n-2} & 0 & 0 \\ 0 & \cos 2\theta_1 & \sin 2\theta_1 \\ 0 & \sin 2\theta_1 & -\cos 2\theta_1 \end{pmatrix}.$$

Similarly, the matrix A_2 of h_2 over the basis $(e_1, \ldots, e_n)$ is of the form

$$A_2 = \begin{pmatrix} I_{n-2} & 0 & 0 \\ 0 & \cos 2\theta_2 & \sin 2\theta_2 \\ 0 & \sin 2\theta_2 & -\cos 2\theta_2 \end{pmatrix}.$$

Observe that both A_1 and A_2 have the eigenvalues -1 and $+1$ with multiplicity $n-1$. The trick is to observe that if we change the last entry in I_{n-2} from $+1$ to -1 (which is possible since $n \geq 3$), we have the following product A_2A_1:

$$\begin{pmatrix} I_{n-3} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \cos 2\theta_2 & \sin 2\theta_2 \\ 0 & 0 & \sin 2\theta_2 & -\cos 2\theta_2 \end{pmatrix} \begin{pmatrix} I_{n-3} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \cos 2\theta_1 & \sin 2\theta_1 \\ 0 & 0 & \sin 2\theta_1 & -\cos 2\theta_1 \end{pmatrix}.$$

Now, the two matrices above are clearly orthogonal, and they have the eigenvalues $-1, -1$, and $+1$ with multiplicity $n-2$, which implies that the corresponding isometries leave invariant a subspace of dimension $n-2$ and act as $-\mathrm{id}$ on its orthogonal complement (which has dimension 2). This means that the above two matrices represent two flips f_1 and f_2 such that $h_2 \circ h_1 = f_2 \circ f_1$. See Figure 27.7. $\square$

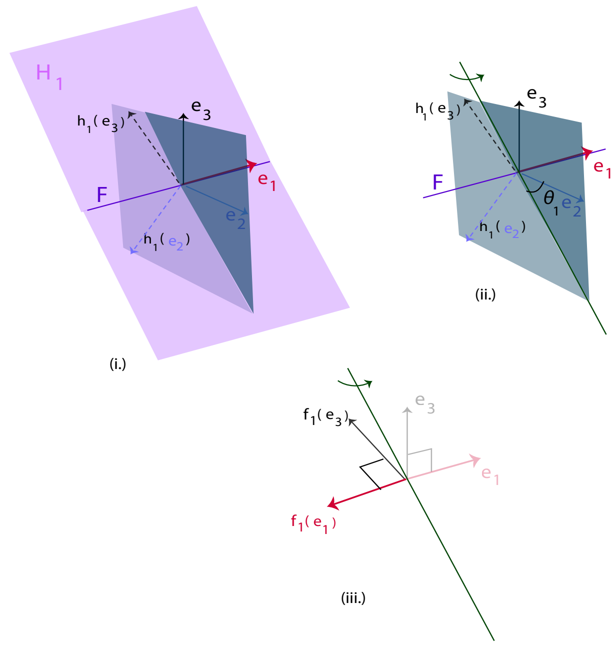


Figure 27.7: The conversion of the hyperplane reflection h_1 into the flip or 180° rotation around the green axis in the e_2e_3-plane. The green axis corresponds to the restriction of the eigenspace associated with eigenvalue 1.

Using Lemma 27.4 and the Cartan–Dieudonné theorem, we obtain the following characterization of rotations when $n \geq 3$.

Theorem 27.5. *Let E be a Euclidean space of dimension $n \geq 3$. Every rotation $f \in \mathbf{SO}(E)$ is the composition of an even number of flips $f = f_{2k} \circ \cdots \circ f_1$, where $2k \leq n$. Furthermore, if $u \neq 0$ is invariant under f (i.e., $u \in \mathrm{Ker}\,(f - \mathrm{id})$), we can pick the last flip f_{2k} such that $u \in F_{2k}^{\perp}$, where F_{2k} is the subspace of dimension $n-2$ determining f_{2k}.*

Proof. By Theorem 27.1, the rotation f can be expressed as an even number of hyperplane reflections $f = s_{2k} \circ s_{2k-1} \circ \cdots \circ s_2 \circ s_1$, with $2k \leq n$. By Lemma 27.4, every composition of two reflections $s_{2i} \circ s_{2i-1}$ can be replaced by the composition of two flips $f_{2i} \circ f_{2i-1}$ $(1 \leq i \leq k)$, which yields $f = f_{2k} \circ \cdots \circ f_1$, where $2k \leq n$.

Assume that $f(u) = u$, with $u \neq 0$. We have already made the remark that in the case where 1 is an eigenvalue of f, the proof of Theorem 27.1 shows that the reflections s_i can be chosen so that $s_i(u) = u$. In particular, if each reflection s_i is a reflection about the hyperplane H_i, we have $u \in H_{2k-1} \cap H_{2k}$. Letting $F = H_{2k-1} \cap H_{2k}$, pick an orthonormal basis $(e_1, \ldots, e_{n-3}, e_{n-2})$ of F, where

$$e_{n-2} = \frac{u}{\|u\|}.$$

The proof of Lemma 27.4 yields two flips f_{2k-1} and f_{2k} such that

$$f_{2k}(e_{n-2}) = -e_{n-2} \quad \text{and} \quad s_{2k} \circ s_{2k-1} = f_{2k} \circ f_{2k-1},$$

since the $(n-2)$th diagonal entry in both matrices is -1, which means that $e_{n-2} \in F_{2k}^{\perp}$, where F_{2k} is the subspace of dimension $n-2$ determining f_{2k}. Since $u = \|u\| e_{n-2}$, we also have $u \in F_{2k}^{\perp}$. $\square$

Remarks:

(1) It is easy to prove that if f is a rotation in $\mathbf{SO}(3)$ and if D is its axis and θ is its angle of rotation, then f is the composition of two flips about lines D_1 and D_2 orthogonal to D and making an angle $\theta/2$.

(2) It is natural to ask what is the minimal number of flips needed to obtain a rotation f (when $n \geq 3$). As for arbitrary isometries, we will prove later that every rotation is the composition of k flips, where

$$k = n - \dim(\mathrm{Ker}\,(f - \mathrm{id})),$$

and that this number is minimal (where $n = \dim(E)$).

We now turn to affine isometries.

27.2 Affine Isometries (Rigid Motions)

In the remaining sections we study affine isometries. First, we characterize the set of fixed points of an affine map. Using this characterization, we prove that every affine isometry f can be written uniquely as

$$f = t \circ g, \quad \text{with} \quad t \circ g = g \circ t,$$

where g is an isometry having a fixed point, and t is a translation by a vector τ such that $\overrightarrow{f}(\tau) = \tau$, and with some additional nice properties (see Theorem 27.10). This is a generalization of a classical result of Chasles about (proper) rigid motions in $\mathbb{R}^3$ (screw motions). We prove a generalization of the Cartan–Dieudonné theorem for the affine isometries: Every isometry in $\mathbf{Is}(n)$ can be written as the composition of at most n affine reflections if it has a fixed point, or else as the composition of at most $n+2$ affine reflections. We also prove that every rigid motion in $\mathbf{SE}(n)$ is the composition of at most n affine flips (for $n \geq 3$). This is somewhat surprising, in view of the previous theorem.

Definition 27.1. Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, a function $f\colon E \to F$ is *an affine isometry (or rigid map)* if it is an affine map and

$$\|\overrightarrow{f(a)f(b)}\| = \|\overrightarrow{ab}\|,$$

for all $a, b \in E$. When $E = F$, an affine isometry $f\colon E \to E$ is also called a *rigid motion*.

Thus, an affine isometry is an affine map that preserves the distance. This is a rather strong requirement. In fact, we will show that for any function $f\colon E \to F$, the assumption that

$$\|\overrightarrow{f(a)f(b)}\| = \|\overrightarrow{ab}\|,$$

for all $a, b \in E$, forces f to be an affine map.

Remark: Sometimes, an affine isometry is defined as a *bijective* affine isometry. When E and F are of finite dimension, the definitions are equivalent.

The following simple lemma is left as an exercise.

Proposition 27.6. *Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, an affine map $f\colon E \to F$ is an affine isometry iff its associated linear map $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{F}$ is an isometry. An affine isometry is a bijection.*

Let us now consider affine isometries $f\colon E \to E$. If $\overrightarrow{f}$ is a rotation, we call f a *proper (or direct) affine isometry*, and if $\overrightarrow{f}$ is an improper linear isometry, we call f an *improper (or skew) affine isometry*. It is easily shown that the set of affine isometries $f\colon E \to E$ forms a group, and those for which $\overrightarrow{f}$ is a rotation is a subgroup. The group of affine isometries, or rigid motions, is a subgroup of the affine group $\mathbf{GA}(E)$, denoted by $\mathbf{Is}(E)$ (or $\mathbf{Is}(n)$ when $E = \mathbb{E}^n$). In Snapper and Troyer [162] the group of rigid motions is denoted by $\mathbf{Mo}(E)$. Since we denote the group of affine bijections as $\mathbf{GA}(E)$, perhaps we should denote the group of affine isometries by $\mathbf{IA}(E)$ (or $\mathbf{EA}(E)$!). The subgroup of $\mathbf{Is}(E)$ consisting of the direct rigid motions is also a subgroup of $\mathbf{SA}(E)$, and it is denoted by $\mathbf{SE}(E)$ (or $\mathbf{SE}(n)$, when $E = \mathbb{E}^n$). The translations are the affine isometries f for which $\overrightarrow{f} = \mathrm{id}$, the identity map on $\overrightarrow{E}$. The following lemma is the counterpart of Lemma 12.12 for isometries between Euclidean vector spaces.

Proposition 27.7. *Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, for every function $f\colon E \to F$, the following properties are equivalent:*

(1) f is an affine map and $\|\overrightarrow{f(a)f(b)}\| = \|\overrightarrow{ab}\|$, for all $a, b \in E$.

(2) $\|\overrightarrow{f(a)f(b)}\| = \|\overrightarrow{ab}\|$, for all $a, b \in E$.

Proof. Obviously, (1) implies (2). In order to prove that (2) implies (1), we proceed as follows. First, we pick some arbitrary point $\Omega \in E$. We define the map $g\colon \overrightarrow{E} \to \overrightarrow{F}$ such that

$$g(u) = \overrightarrow{f(\Omega)f(\Omega + u)}$$

for all $u \in E$. Since

$$f(\Omega) + g(u) = f(\Omega) + \overrightarrow{f(\Omega)f(\Omega + u)} = f(\Omega + u)$$

for all $u \in \vec{E}$, f will be affine if we can show that g is linear, and f will be an affine isometry if we can show that g is a linear isometry.

Observe that

$$\begin{aligned} g(v) - g(u) &= \overrightarrow{f(\Omega)f(\Omega + v)} - \overrightarrow{f(\Omega)f(\Omega + u)} \\ &= \overrightarrow{f(\Omega + u)f(\Omega + v)}. \end{aligned}$$

Then, the hypothesis

$$\|\overrightarrow{f(a)f(b)}\| = \|\overrightarrow{ab}\|$$

for all $a, b \in E$, implies that

$$\|g(v) - g(u)\| = \|\overrightarrow{f(\Omega + u)f(\Omega + v)}\| = \|\overrightarrow{(\Omega + u)(\Omega + v)}\| = \|v - u\|.$$

Thus, g preserves the distance. Also, by definition, we have

$$g(0) = 0.$$

Thus, we can apply Lemma 12.12, which shows that g is indeed a linear isometry, and thus f is an affine isometry. $\square$

In order to understand the structure of affine isometries, it is important to investigate the fixed points of an affine map.

27.3 Fixed Points of Affine Maps

Recall that $E\big(1, \vec{f}\big)$ denotes the eigenspace of the linear map $\vec{f}$ associated with the scalar 1, that is, the subspace consisting of all vectors $u \in \vec{E}$ such that $\vec{f}(u) = u$. Clearly, $\mathrm{Ker}\,\big(\vec{f} - \mathrm{id}\big) = E\big(1, \vec{f}\big)$. Given some origin $\Omega \in E$, since

$$f(a) = f(\Omega + \overrightarrow{\Omega a}) = f(\Omega) + \vec{f}(\overrightarrow{\Omega a}),$$

we have $\overrightarrow{f(\Omega)f(a)} = \vec{f}(\overrightarrow{\Omega a})$, and thus

$$\overrightarrow{\Omega f(a)} = \overrightarrow{\Omega f(\Omega)} + \vec{f}(\overrightarrow{\Omega a}).$$

From the above, we get

$$\overrightarrow{\Omega f(a)} - \overrightarrow{\Omega a} = \overrightarrow{\Omega f(\Omega)} + \vec{f}(\overrightarrow{\Omega a}) - \overrightarrow{\Omega a}.$$

Using this, we show the following lemma, which holds for arbitrary affine spaces of finite dimension and for arbitrary affine maps.

Proposition 27.8. *Let E be any affine space of finite dimension. For every affine map $f\colon E \to E$, let* $\mathrm{Fix}(f) = \{a \in E \mid f(a) = a\}$ *be the set of fixed points of f. The following properties hold:*

(1) If f has some fixed point a, so that $\mathrm{Fix}(f) \neq \emptyset$, *then* $\mathrm{Fix}(f)$ *is an affine subspace of E such that*

$$\mathrm{Fix}(f) = a + E\big(1, \overrightarrow{f}\big) = a + \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big),$$

where $E\big(1, \overrightarrow{f}\big)$ is the eigenspace of the linear map $\overrightarrow{f}$ for the eigenvalue 1.

(2) The affine map f has a unique fixed point iff $E\big(1, \overrightarrow{f}\big) = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) = \{0\}$.

Proof. (1) Since the identity

$$\overrightarrow{\Omega f(b)} - \overrightarrow{\Omega b} = \overrightarrow{\Omega f(\Omega)} + \overrightarrow{f}\,(\overrightarrow{\Omega b}) - \overrightarrow{\Omega b}$$

holds for all $\Omega, b \in E$, if $f(a) = a$, then $\overrightarrow{af(a)} = 0$, and thus, letting $\Omega = a$, for any $b \in E$ we have

$$\overrightarrow{af(b)} - \overrightarrow{ab} = \overrightarrow{af(a)} + \overrightarrow{f}\,(\overrightarrow{ab}) - \overrightarrow{ab} = \overrightarrow{f}\,(\overrightarrow{ab}) - \overrightarrow{ab},$$

and so

$$f(b) = b$$

iff

$$\overrightarrow{af(b)} - \overrightarrow{ab} = 0$$

iff

$$\overrightarrow{f}\,(\overrightarrow{ab}) - \overrightarrow{ab} = 0$$

iff

$$\overrightarrow{ab} \in E\big(1, \overrightarrow{f}\big) = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big),$$

which proves that

$$\mathrm{Fix}(f) = a + E\big(1, \overrightarrow{f}\big) = a + \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big).$$

(2) Again, fix some origin Ω. Some a satisfies $f(a) = a$ iff

$$\overrightarrow{\Omega f(a)} - \overrightarrow{\Omega a} = 0$$

iff

$$\overrightarrow{\Omega f(\Omega)} + \overrightarrow{f}\,(\overrightarrow{\Omega a}) - \overrightarrow{\Omega a} = 0,$$

which can be rewritten as

$$\big(\overrightarrow{f} - \mathrm{id}\big)(\overrightarrow{\Omega a}) = -\overrightarrow{\Omega f(\Omega)}.$$

We have $E\big(1, \overrightarrow{f}\big) = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) = \{0\}$ iff $\overrightarrow{f} - \mathrm{id}$ is injective, and since $\overrightarrow{E}$ has finite dimension, $\overrightarrow{f} - \mathrm{id}$ is also surjective, and thus, there is indeed some $a \in E$ such that

$$\big(\overrightarrow{f} - \mathrm{id}\big)(\overrightarrow{\Omega a}) = -\overrightarrow{\Omega f(\Omega)},$$

and it is unique, since $\overrightarrow{f} - \mathrm{id}$ is injective. Conversely, if f has a unique fixed point, say a, from

$$\big(\overrightarrow{f} - \mathrm{id}\big)(\overrightarrow{\Omega a}) = -\overrightarrow{\Omega f(\Omega)},$$

we have $\big(\overrightarrow{f} - \mathrm{id}\big)(\overrightarrow{\Omega a}) = 0$ iff $f(\Omega) = \Omega$, and since a is the unique fixed point of f, we must have $a = \Omega$, which shows that $\overrightarrow{f} - \mathrm{id}$ is injective. $\square$

Remark: The fact that E has finite dimension is used only to prove (2), and (1) holds in general.

If an affine isometry f leaves some point fixed, we can take such a point Ω as the origin, and then $f(\Omega) = \Omega$ and we can view f as a rotation or an improper orthogonal transformation, depending on the nature of $\overrightarrow{f}$. Note that it is quite possible that $\mathrm{Fix}(f) = \emptyset$. For example, nontrivial translations have no fixed points. A more interesting example is provided by the composition of a plane reflection about a line composed with a a nontrivial translation parallel to this line.

Otherwise, we will see in Theorem 27.10 that every affine isometry is the (commutative) composition of a translation with an affine isometry that always has a fixed point.

27.4 Affine Isometries and Fixed Points

Let E be an affine space. Given any two affine subspaces F, G, if F and G are orthogonal complements in E, which means that $\overrightarrow{F}$ and $\overrightarrow{G}$ are orthogonal subspaces of $\overrightarrow{E}$ such that $\overrightarrow{E} = \overrightarrow{F} \oplus \overrightarrow{G}$, for any point $\Omega \in F$, we define $q\colon E \to \overrightarrow{G}$ such that

$$q(a) = p_{\overrightarrow{G}}(\overrightarrow{\Omega a}).$$

Note that $q(a)$ is independent of the choice of $\Omega \in F$, since we have

$$\overrightarrow{\Omega a} = p_{\overrightarrow{F}}(\overrightarrow{\Omega a}) + p_{\overrightarrow{G}}(\overrightarrow{\Omega a}),$$

and for any $\Omega_1 \in F$, we have

$$\overrightarrow{\Omega_1 a} = \overrightarrow{\Omega_1 \Omega} + p_{\overrightarrow{F}}(\overrightarrow{\Omega a}) + p_{\overrightarrow{G}}(\overrightarrow{\Omega a}),$$

and since $\overrightarrow{\Omega_1 \Omega} \in \overrightarrow{F}$, this shows that

$$p_{\overrightarrow{G}}(\overrightarrow{\Omega_1 a}) = p_{\overrightarrow{G}}(\overrightarrow{\Omega a}).$$

Then the map $g\colon E \to E$ such that $g(a) = a - 2q(a)$, or equivalently

$$\overrightarrow{ag(a)} = -2q(a) = -2p_{\overrightarrow{G}}(\overrightarrow{\Omega a}),$$

does not depend on the choice of $\Omega \in F$. If we identify E to $\overrightarrow{E}$ by choosing any origin Ω in F, we note that g is identified with the symmetry with respect to $\overrightarrow{F}$ and parallel to $\overrightarrow{G}$. Thus, the map g is an affine isometry, and it is called the *affine orthogonal symmetry about F*. Since

$$g(a) = \Omega + \overrightarrow{\Omega a} - 2p_{\overrightarrow{G}}(\overrightarrow{\Omega a})$$

for all $\Omega \in F$ and for all $a \in E$, we note that the linear map $\overrightarrow{g}$ associated with g is the (linear) symmetry about the subspace $\overrightarrow{F}$ (the direction of F), and parallel to $\overrightarrow{G}$ (the direction of G).

Remark: The map $p\colon E \to F$ such that $p(a) = a - q(a)$, or equivalently

$$\overrightarrow{ap(a)} = -q(a) = -p_{\overrightarrow{G}}(\overrightarrow{\Omega a}),$$

is also independent of $\Omega \in F$, and it is called the *affine orthogonal projection onto F*.

The following amusing lemma shows the extra power afforded by affine orthogonal symmetries: Translations are subsumed! Given two parallel affine subspaces F_1 and F_2 in E, letting $\overrightarrow{F}$ be the common direction of F_1 and F_2 and $\overrightarrow{G} = \overrightarrow{F}^{\perp}$ be its orthogonal complement, for any $a \in F_1$, the affine subspace $a + \overrightarrow{G}$ intersects F_2 in a single point b (see Lemma 24.16). We define the *distance between F_1 and F_2* as $\|\overrightarrow{ab}\|$. It is easily seen that the distance between F_1 and F_2 is independent of the choice of a in F_1, and that it is the minimum of $\|\overrightarrow{xy}\|$ for all $x \in F_1$ and all $y \in F_2$.

Proposition 27.9. *Given any affine space E, if $f\colon E \to E$ and $g\colon E \to E$ are affine orthogonal symmetries about parallel affine subspaces F_1 and F_2, then $g \circ f$ is a translation defined by the vector $2\overrightarrow{ab}$, where $\overrightarrow{ab}$ is any vector perpendicular to the common direction $\overrightarrow{F}$ of F_1 and F_2 such that $\|\overrightarrow{ab}\|$ is the distance between F_1 and F_2, with $a \in F_1$ and $b \in F_2$. Conversely, every translation by a vector τ is obtained as the composition of two affine orthogonal symmetries about parallel affine subspaces F_1 and F_2 whose common direction is orthogonal to $\tau = \overrightarrow{ab}$, for some $a \in F_1$ and some $b \in F_2$ such that the distance between F_1 and F_2 is $\|\overrightarrow{ab}\|/2$.*

Proof. We observed earlier that the linear maps $\overrightarrow{f}$ and $\overrightarrow{g}$ associated with f and g are the linear reflections about the directions of F_1 and F_2. However, F_1 and F_2 have the same direction, and so $\overrightarrow{f} = \overrightarrow{g}$. Since $\overrightarrow{g \circ f} = \overrightarrow{g} \circ \overrightarrow{f}$ and since $\overrightarrow{f} \circ \overrightarrow{g} = \overrightarrow{f} \circ \overrightarrow{f} = \mathrm{id}$, because every reflection is an involution, we have $\overrightarrow{g \circ f} = \mathrm{id}$, proving that $g \circ f$ is a translation. If we pick $a \in F_1$, then $g \circ f(a) = g(a)$, the affine reflection of $a \in F_1$ about F_2, and it is easily checked that $g \circ f$ is the translation by the vector $\tau = \overrightarrow{ag(a)}$ whose norm is twice the distance between F_1 and F_2. The second part of the lemma is left as an easy exercise. $\square$

We conclude our quick study of affine isometries by proving a result that plays a major role in characterizing the affine isometries. This result may be viewed as a generalization of Chasles's theorem about the direct rigid motions in $\mathbb{E}^3$.

Theorem 27.10. *Let E be a Euclidean affine space of finite dimension n. For every affine isometry $f\colon E \to E$, there is a unique affine isometry $g\colon E \to E$ and a unique translation $t = t_\tau$, with $\overrightarrow{f}(\tau) = \tau$ (i.e., $\tau \in \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$), such that the set $\mathrm{Fix}(g) = \{a \in E \mid g(a) = a\}$ of fixed points of g is a nonempty affine subspace of E of direction*

$$\overrightarrow{G} = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) = E\big(1, \overrightarrow{f}\big),$$

and such that

$$f = t \circ g \quad \textit{and} \quad t \circ g = g \circ t.$$

Furthermore, we have the following additional properties:

(a) $f = g$ and $\tau = 0$ iff f has some fixed point, i.e., iff $\mathrm{Fix}(f) \neq \emptyset$.

(b) If f has no fixed points, i.e., $\mathrm{Fix}(f) = \emptyset$, then $\dim\big(\mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)\big) \geq 1$.

Proof. The proof rests on the following two key facts:

(1) If we can find some $x \in E$ such that $\overrightarrow{xf(x)} = \tau$ belongs to $\mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$, we get the existence of g and τ.

(2) $\overrightarrow{E} = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) \oplus \mathrm{Im}\,\big(\overrightarrow{f} - \mathrm{id}\big)$, and the spaces $\mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$ and $\mathrm{Im}\,\big(\overrightarrow{f} - \mathrm{id}\big)$ are orthogonal. This implies the uniqueness of g and τ.

First, we prove that for every isometry $h\colon \overrightarrow{E} \to \overrightarrow{E}$, $\mathrm{Ker}\,(h - \mathrm{id})$ and $\mathrm{Im}\,(h - \mathrm{id})$ are orthogonal and that

$$\overrightarrow{E} = \mathrm{Ker}\,(h - \mathrm{id}) \oplus \mathrm{Im}\,(h - \mathrm{id}).$$

Recall that

$$\dim\big(\overrightarrow{E}\big) = \dim(\mathrm{Ker}\,\varphi) + \dim(\mathrm{Im}\,\varphi),$$

for any linear map $\varphi\colon \overrightarrow{E} \to \overrightarrow{E}$; see Theorem 6.16. To show that we have a direct sum, we prove orthogonality. Let $u \in \mathrm{Ker}\,(h - \mathrm{id})$, so that $h(u) = u$, let $v \in \overrightarrow{E}$, and compute

$$u \cdot (h(v) - v) = u \cdot h(v) - u \cdot v = h(u) \cdot h(v) - u \cdot v = 0,$$

since $h(u) = u$ and h is an isometry.

Next, assume that there is some $x \in E$ such that $\overrightarrow{xf(x)} = \tau$ belongs to the space $\mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$. If we define $g\colon E \to E$ such that

$$g = t_{(-\tau)} \circ f,$$

we have

$$g(x) = f(x) - \tau = x,$$

since $\overrightarrow{xf(x)} = \tau$ is equivalent to $x = f(x) - \tau$. As a composition of affine isometries, g is an affine isometry, x is a fixed point of g, and since $\tau \in \operatorname{Ker}\big(\overrightarrow{f} - \mathrm{id}\big)$, we have

$$\overrightarrow{f}(\tau) = \tau,$$

and since

$$g(b) = f(b) - \tau$$

for all $b \in E$, we have $\overrightarrow{g} = \overrightarrow{f}$. Since g has some fixed point x, by Lemma 27.8, $\mathrm{Fix}(g)$ is an affine subspace of E with direction $\operatorname{Ker}\big(\overrightarrow{g} - \mathrm{id}\big) = \operatorname{Ker}\big(\overrightarrow{f} - \mathrm{id}\big)$. We also have $f(b) = g(b) + \tau$ for all $b \in E$, and thus

$$(g \circ t_\tau)(b) = g(b + \tau) = g(b) + \overrightarrow{g}(\tau) = g(b) + \overrightarrow{f}(\tau) = g(b) + \tau = f(b),$$

and

$$(t_\tau \circ g)(b) = g(b) + \tau = f(b),$$

which proves that $t \circ g = g \circ t$.

To prove the existence of x as above, pick any arbitrary point $a \in E$. Since

$$\overrightarrow{E} = \operatorname{Ker}\big(\overrightarrow{f} - \mathrm{id}\big) \oplus \operatorname{Im}\big(\overrightarrow{f} - \mathrm{id}\big),$$

there is a unique vector $\tau \in \operatorname{Ker}\big(\overrightarrow{f} - \mathrm{id}\big)$ and some $v \in \overrightarrow{E}$ such that

$$\overrightarrow{af(a)} = \tau + \overrightarrow{f}(v) - v.$$

For any $x \in E$, since we also have

$$\overrightarrow{xf(x)} = \overrightarrow{xa} + \overrightarrow{af(a)} + \overrightarrow{f(a)f(x)} = \overrightarrow{xa} + \overrightarrow{af(a)} + \overrightarrow{f}(\overrightarrow{ax}),$$

we get

$$\overrightarrow{xf(x)} = \overrightarrow{xa} + \tau + \overrightarrow{f}(v) - v + \overrightarrow{f}(\overrightarrow{ax}),$$

which can be rewritten as

$$\overrightarrow{xf(x)} = \tau + \big(\overrightarrow{f} - \mathrm{id}\big)(v + \overrightarrow{ax}).$$

If we let $\overrightarrow{ax} = -v$, that is, $x = a - v$, we get

$$\overrightarrow{xf(x)} = \tau,$$

with $\tau \in \operatorname{Ker}\big(\overrightarrow{f} - \mathrm{id}\big)$.

Finally, we show that τ is unique. Assume two decompositions (g_1, τ_1) and (g_2, τ_2). Since $\overrightarrow{f} = \overrightarrow{g_1}$, we have $\mathrm{Ker}\,(\overrightarrow{g_1} - \mathrm{id}) = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$. Since g_1 has some fixed point b, we get

$$f(b) = g_1(b) + \tau_1 = b + \tau_1,$$

that is, $\overrightarrow{bf(b)} = \tau_1$, and $\overrightarrow{bf(b)} \in \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$, since $\tau_1 \in \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$. Similarly, for some fixed point c of g_2, we get $\overrightarrow{cf(c)} = \tau_2$ and $\overrightarrow{cf(c)} \in \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big)$. Then we have

$$\tau_2 - \tau_1 = \overrightarrow{cf(c)} - \overrightarrow{bf(b)} = \overrightarrow{cb} - \overrightarrow{f(c)f(b)} = \overrightarrow{cb} - \overrightarrow{f}(\overrightarrow{cb}),$$

which shows that

$$\tau_2 - \tau_1 \in \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) \cap \mathrm{Im}\,\big(\overrightarrow{f} - \mathrm{id}\big),$$

and thus that $\tau_2 = \tau_1$, since we have shown that

$$\overrightarrow{E} = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) \oplus \mathrm{Im}\,\big(\overrightarrow{f} - \mathrm{id}\big).$$

The fact that (a) holds is a consequence of the uniqueness of g and τ, since f and 0 clearly satisfy the required conditions. That (b) holds follows from Lemma 27.8 (2), since the affine map f has a unique fixed point iff $E\big(1, \overrightarrow{f}\big) = \mathrm{Ker}\,\big(\overrightarrow{f} - \mathrm{id}\big) = \{0\}$. $\square$

The determination of x is illustrated in Figure 27.8.

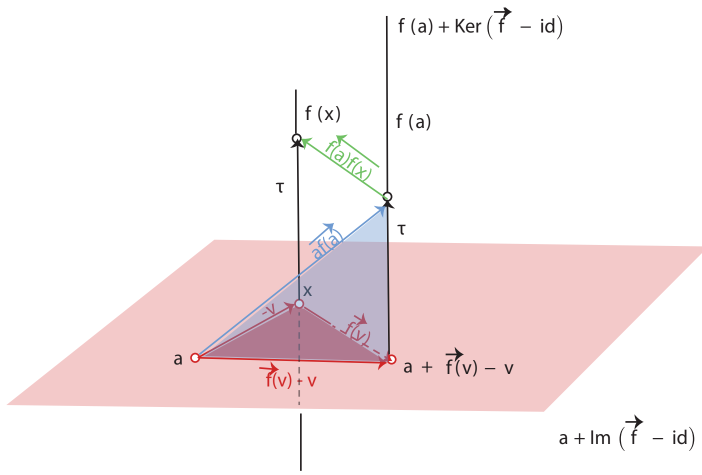


Figure 27.8: Affine rigid motion as $f = t \circ g$, where g has some fixed point x.

Remarks:

(1) Note that $\mathrm{Ker}\left(\overrightarrow{f} - \mathrm{id}\right) = \{0\}$ iff $\mathrm{Fix}(g)$ consists of a single element, which is the unique fixed point of f. However, even if f is not a translation, f may not have any fixed points. For example, this happens when E is the affine Euclidean plane and f is the composition of a reflection about a line composed with a nontrivial translation parallel to this line.

(2) The fact that E has finite dimension is used only to prove (b).

(3) It is easily checked that $\mathrm{Fix}(g)$ consists of the set of points x such that $\|\overrightarrow{xf(x)}\|$ is minimal.

In the affine Euclidean plane it is easy to see that the affine isometries (besides the identity) are classified as follows. An affine isometry f that has a fixed point is a rotation if it is a direct isometry; otherwise, it is an affine reflection about a line. If f has no fixed point, then it is either a nontrivial translation or the composition of an affine reflection about a line with a nontrivial translation parallel to this line.

In an affine space of dimension 3 it is easy to see that the affine isometries (besides the identity) are classified as follows. There are three kinds of affine isometries that have a fixed point. A proper affine isometry with a fixed point is a rotation around a line D (its set of fixed points), as illustrated in Figure 27.9.

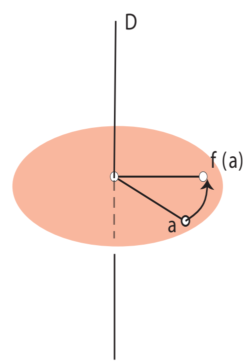


Figure 27.9: 3D proper affine rigid motion with line D of fixed points (rotation).

An improper affine isometry with a fixed point is either an affine reflection about a plane H (the set of fixed points) or the composition of a rotation followed by an affine reflection about a plane H orthogonal to the axis of rotation D, as illustrated in Figures 27.10 and 27.11. In the second case, there is a single fixed point $O = D \cap H$.

There are three types of affine isometries with no fixed point. The first kind is a nontrivial translation. The second kind is the composition of a rotation followed by a nontrivial translation parallel to the axis of rotation D. Such an affine rigid motion is proper, and is called a *screw motion*. A screw motion is illustrated in Figure 27.12.

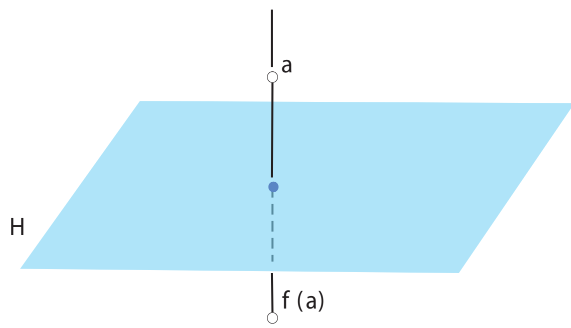


Figure 27.10: 3D improper affine rigid motion with a plane H of fixed points (reflection).

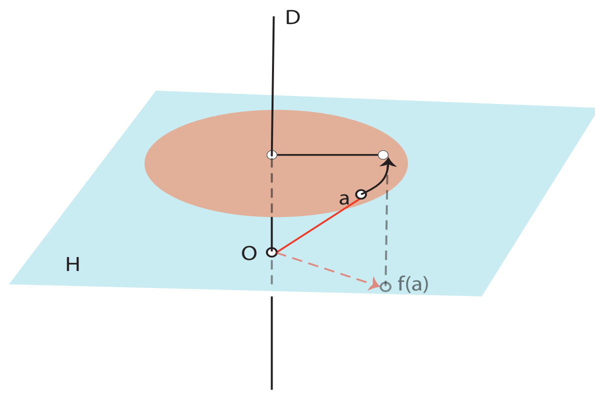


Figure 27.11: $3D$ improper affine rigid motion with a unique fixed point.

The third kind is the composition of an affine reflection about a plane followed by a nontrivial translation by a vector parallel to the direction of the plane of the reflection, as illustrated in Figure 27.13.

This last transformation is an improper affine isometry.

27.5 The Cartan–Dieudonné Theorem for Affine Isometries

The Cartan–Dieudonné theorem also holds for affine isometries, with a small twist due to translations. The reader is referred to Berger [11], Snapper and Troyer [162], or Tisseron [175] for a detailed treatment of the Cartan–Dieudonné theorem and its variants.

Theorem 27.11. *Let E be an affine Euclidean space of dimension $n \geq 1$. Every affine isometry $f \in \mathbf{Is}(E)$ that has a fixed point and is not the identity is the composition of at most n affine reflections. Every affine isometry $f \in \mathbf{Is}(E)$ that has no fixed point is the*

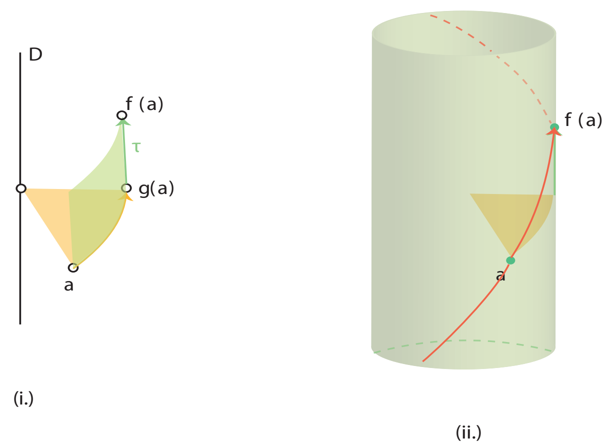


Figure 27.12: 3D proper affine rigid motion with no fixed point (screw motion). The second illustration demonstrates that a screw motion produces a helix path along the surface of a cylinder.

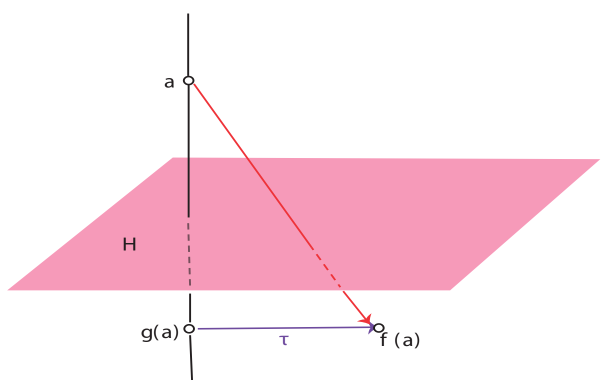


Figure 27.13: 3D improper affine rigid motion with no fixed points.

composition of at most $n+2$ affine reflections. When $n \geq 2$, the identity is the composition of any reflection with itself.

Proof. First, we use Theorem 27.10. If f has a fixed point Ω, we choose Ω as an origin and work in the vector space E_Ω. Since f behaves as a linear isometry, the result follows from Theorem 27.1. More specifically, we can write $\overrightarrow{f} = \overrightarrow{s_k} \circ \cdots \circ \overrightarrow{s_1}$ for $k \leq n$ hyperplane reflections $\overrightarrow{s_i}$. We define the affine reflections s_i such that

$$s_i(a) = \Omega + \overrightarrow{s_i}(\overrightarrow{\Omega a})$$

for all $a \in E$, and we note that $f = s_k \circ \cdots \circ s_1$, since

$$f(a) = \Omega + \overrightarrow{s_k} \circ \cdots \circ \overrightarrow{s_1}(\overrightarrow{\Omega a})$$

for all $a \in E$. If f has no fixed point, then $f = t \circ g$ for some affine isometry g that has a fixed point Ω and some translation $t = t_\tau$, with $\overrightarrow{f}(\tau) = \tau$. By the argument just given, we can write $g = s_k \circ \cdots \circ s_1$ for some affine reflections (at most n). However, by Lemma 27.9, the translation $t = t_\tau$ can be achieved by two affine reflections about parallel hyperplanes, and thus $f = s_{k+2} \circ \cdots \circ s_1$, for some affine reflections (at most $n+2$). ☐

When $n \geq 3$, we can also characterize the affine isometries in $\mathbf{SE}(n)$ in terms of affine flips. Remarkably, not only we can do without translations, but we can even bound the number of affine flips by n.

Theorem 27.12. *Let E be a Euclidean affine space of dimension $n \geq 3$. Every affine rigid motion $f \in \mathbf{SE}(E)$ is the composition of an even number of affine flips $f = f_{2k} \circ \cdots \circ f_1$, where $2k \leq n$.*

Proof. As in the proof of Theorem 27.11, we distinguish between the two cases where f has some fixed point or not. If f has a fixed point Ω, we apply Theorem 27.5. More specifically, we can write $\overrightarrow{f} = \overrightarrow{f_{2k}} \circ \cdots \circ \overrightarrow{f_1}$ for some flips $\overrightarrow{f_i}$. We define the affine flips f_i such that

$$f_i(a) = \Omega + \overrightarrow{f_i}(\overrightarrow{\Omega a})$$

for all $a \in E$, and we note that $f = f_{2k} \circ \cdots \circ f_1$, since

$$f(a) = \Omega + \overrightarrow{f_{2k}} \circ \cdots \circ \overrightarrow{f_1}(\overrightarrow{\Omega a})$$

for all $a \in E$.

If f does not have a fixed point, as in the proof of Theorem 27.11, we get

$$f = t_\tau \circ f_{2k} \circ \cdots \circ f_1,$$

for some affine flips f_i. We need to get rid of the translation. However, $\overrightarrow{f}(\tau) = \tau$, and by the second part of Theorem 27.5, we can assume that $\tau \in \overrightarrow{F_{2k}}^{\perp}$, where $\overrightarrow{F_{2k}}$ is the direction of the affine subspace defining the affine flip f_{2k}. Finally, appealing to Lemma 27.9, since $\tau \in \overrightarrow{F_{2k}}^{\perp}$, the translation t_τ can be expressed as the composition $f'_{2k} \circ f'_{2k-1}$ of two affine flips f'_{2k-1} and f'_{2k} about the two parallel subspaces $\Omega + \overrightarrow{F_{2k}}$ and $\Omega + \tau/2 + \overrightarrow{F_{2k}}$, whose distance is $\|\tau\|/2$. However, since f'_{2k-1} and f_{2k} are both the identity on $\Omega + \overrightarrow{F_{2k}}$, we must have $f'_{2k-1} = f_{2k}$, and thus

$$\begin{aligned} f &= t_\tau \circ f_{2k} \circ f_{2k-1} \circ \cdots \circ f_1 \\ &= f'_{2k} \circ f'_{2k-1} \circ f_{2k} \circ f_{2k-1} \circ \cdots \circ f_1 \\ &= f'_{2k} \circ f_{2k-1} \circ \cdots \circ f_1, \end{aligned}$$

since $f'_{2k-1} = f_{2k}$ and $f'_{2k-1} \circ f_{2k} = f_{2k} \circ f_{2k} = \mathrm{id}$, since f_{2k} is an affine symmetry. $\square$

Remark: It is easy to prove that if f is a screw motion in $\mathbf{SE}(3)$, D its axis, θ is its angle of rotation, and τ the translation along the direction of D, then f is the composition of two affine flips about lines D_1 and D_2 orthogonal to D, at a distance $\|\tau\|/2$ and making an angle $\theta/2$.

Chapter 28

Isometries of Hermitian Spaces

28.1 The Cartan–Dieudonné Theorem, Hermitian Case

The Cartan-Dieudonné theorem can be generalized (Theorem 28.2), but this requires allowing new types of hyperplane reflections that we call Hermitian reflections. After doing so, every isometry in $\mathbf{U}(n)$ can always be written as a composition of at most n Hermitian reflections (for $n \geq 2$). Better yet, every rotation in $\mathbf{SU}(n)$ can be expressed as the composition of at most $2n-2$ (standard) hyperplane reflections! This implies that every unitary transformation in $\mathbf{U}(n)$ is the composition of at most $2n-1$ isometries, with at most one Hermitian reflection, the other isometries being (standard) hyperplane reflections. The crucial Proposition 13.2 is false as is, and needs to be amended. The QR-decomposition of arbitrary complex matrices in terms of Householder matrices can also be generalized, using a trick.

In order to generalize the Cartan–Dieudonné theorem and the QR-decomposition in terms of Householder transformations, we need to introduce new kinds of hyperplane reflections. This is not really surprising, since in the Hermitian case, there are improper isometries whose determinant can be any unit complex number. Hyperplane reflections are generalized as follows.

Definition 28.1. Let E be a Hermitian space of finite dimension. For any hyperplane H, for any nonnull vector w orthogonal to H, so that $E = H \oplus G$, where $G = \mathbb{C}w$, a *Hermitian reflection about H of angle θ* is a linear map of the form $\rho_{H,\theta}\colon E \to E$, defined such that

$$\rho_{H,\theta}(u) = p_H(u) + e^{i\theta} p_G(u),$$

for any unit complex number $e^{i\theta} \neq 1$ (i.e. $\theta \neq k2\pi$). For any nonzero vector $w \in E$, we denote by $\rho_{w,\theta}$ the Hermitian reflection given by $\rho_{H,\theta}$, where H is the hyperplane orthogonal to w.

Since $u = p_H(u) + p_G(u)$, the Hermitian reflection $\rho_{w,\theta}$ is also expressed as

$$\rho_{w,\theta}(u) = u + (e^{i\theta} - 1)p_G(u),$$

or as

$$\rho_{w,\theta}(u) = u + (e^{i\theta} - 1)\,\frac{(u \cdot w)}{\|w\|^2}\,w.$$

Note that the case of a standard hyperplane reflection is obtained when $e^{i\theta} = -1$, i.e., $\theta = \pi$.

We leave as an easy exercise to check that $\rho_{w,\theta}$ is indeed an isometry, and that the inverse of $\rho_{w,\theta}$ is $\rho_{w,-\theta}$. If we pick an orthonormal basis $(e_1, \ldots, e_n)$ such that $(e_1, \ldots, e_{n-1})$ is an orthonormal basis of H, the matrix of $\rho_{w,\theta}$ is

$$\begin{pmatrix} I_{n-1} & 0 \\ 0 & e^{i\theta} \end{pmatrix}$$

We now come to the main surprise. Given any two distinct vectors u and v such that $\|u\| = \|v\|$, there isn't always a hyperplane reflection mapping u to v, but this can be done using two Hermitian reflections!

Proposition 28.1. *Let E be any nontrivial Hermitian space.*

(1) For any two vectors $u, v \in E$ such that $u \neq v$ and $\|u\| = \|v\|$, if $u \cdot v = e^{i\theta}|u \cdot v|$, then the (usual) reflection s about the hyperplane orthogonal to the vector $v - e^{-i\theta}u$ is such that $s(u) = e^{i\theta}v$.

(2) For any nonnull vector $v \in E$, for any unit complex number $e^{i\theta} \neq 1$, there is a Hermitian reflection $\rho_{v,\theta}$ such that

$$\rho_{v,\theta}(v) = e^{i\theta}v.$$

As a consequence, for u and v as in (1), we have $\rho_{v,-\theta} \circ s(u) = v$.

Proof. (1) Consider the (usual) reflection about the hyperplane orthogonal to $w = v - e^{-i\theta}u$. We have

$$s(u) = u - 2\,\frac{(u \cdot (v - e^{-i\theta}u))}{\|v - e^{-i\theta}u\|^2}\,(v - e^{-i\theta}u).$$

We need to compute

$$-2u \cdot (v - e^{-i\theta}u) \quad \text{and} \quad (v - e^{-i\theta}u) \cdot (v - e^{-i\theta}u).$$

Since $u \cdot v = e^{i\theta}|u \cdot v|$, we have

$$e^{-i\theta}u \cdot v = |u \cdot v| \quad \text{and} \quad e^{i\theta}v \cdot u = |u \cdot v|.$$

Using the above and the fact that $\|u\| = \|v\|$, we get

$$\begin{aligned} -2u \cdot (v - e^{-i\theta}u) &= 2e^{i\theta}\,\|u\|^2 - 2u \cdot v, \\ &= 2e^{i\theta}(\|u\|^2 - |u \cdot v|), \end{aligned}$$

and

$$\begin{aligned}(v - e^{-i\theta}u) \cdot (v - e^{-i\theta}u) &= \|v\|^2 + \|u\|^2 - e^{-i\theta}u \cdot v - e^{i\theta}v \cdot u, \\ &= 2(\|u\|^2 - |u \cdot v|),\end{aligned}$$

and thus,

$$-2\,\frac{(u \cdot (v - e^{-i\theta}u))}{\|(v - e^{-i\theta}u)\|^2}\,(v - e^{-i\theta}u) = e^{i\theta}(v - e^{-i\theta}u).$$

But then,

$$s(u) = u + e^{i\theta}(v - e^{-i\theta}u) = u + e^{i\theta}v - u = e^{i\theta}v,$$

and $s(u) = e^{i\theta}v$, as claimed.

(2) This part is easier. Consider the Hermitian reflection

$$\rho_{v,\theta}(u) = u + (e^{i\theta} - 1)\,\frac{(u \cdot v)}{\|v\|^2}\,v.$$

We have

$$\begin{aligned}\rho_{v,\theta}(v) &= v + (e^{i\theta} - 1)\,\frac{(v \cdot v)}{\|v\|^2}\,v, \\ &= v + (e^{i\theta} - 1)v, \\ &= e^{i\theta}v.\end{aligned}$$

Thus, $\rho_{v,\theta}(v) = e^{i\theta}v$. Since $\rho_{v,\theta}$ is linear, changing the argument v to $e^{i\theta}v$, we get

$$\rho_{v,-\theta}(e^{i\theta}v) = v,$$

and thus, $\rho_{v,-\theta} \circ s(u) = v$. $\square$

Remarks:

(1) If we use the vector $v + e^{-i\theta}u$ instead of $v - e^{-i\theta}u$, we get $s(u) = -e^{i\theta}v$.

(2) Certain authors, such as Kincaid and Cheney [102] and Ciarlet [41], use the vector $u + e^{i\theta}v$ instead of our vector $v + e^{-i\theta}u$. The effect of this choice is that they also get $s(u) = -e^{i\theta}v$.

(3) If $v = \|u\|\, e_1$, where e_1 is a basis vector, $u \cdot e_1 = a_1$, where a_1 is just the coefficient of u over the basis vector e_1. Then, since $u \cdot e_1 = e^{i\theta}|a_1|$, the choice of the plus sign in the vector $\|u\|\, e_1 + e^{-i\theta}u$ has the effect that the coefficient of this vector over e_1 is $\|u\| + |a_1|$, and no cancellations takes place, which is preferable for numerical stability (we need to divide by the square norm of this vector).

The last part of Proposition 28.1 shows that the Cartan–Dieudonné is salvaged, since we can send u to v by a sequence of two Hermitian reflections when $u \neq v$ and $\|u\| = \|v\|$, and since the inverse of a Hermitian reflection is a Hermitian reflection. Actually, because we are over the complex field, a linear map always have (complex) eigenvalues, and we can get a slightly improved result.

Theorem 28.2. *Let E be a Hermitian space of dimension $n \geq 1$. Every isometry $f \in \mathbf{U}(E)$ is the composition $f = \rho_n \circ \rho_{n-1} \circ \cdots \circ \rho_1$ of n isometries ρ_j, where each ρ_j is either the identity or a Hermitian reflection (possibly a standard hyperplane reflection). When $n \geq 2$, the identity is the composition of any hyperplane reflection with itself.*

Proof. We prove by induction on n that there is an orthonormal basis of eigenvectors $(u_1, \ldots, u_n)$ of f such that
$$f(u_j) = e^{i\theta_j} u_j,$$
where $e^{i\theta_j}$ is an eigenvalue associated with u_j, for all j, $1 \leq j \leq n$.

When $n = 1$, every isometry $f \in \mathbf{U}(E)$ is either the identity or a Hermitian reflection ρ_θ, since for any nonnull vector u, we have $f(u) = e^{i\theta} u$ for some θ. We let u_1 be any nonnull unit vector.

Let us now consider the case where $n \geq 2$. Since $\mathbb{C}$ is algebraically closed, the characteristic polynomial $\det(f - \lambda \mathrm{id})$ of f has n complex roots which must be the form $e^{i\theta}$, since they have absolute value 1. Pick any such eigenvalue $e^{i\theta_1}$, and pick any eigenvector $u_1 \neq 0$ of f for $e^{i\theta_1}$ of unit length. If $F = \mathbb{C}u_1$ is the subspace spanned by u_1, we have $f(F) = F$, since $f(u_1) = e^{i\theta_1} u_1$. Since $f(F) = F$ and f is an isometry, it is easy to see that $f(F^\perp) \subseteq F^\perp$, and by Proposition 14.13, we have $E = F \oplus F^\perp$. Furthermore, it is obvious that the restriction of f to $F^\perp$ is unitary. Since $\dim(F^\perp) = n - 1$, we can apply the induction hypothesis to $F^\perp$, and we get an orthonormal basis of eigenvectors $(u_2, \ldots, u_n)$ for $F^\perp$ such that
$$f(u_j) = e^{i\theta_j} u_j,$$
where $e^{i\theta_j}$ is an eigenvalue associated with u_j, for all j, $2 \leq j \leq n$ Since $E = F \oplus F^\perp$ and $F = \mathbb{C}u_1$, the claim is proved. But then, if ρ_j is the Hermitian reflection about the hyperplane H_j orthogonal to u_j and of angle θ_j, it is obvious that
$$f = \rho_{\theta_n} \circ \cdots \circ \rho_{\theta_1}.$$
When $n \geq 2$, we have $\mathrm{id} = s \circ s$ for every reflection s. $\square$

Remarks:

(1) Any isometry $f \in \mathbf{U}(n)$ can be express as $f = \rho_\theta \circ g$, where $g \in \mathbf{SU}(n)$ is a rotation, and ρ_θ is a Hermitian reflection. Indeed, by the above theorem, with respect to the basis $(u_1, \ldots, u_n)$, $\det(f) = e^{i(\theta_1 + \cdots + \theta_n)}$, and letting $\theta = \theta_1 + \cdots + \theta_n$ and ρ_θ be the Hermitian

reflection about the hyperplane orthogonal to u_1 and of angle θ, since $\rho_\theta \circ \rho_{-\theta} = \mathrm{id}$, we have

$$f = (\rho_\theta \circ \rho_{-\theta}) \circ f = \rho_\theta \circ (\rho_{-\theta} \circ f).$$

Letting $g = \rho_{-\theta} \circ f$, it is obvious that $\det(g) = 1$. As a consequence, there is a bijection between $S^1 \times \mathbf{SU}(n)$ and $\mathbf{U}(n)$, where S^1 is the unit circle (which corresponds to the group of complex numbers $e^{i\theta}$ of unit length). In fact, it is a homeomorphism.

(2) We abandoned the style of proof used in theorem 27.1, because in the Hermitian case, eigenvalues and eigenvectors always exist, and the proof is simpler that way (in the real case, an isometry may not have any real eigenvalues!). The sacrifice is that the theorem yields no information on the number of (standard) hyperplane reflections. We shall rectify this situation shortly.

We will now reveal the beautiful trick (found in Mneimné and Testard [127]) that allows us to prove that every rotation in $\mathbf{SU}(n)$ is the composition of at most $2n-2$ (standard) hyperplane reflections. For what follows, it is more convenient to denote a standard reflection about the hyperplane H as h_u (it is trivial that these do not depend on the choice of u in $H^\perp$). Then, given any two distinct orthogonal vectors u, v such that $\|u\| = \|v\|$, consider the composition $\rho_{v,-\theta} \circ \rho_{u,\theta}$. The trick is that this composition can be expressed as two standard hyperplane reflections! This wonderful fact is proved in the next Proposition.

Proposition 28.3. *Let E be a nontrivial Hermitian space. For any two distinct orthogonal vectors u, v such that $\|u\| = \|v\|$, we have*

$$\rho_{v,-\theta} \circ \rho_{u,\theta} = h_{v-u} \circ h_{v-e^{-i\theta}u} = h_{u+v} \circ h_{u+e^{i\theta}v}.$$

Proof. Since u and v are orthogonal, each one is in the hyperplane orthogonal to the other, and thus,

$$\begin{aligned}
\rho_{u,\theta}(u) &= e^{i\theta}u,\\
\rho_{u,\theta}(v) &= v,\\
\rho_{v,-\theta}(u) &= u,\\
\rho_{v,-\theta}(v) &= e^{-i\theta}v,\\
h_{v-u}(u) &= v,\\
h_{v-u}(v) &= u,\\
h_{v-e^{-i\theta}u}(u) &= e^{i\theta}v,\\
h_{v-e^{-i\theta}u}(v) &= e^{-i\theta}u.
\end{aligned}$$

Consequently, using linearity,

$$\begin{aligned}
\rho_{v,-\theta} \circ \rho_{u,\theta}(u) &= e^{i\theta}u,\\
\rho_{v,-\theta} \circ \rho_{u,\theta}(v) &= e^{-i\theta}v,\\
h_{v-u} \circ h_{v-e^{-i\theta}u}(u) &= e^{i\theta}u,\\
h_{v-u} \circ h_{v-e^{-i\theta}u}(v) &= e^{-i\theta}v,
\end{aligned}$$

and since both $\rho_{v,-\theta}\circ\rho_{u,\theta}$ and $h_{v-u}\circ h_{v-e^{-i\theta}u}$ are the identity on the orthogonal complement of $\{u,v\}$, they are equal. Since we also have

$$\begin{aligned} h_{u+v}(u) &= -v, \\ h_{u+v}(v) &= -u, \\ h_{u+e^{i\theta}v}(u) &= -e^{i\theta}v, \\ h_{u+e^{i\theta}v}(v) &= -e^{-i\theta}u, \end{aligned}$$

it is immediately verified that

$$h_{v-u}\circ h_{v-e^{-i\theta}u} = h_{u+v}\circ h_{u+e^{i\theta}v}.$$

□

We will use Proposition 28.3 as follows.

Proposition 28.4. *Let E be a nontrivial Hermitian space, and let $(u_1,\ldots,u_n)$ be some orthonormal basis for E. For any $\theta_1,\ldots,\theta_n$ such that $\theta_1+\cdots+\theta_n=0$, if $f\in\mathbf{U}(n)$ is the isometry defined such that*

$$f(u_j) = e^{i\theta_j}u_j,$$

for all j, $1\leq j\leq n$, then f is a rotation ($f\in\mathbf{SU}(n)$), and

$$\begin{aligned} f &= \rho_{u_n,\theta_n}\circ\cdots\circ\rho_{u_1,\theta_1} \\ &= \rho_{u_n,-(\theta_1+\cdots+\theta_{n-1})}\circ\rho_{u_{n-1},\theta_1+\cdots+\theta_{n-1}}\circ\cdots\circ\rho_{u_2,-\theta_1}\circ\rho_{u_1,\theta_1} \\ &= h_{u_n-u_{n-1}}\circ h_{u_n-e^{-i(\theta_1+\cdots+\theta_{n-1})}u_{n-1}}\circ\cdots\circ h_{u_2-u_1}\circ h_{u_2-e^{-i\theta_1}u_1} \\ &= h_{u_{n-1}+u_n}\circ h_{u_{n-1}+e^{i(\theta_1+\cdots+\theta_{n-1})}u_n}\circ\cdots\circ h_{u_1+u_2}\circ h_{u_1+e^{i\theta_1}u_2}. \end{aligned}$$

Proof. It is obvious from the definitions that

$$f = \rho_{u_n,\theta_n}\circ\cdots\circ\rho_{u_1,\theta_1},$$

and since the determinant of f is

$$D(f) = e^{i\theta_1}\cdots e^{i\theta_n} = e^{i(\theta_1+\cdots+\theta_n)}$$

and $\theta_1+\cdots+\theta_n=0$, we have $D(f)=e^0=1$, and f is a rotation. Letting

$$f_k = \rho_{u_k,-(\theta_1+\cdots+\theta_{k-1})}\circ\rho_{u_{k-1},\theta_1+\cdots+\theta_{k-1}}\circ\cdots\circ\rho_{u_3,-(\theta_1+\theta_2)}\circ\rho_{u_2,\theta_1+\theta_2}\circ\rho_{u_2,-\theta_1}\circ\rho_{u_1,\theta_1},$$

we prove by induction on k, $2\leq k\leq n$, that

$$f_k(u_j) = \begin{cases} e^{i\theta_j}u_j & \text{if } 1\leq j\leq k-1, \\ e^{-i(\theta_1+\cdots+\theta_{k-1})}u_k & \text{if } j=k, \text{ and} \\ u_j & \text{if } k+1\leq j\leq n. \end{cases}$$

The base case was treated in Proposition 28.3. Now, the proof of Proposition 28.3 also showed that

$$\begin{aligned}\rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\,\theta_1+\cdots+\theta_k}(u_k) &= e^{i(\theta_1+\cdots+\theta_k)} u_k,\\ \rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\,\theta_1+\cdots+\theta_k}(u_{k+1}) &= e^{-i(\theta_1+\cdots+\theta_k)} u_{k+1},\end{aligned}$$

and thus, using the induction hypothesis for k ($2 \leq k \leq n-1$), we have

$$\begin{aligned}f_{k+1}(u_j) &= \rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\theta_1+\cdots+\theta_k} \circ f_k(u_j) = e^{i\theta_j} u_j, \quad 1 \leq j \leq k-1,\\ f_{k+1}(u_k) &= \rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\theta_1+\cdots+\theta_k} \circ f_k(u_k) = e^{i(\theta_1+\cdots+\theta_k)} e^{-i(\theta_1+\cdots+\theta_{k-1})} u_k = e^{i\theta_k} u_k,\\ f_{k+1}(u_{k+1}) &= \rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\theta_1+\cdots+\theta_k} \circ f_k(u_{k+1}) = e^{-i(\theta_1+\cdots+\theta_k)} u_{k+1},\\ f_{k+1}(u_j) &= \rho_{u_{k+1},\,-(\theta_1+\cdots+\theta_k)} \circ \rho_{u_k,\theta_1+\cdots+\theta_k} \circ f_k(u_j) = u_j, \quad k+1 \leq j \leq n,\end{aligned}$$

which proves the induction step.

As a summary, we proved that

$$f_n(u_j) = \begin{cases} e^{i\theta_j} u_j & \text{if } 1 \leq j \leq n-1,\\ e^{-i(\theta_1+\cdots+\theta_{n-1})} u_n & \text{when } j = n,\end{cases}$$

but since $\theta_1 + \cdots + \theta_n = 0$, we have $\theta_n = -(\theta_1 + \cdots + \theta_{n-1})$, and the last expression is in fact

$$f_n(u_n) = e^{i\theta_n} u_n.$$

Therefore, we proved that

$$f = \rho_{u_n,\theta_n} \circ \cdots \circ \rho_{u_1,\theta_1} = \rho_{u_n,\,-(\theta_1+\cdots+\theta_{n-1})} \circ \rho_{u_{n-1},\,\theta_1+\cdots+\theta_{n-1}} \circ \cdots \circ \rho_{u_2,\,-\theta_1} \circ \rho_{u_1,\theta_1},$$

and using Proposition 28.3, we also have

$$\begin{aligned}f &= \rho_{u_n,\,-(\theta_1+\cdots+\theta_{n-1})} \circ \rho_{u_{n-1},\,\theta_1+\cdots+\theta_{n-1}} \circ \cdots \circ \rho_{u_2,\,-\theta_1} \circ \rho_{u_1,\theta_1}\\ &= h_{u_n - u_{n-1}} \circ h_{u_n - e^{-i(\theta_1+\cdots+\theta_{n-1})} u_{n-1}} \circ \cdots \circ h_{u_2 - u_1} \circ h_{u_2 - e^{-i\theta_1} u_1}\\ &= h_{u_{n-1}+u_n} \circ h_{u_{n-1} + e^{i(\theta_1+\cdots+\theta_{n-1})} u_n} \circ \cdots \circ h_{u_1+u_2} \circ h_{u_1 + e^{i\theta_1} u_2},\end{aligned}$$

which completes the proof. □

We finally get our improved version of the Cartan–Dieudonné theorem.

Theorem 28.5. *Let E be a Hermitian space of dimension $n \geq 1$. Every rotation $f \in \mathbf{SU}(E)$ different from the identity is the composition of at most $2n-2$ standard hyperplane reflections. Every isometry $f \in \mathbf{U}(E)$ different from the identity is the composition of at most $2n-1$ isometries, all standard hyperplane reflections, except for possibly one Hermitian reflection. When $n \geq 2$, the identity is the composition of any reflection with itself.*

Proof. By Theorem 28.2, $f \in \mathbf{SU}(n)$ can be written as a composition

$$\rho_{u_n,\theta_n} \circ \cdots \circ \rho_{u_1,\theta_1},$$

where $(u_1, \ldots, u_n)$ is an orthonormal basis of eigenvectors. Since f is a rotation, $\det(f) = 1$, and this implies that $\theta_1 + \cdots + \theta_n = 0$. By Proposition 28.4,

$$f = h_{u_n - u_{n-1}} \circ h_{u_n - e^{-i(\theta_1+\cdots+\theta_{n-1})}u_{n-1}} \circ \cdots \circ h_{u_2 - u_1} \circ h_{u_2 - e^{-i\theta_1}u_1},$$

a composition of $2n - 2$ hyperplane reflections. In general, if $f \in \mathbf{U}(n)$, by the remark after Theorem 28.2, f can be written as $f = \rho_\theta \circ g$, where $g \in \mathbf{SU}(n)$ is a rotation, and ρ_θ is a Hermitian reflection. We conclude by applying what we just proved to g. ☐

As a corollary of Theorem 28.5, the following interesting result can be shown (this is not hard, do it!). First, recall that a linear map $f\colon E \to E$ is *self-adjoint* (or *Hermitian*) iff $f = f^*$. Then, the subgroup of $\mathbf{U}(n)$ generated by the Hermitian isometries is equal to the group

$$\mathbf{SU}(n)^{\pm} = \{f \in \mathbf{U}(n) \mid \det(f) = \pm 1\}.$$

Equivalently, $\mathbf{SU}(n)^{\pm}$ is equal to the subgroup of $\mathbf{U}(n)$ generated by the hyperplane reflections.

This problem had been left open by Dieudonné in [49]. Evidently, it was settled since the publication of the third edition of the book [49].

Inspection of the proof of Proposition 27.4 reveals that this Proposition also holds for Hermitian spaces. Thus, when $n \geq 3$, the composition of any two hyperplane reflections is equal to the composition of two flips. As a consequence, a version of Theorem 27.5 holds for rotations in a Hermitian space of dimension at least 3.

Theorem 28.6. *Let E be a Hermitan space of dimension $n \geq 3$. Every rotation $f \in \mathbf{SU}(E)$ is the composition of an even number of flips $f = f_{2k} \circ \cdots \circ f_1$, where $k \leq n-1$. Furthermore, if $u \neq 0$ is invariant under f (i.e. $u \in \mathrm{Ker}\,(f - \mathrm{id})$), we can pick the last flip f_{2k} such that $u \in F_{2k}^{\perp}$, where F_{2k} is the subspace of dimension $n - 2$ determining f_{2k}.*

Proof. It is identical to that of Theorem 27.5, except that it uses Theorem 28.5 instead of Theorem 27.1. The second part of the Proposition also holds, because if $u \neq 0$ is an eigenvector of f for 1, then u is one of the vectors in the orthonormal basis of eigenvectors used in 28.2. The details are left as an exercise. ☐

We now show that the QR-decomposition in terms of (complex) Householder matrices holds for complex matrices. We need the version of Proposition 28.1 and a trick at the end of the argument, but the proof is basically unchanged.

Proposition 28.7. *Let E be a nontrivial Hermitian space of dimension n. Given any orthonormal basis $(e_1, \ldots, e_n)$, for any n-tuple of vectors $(v_1, \ldots, v_n)$, there is a sequence of $n-1$ isometries $h_1, \ldots, h_{n-1}$, such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by*

$$r_j = h_{n-1} \circ \cdots \circ h_2 \circ h_1(v_j) \quad 1 \leq j \leq n,$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $(1 \leq j \leq n)$. Equivalently, the matrix R whose columns are the components of the r_j over the basis $(e_1, \ldots, e_n)$ is an upper triangular matrix. Furthermore, if we allow one more isometry h_n of the form

$$h_n = \rho_{e_n, \varphi_n} \circ \cdots \circ \rho_{e_1, \varphi_1}$$

after $h_1, \ldots, h_{n-1}$, we can ensure that the diagonal entries of R are nonnegative.

Proof. The proof is very similar to the proof of Proposition 13.3, but it needs to be modified a little bit since Proposition 28.1 is weaker than Proposition 13.2. We explain how to modify the induction step, leaving the base case and the rest of the proof as an exercise.

As in the proof of Proposition 13.3, the vectors $(e_1, \ldots, e_k)$ form a basis for the subspace denoted as U'_k, the vectors $(e_{k+1}, \ldots, e_n)$ form a basis for the subspace denoted as U''_k, the subspaces U'_k and U''_k are orthogonal, and $E = U'_k \oplus U''_k$. Let

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}).$$

We can write

$$u_{k+1} = u'_{k+1} + u''_{k+1},$$

where $u'_{k+1} \in U'_k$ and $u''_{k+1} \in U''_k$. Let

$$r_{k+1,k+1} = \left\| u''_{k+1} \right\|, \quad \text{and} \quad e^{i\theta_{k+1}} |u''_{k+1} \cdot e_{k+1}| = u''_{k+1} \cdot e_{k+1}.$$

If $u''_{k+1} = e^{i\theta_{k+1}} r_{k+1,k+1}\, e_{k+1}$, we let $h_{k+1} = \mathrm{id}$. Otherwise, by Proposition 28.1, there is a unique hyperplane reflection h_{k+1} such that

$$h_{k+1}(u''_{k+1}) = e^{i\theta_{k+1}} r_{k+1,k+1}\, e_{k+1},$$

where h_{k+1} is the reflection about the hyperplane H_{k+1} orthogonal to the vector

$$w_{k+1} = r_{k+1,k+1}\, e_{k+1} - e^{-i\theta_{k+1}} u''_{k+1}.$$

At the end of the induction, we have a triangular matrix R, but the diagonal entries $e^{i\theta_j} r_{j,j}$ of R may be complex. Letting

$$h_{n+1} = \rho_{e_n, -\theta_n} \circ \cdots \circ \rho_{e_1, -\theta_1},$$

we observe that the diagonal entries of the matrix of vectors

$$r'_j = h_{n+1} \circ h_n \circ \cdots \circ h_2 \circ h_1(v_j)$$

is triangular with nonnegative entries. $\square$

Remark: For numerical stability, it is preferable to use $w_{k+1} = r_{k+1,k+1}\, e_{k+1} + e^{-i\theta_{k+1}} u''_{k+1}$ instead of $w_{k+1} = r_{k+1,k+1}\, e_{k+1} - e^{-i\theta_{k+1}} u''_{k+1}$. The effect of that choice is that the diagonal entries in R will be of the form $-e^{i\theta_j} r_{j,j} = e^{i(\theta_j+\pi)} r_{j,j}$. Of course, we can make these entries nonegative by applying

$$h_{n+1} = \rho_{e_n, \pi-\theta_n} \circ \cdots \circ \rho_{e_1,\pi-\theta_1}$$

after h_n.

As in the Euclidean case, Proposition 28.7 immediately implies the QR-decomposition for arbitrary complex $n \times n$-matrices, where Q is now unitary (see Kincaid and Cheney [102], Golub and Van Loan [80], Trefethen and Bau [176], or Ciarlet [41]).

Proposition 28.8. *For every complex $n \times n$-matrix A, there is a sequence $H_1, \ldots, H_{n-1}$ of matrices, where each H_i is either a Householder matrix or the identity, and an upper triangular matrix R, such that*

$$R = H_{n-1} \cdots H_2 H_1 A.$$

As a corollary, there is a pair of matrices Q, R, where Q is unitary and R is upper triangular, such that $A = QR$ (a QR-decomposition of A). Furthermore, R can be chosen so that its diagonal entries are nonnegative. This can be achieved by a diagonal matrix D with entries such that $|d_{ii}| = 1$ for $i = 1, \ldots, n$, and we have $A = \widetilde{Q}\widetilde{R}$ with

$$\widetilde{Q} = H_1 \cdots H_{n-1} D, \quad \widetilde{R} = D^* R,$$

where $\widetilde{R}$ is upper triangular and has nonnegative diagonal entries

Proof. It is essentially identical to the proof of Proposition 13.4, and we leave the details as an exercise. For the last statement, observe that $h_n \circ \cdots \circ h_1$ is also an isometry. ☐

As in the Euclidean case, the QR-decomposition has applications to least squares problems. It is also possible to convert any complex matrix to bidiagonal form.

28.2 Affine Isometries (Rigid Motions)

In this section, we study very briefly the affine isometries of a Hermitian space. Most results holding for Euclidean affine spaces generalize without any problems to Hermitian spaces.

The characterization of the set of fixed points of an affine map is unchanged. Similarly, every affine isometry f (of a Hermitian space) can be written uniquely as

$$f = t \circ g, \quad \text{with} \quad t \circ g = g \circ t,$$

where g is an isometry having a fixed point, and t is a translation by a vector τ such that $\overrightarrow{f}(\tau) = \tau$, and with some additional nice properties (see Proposition 28.13). A generalization of the Cartan–Dieudonné theorem can easily be shown: every affine isometry in $\mathbf{Is}(n, \mathbb{C})$ can be written as the composition of at most $2n-1$ isometries if it has a fixed point, or else as the composition of at most $2n+1$ isometries, where all these isometries are affine hyperplane reflections except for possibly one affine Hermitian reflection. We also prove that every rigid motion in $\mathbf{SE}(n, \mathbb{C})$ is the composition of at most $2n-2$ flips (for $n \geq 3$).

Definition 28.2. Given any two nontrivial Hermitian affine spaces E and F of the same finite dimension n, a function $f \colon E \to F$ is *an affine isometry (or rigid map)* iff it is an affine map and

$$\left\| \overrightarrow{f(a)f(b)} \right\| = \left\| \overrightarrow{ab} \right\|,$$

for all $a, b \in E$. When $E = F$, an affine isometry $f \colon E \to E$ is also called a *rigid motion*.

Thus, an affine isometry is an affine map that preserves the distance. This is a rather strong requirement, but unlike the Euclidean case, not strong enough to force f to be an affine map.

The following simple Proposition is left as an exercise.

Proposition 28.9. *Given any two nontrivial Hermitian affine spaces E and F of the same finite dimension n, an affine map $f \colon E \to F$ is an affine isometry iff its associated linear map $\overrightarrow{f} \colon \overrightarrow{E} \to \overrightarrow{F}$ is an isometry. An affine isometry is a bijection.*

As in the Euclidean case, given an affine isometry $f \colon E \to E$, if $\overrightarrow{f}$ is a rotation, we call f a *proper (or direct) affine isometry*, and if $\overrightarrow{f}$ is a an improper linear isometry, we call f a *an improper (or skew) affine isometry*. It is easily shown that the set of affine isometries $f \colon E \to E$ forms a group, and those for which $\overrightarrow{f}$ is a rotation is a subgroup. The group of affine isometries, or rigid motions, is a subgroup of the affine group $\mathbf{GA}(E, \mathbb{C})$ denoted as $\mathbf{Is}(E, \mathbb{C})$ (or $\mathbf{Is}(n, \mathbb{C})$ when $E = \mathbb{C}^n$). The subgroup of $\mathbf{Is}(E, \mathbb{C})$ consisting of the direct rigid motions is also a subgroup of $\mathbf{SA}(E, \mathbb{C})$, and it is denoted as $\mathbf{SE}(E, \mathbb{C})$ (or $\mathbf{SE}(n, \mathbb{C})$, when $E = \mathbb{C}^n$). The translations are the affine isometries f for which $\overrightarrow{f} = \mathrm{id}$, the identity map on $\overrightarrow{E}$. The following Proposition is the counterpart of Proposition 14.14 for isometries between Hermitian vector spaces.

Proposition 28.10. *Given any two nontrivial Hermitian affine spaces E and F of the same finite dimension n, for every function $f \colon E \to F$, the following properties are equivalent:*

(1) f is an affine map and $\left\| \overrightarrow{f(a)f(b)} \right\| = \left\| \overrightarrow{ab} \right\|$, for all $a, b \in E$.

(2) $\left\| \overrightarrow{f(a)f(b)} \right\| = \left\| \overrightarrow{ab} \right\|$, and there is some $\Omega \in E$ such that

$$f(\Omega + i\overrightarrow{ab}) = f(\Omega) + i(\overrightarrow{f(\Omega)f(\Omega + \overrightarrow{ab})}),$$

for all $a, b \in E$.

Proof. Obviously, (1) implies (2). The proof that that (2) implies (1) is similar to the proof of Proposition 27.7, but uses Proposition 14.14 instead of Proposition 12.12. The details are left as an exercise. ☐

Inspection of the proof shows immediately that Proposition 27.8 holds for Hermitian spaces. For the sake of completeness, we restate the Proposition in the complex case.

Proposition 28.11. *Let E be any complex affine space of finite dimension For every affine map $f\colon E \to E$, let $Fix(f) = \{a \in E \mid f(a) = a\}$ be the set of fixed points of f. The following properties hold:*

(1) If f has some fixed point a, so that $Fix(f) \neq \emptyset$, then $Fix(f)$ is an affine subspace of E such that
$$Fix(f) = a + E(1, \overrightarrow{f}) = a + \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}),$$
where $E(1, \overrightarrow{f})$ is the eigenspace of the linear map $\overrightarrow{f}$ for the eigenvalue 1.

(2) The affine map f has a unique fixed point iff $E(1, \overrightarrow{f}) = \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = \{0\}$.

Affine orthogonal symmetries are defined just as in the Euclidean case, and Proposition 27.9 also applies to complex affine spaces.

Proposition 28.12. *Given any affine complex space E, if $f\colon E \to E$ and $g\colon E \to E$ are affine orthogonal symmetries about parallel affine subspaces F_1 and F_2, then $g \circ f$ is a translation defined by the vector $2\overrightarrow{ab}$, where $\overrightarrow{ab}$ is any vector perpendicular to the common direction $\overrightarrow{F}$ of F_1 and F_2 such that $\left\|\overrightarrow{ab}\right\|$ is the distance between F_1 and F_2, with $a \in F_1$ and $b \in F_2$. Conversely, every translation by a vector τ is obtained as the composition of two affine orthogonal symmetries about parallel affine subspaces F_1 and F_2 whose common direction is orthogonal to $\tau = \overrightarrow{ab}$, for some $a \in F_1$ and some $b \in F_2$ such that the distance betwen F_1 and F_2 is $\left\|\overrightarrow{ab}\right\|/2$.*

It is easy to check that the proof of Proposition 27.10 also holds in the Hermitian case.

Proposition 28.13. *Let E be a Hermitian affine space of finite dimension n. For every affine isometry $f\colon E \to E$, there is a unique affine isometry $g\colon E \to E$ and a unique translation $t = t_\tau$, with $\overrightarrow{f}(\tau) = \tau$ (i.e., $\tau \in \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})$), such that the set $Fix(g) = \{a \in E, \mid g(a) = a\}$ of fixed points of g is a nonempty affine subspace of E of direction*
$$\overrightarrow{G} = \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = E(1, \overrightarrow{f}),$$
and such that
$$f = t \circ g \quad \text{and} \quad t \circ g = g \circ t.$$
Furthermore, we have the following additional properties:

(a) $f = g$ and $\tau = 0$ iff f has some fixed point, i.e., iff $Fix(f) \neq \emptyset$.

(b) If f has no fixed points, i.e., $Fix(f) = \emptyset$, then $\dim(\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})) \geq 1$.

The remarks made in the Euclidean case also apply to the Hermitian case. In particular, the fact that E has finite dimension is only used to prove (b).

A version of the Cartan–Dieudonné also holds for affine isometries, but it may not be possible to get rid of Hermitian reflections entirely.

Theorem 28.14. *Let E be an affine Hermitian space of dimension $n \geq 1$. Every affine isometry in $\mathbf{Is}(n, \mathbb{C})$ can be written as the composition of at most $2n - 1$ affine isometries if it has a fixed point, or else as the composition of at most $2n + 1$ affine isometries, where all these isometries are affine hyperplane reflections except for possibly one affine Hermitian reflection. When $n \geq 2$, the identity is the composition of any reflection with itself.*

Proof. The proof is very similar to the proof of Theorem 27.11, except that it uses Theorem 28.5 instead of Theorem 27.1. The details are left as an exercise. □

When $n \geq 3$, as in the Euclidean case, we can characterize the affine isometries in $\mathbf{SE}(n, \mathbb{C})$ in terms of flips, and we can even bound the number of flips by $2n - 2$.

Theorem 28.15. *Let E be a Hermitian affine space of dimension $n \geq 3$. Every rigid motion $f \in \mathbf{SE}(E, \mathbb{C})$ is the composition of an even number of affine flips $f = f_{2k} \circ \cdots \circ f_1$, where $k \leq n - 1$.*

Proof. It is very similar to the proof of theorem 27.12, but it uses Proposition 28.6 instead of Proposition 27.5. The details are left as an exercise. □

A more detailed study of the rigid motions of Hermitian spaces of dimension 2 and 3 would seem worthwhile, but we are not aware of any reference on this subject.

Chapter 29

The Geometry of Bilinear Forms; Witt's Theorem; The Cartan–Dieudonné Theorem

29.1 Bilinear Forms

In this chapter, we study the structure of a K-vector space E endowed with a nondegenerate bilinear form $\varphi\colon E \times E \to K$ (for any field K), which can be viewed as a kind of generalized inner product. Unlike the case of an inner product, there may be nonzero vectors $u \in E$ such that $\varphi(u, u) = 0$ so the map $u \mapsto \varphi(u, u)$ can no longer be interpreted as a notion of square length (also, $\varphi(u, u)$ may not be real and positive!). However, the notion of orthogonality survives: we say that $u, v \in E$ are orthogonal iff $\varphi(u, v) = 0$. Under some additional conditions on φ, it is then possible to split E into orthogonal subspaces having some special properties. It turns out that the special cases where φ is symmetric (or Hermitian) or skew-symmetric (or skew-Hermitian) can be handled uniformly using a deep theorem due to Witt (the Witt decomposition theorem (1936)).

We begin with the very general situation of a bilinear form $\varphi\colon E \times F \to K$, where K is an arbitrary field, possibly of characteristric 2. Actually, even though at first glance this may appear to be an unnecessary abstraction, it turns out that this situation arises in attempting to prove properties of a bilinear map $\varphi\colon E \times E \to K$, because it may be necessary to restrict φ to different subspaces U and V of E. This general approach was pioneered by Chevalley [37], E. Artin [6], and Bourbaki [24]. The third source was a major source of inspiration, and many proofs are taken from it. Other useful references include Snapper and Troyer [162], Berger [12], Jacobson [98], Grove [83], Taylor [174], and Berndt [14].

Definition 29.1. Given two vector spaces E and F over a field K, a map $\varphi\colon E \times F \to K$ is a *bilinear form* iff the following conditions hold: For all $u, u_1, u_2 \in E$, all $v, v_1, v_2 \in F$, for all $\lambda, \mu \in K$, we have

$$\begin{aligned}\varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v)\\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2)\\ \varphi(\lambda u, v) &= \lambda\varphi(u, v)\\ \varphi(u, \mu v) &= \mu\varphi(u, v).\end{aligned}$$

A bilinear form as in Definition 29.1 is sometimes called a *pairing*. The first two conditions imply that $\varphi(0, v) = \varphi(u, 0) = 0$ for all $u \in E$ and all $v \in F$.

If $E = F$, observe that

$$\begin{aligned}\varphi(\lambda u + \mu v, \lambda u + \mu v) &= \lambda\varphi(u, \lambda u + \mu v) + \mu\varphi(v, \lambda u + \mu v)\\ &= \lambda^2\varphi(u, u) + \lambda\mu\varphi(u, v) + \lambda\mu\varphi(v, u) + \mu^2\varphi(v, v).\end{aligned}$$

If we let $\lambda = \mu = 1$, we get

$$\varphi(u + v, u + v) = \varphi(u, u) + \varphi(u, v) + \varphi(v, u) + \varphi(v, v).$$

If φ is *symmetric*, which means that

$$\varphi(u, v) = \varphi(v, u) \quad \text{for all } u, v \in E,$$

then

$$2\varphi(u, v) = \varphi(u + v, u + v) - \varphi(u, u) - \varphi(v, v). \tag{$*$}$$

The function Φ defined such that

$$\Phi(u) = \varphi(u, u) \quad u \in E,$$

is called the *quadratic form* associated with φ. If the field K is not of characteristic 2, then φ is completely determined by its quadratic form Φ. The symmetric bilinear form φ is called the *polar form* of Φ. This suggests the following definition.

Definition 29.2. A function $\Phi\colon E \to K$ is a *quadratic form* on E if the following conditions hold:

(1) We have $\Phi(\lambda u) = \lambda^2\Phi(u)$, for all $u \in E$ and all $\lambda \in E$.

(2) The map φ' given by $\varphi'(u, v) = \Phi(u + v) - \Phi(u) - \Phi(v)$ is bilinear. Obviously, the map φ' is symmetric.

Since $\Phi(x + x) = \Phi(2x) = 4\Phi(x)$, we have

$$\varphi'(u, u) = 2\Phi(u) \quad u \in E.$$

If the field K is not of characteristic 2, then $\varphi = \frac{1}{2}\varphi'$ is the unique symmetric bilinear form such that that $\varphi(u,u) = \Phi(u)$ for all $u \in E$. The bilinear form $\varphi = \frac{1}{2}\varphi'$ is called the *polar form* of Φ. In this case, there is a bijection between the set of bilinear forms on E and the set of quadratic forms on E.

If K is a field of characteristic 2, then φ' is *alternating*, which means that

$$\varphi'(u,u) = 0 \quad \text{for all } u \in E.$$

Thus if K is a field of characteristic 2, then Φ cannot be recovered from the symmetric bilinear form φ'.

If $(e_1, \ldots, e_n)$ is a basis of E, it is easy to show that

$$\Phi\Big(\sum_{i=1}^n \lambda_i e_i\Big) = \sum_{i=1}^n \lambda_i^2 \Phi(e_i) + \sum_{i \neq j} \lambda_i \lambda_j \varphi'(e_i, e_j).$$

This shows that the quadratic form Φ is completely determined by the scalars $\Phi(e_i)$ and $\varphi'(e_i, e_j)$ $(i \neq j)$. Furthermore, given any bilinear form $\psi\colon E \times E \to K$ (not necessarily symmetric) we can define a quadratic form Φ by setting $\Phi(x) = \psi(x,x)$, and we immediately check that the symmetric bilinear form φ' associated with Φ is given by $\varphi'(u,v) = \psi(u,v) + \psi(v,u)$. Using the above facts, it is not hard to prove that given any quadratic form Φ, there is some (nonsymmetric) bilinear form ψ such that $\Phi(u) = \psi(u,u)$ for all $u \in E$ (see Bourbaki [24], Section §3.4, Proposition 2). Thus, quadratic forms are more general than symmetric bilinear forms (except in characteristic $\neq 2$).

Definition 29.3. Given any bilinear form $\varphi\colon E \times E \to K$ where K is a field of any characteristic, we say that φ is *alternating* if

$$\varphi(u,u) = 0 \quad \text{for all } u \in E,$$

and *skew-symmetric* if

$$\varphi(v,u) = -\varphi(u,v) \quad \text{for all } u, v \in E.$$

If K is a field of any characteristic, the identity

$$\varphi(u+v, u+v) = \varphi(u,u) + \varphi(u,v) + \varphi(v,u) + \varphi(v,v)$$

shows that if φ is alternating, then

$$\varphi(v,u) = -\varphi(u,v) \quad \text{for all } u, v \in E,$$

that is, φ is skew-symmetric. Conversely, if the field K is not of characteristic 2, then a skew-symmetric bilinear map is alternating, since $\varphi(u,u) = -\varphi(u,u)$ implies $\varphi(u,u) = 0$.

An important consequence of bilinearity is that a pairing yields a linear map from E into F^* and a linear map from F into E^* (where $E^* = \mathrm{Hom}_K(E,K)$, the *dual* of E, is the set of linear maps from E to K, called *linear forms*).

Definition 29.4. Given a bilinear map $\varphi\colon E \times F \to K$, for every $u \in E$, let $l_\varphi(u)$ be the linear form in F^* given by

$$l_\varphi(u)(y) = \varphi(u, y) \quad \text{for all } y \in F,$$

and for every $v \in F$, let $r_\varphi(v)$ be the linear form in E^* given by

$$r_\varphi(v)(x) = \varphi(x, v) \quad \text{for all } x \in E.$$

Because φ is bilinear, the maps $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear.

Definition 29.5. A bilinear map $\varphi\colon E \times F \to K$ is said to be *nondegenerate* iff the following conditions hold:

(1) For every $u \in E$, if $\varphi(u, v) = 0$ for all $v \in F$, then $u = 0$, and

(2) For every $v \in F$, if $\varphi(u, v) = 0$ for all $u \in E$, then $v = 0$.

The following proposition shows the importance of l_φ and r_φ.

Proposition 29.1. *Given a bilinear map $\varphi\colon E \times F \to K$, the following properties hold:*

(a) The map l_φ is injective iff Property (1) of Definition 29.5 holds.

(b) The map r_φ is injective iff Property (2) of Definition 29.5 holds.

(c) The bilinear form φ is nondegenerate and iff l_φ and r_φ are injective.

(d) If the bilinear form φ is nondegenerate and if E and F have finite dimensions, then $\dim(E) = \dim(F)$, and $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are linear isomorphisms.

Proof. (a) Assume that (1) of Definition 29.5 holds. If $l_\varphi(u) = 0$, then $l_\varphi(u)$ is the linear form whose value is 0 for all y; that is,

$$l_\varphi(u)(y) = \varphi(u, y) = 0 \quad \text{for all } y \in F,$$

and by (1) of Definition 29.5, we must have $u = 0$. Therefore, l_φ is injective. Conversely, if l_φ is injective, and if

$$l_\varphi(u)(y) = \varphi(u, y) = 0 \quad \text{for all } y \in F,$$

then $l_\varphi(u)$ is the zero form, and by injectivity of l_φ, we get $u = 0$; that is, (1) of Definition 29.5 holds.

(b) The proof is obtained by swapping the arguments of φ.

(c) This follows from (a) and (b).

(d) If E and F are finite dimensional, then $\dim(E) = \dim(E^*)$ and $\dim(F) = \dim(F^*)$. Since φ is nondegenerate, $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are injective, so $\dim(E) \leq \dim(F^*) = \dim(F)$ and $\dim(F) \leq \dim(E^*) = \dim(E)$, which implies that

$$\dim(E) = \dim(F),$$

and thus, $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are bijective. $\square$

As a corollary of Proposition 29.1, we have the following characterization of a nondegenerate bilinear map. The proof is left as an exercise.

Proposition 29.2. *Given a bilinear map $\varphi\colon E \times F \to K$, if E and F have the same finite dimension, then the following properties are equivalent:*

(1) The map l_φ is injective.

(2) The map l_φ is surjective.

(3) The map r_φ is injective.

(4) The map r_φ is surjective.

(5) The bilinear form φ is nondegenerate.

Observe that in terms of the canonical pairing between E^* and E given by

$$\langle f, u\rangle = f(u), \quad f \in E^*, u \in E,$$

(and the canonical pairing between F^* and F), we have

$$\varphi(u, v) = \langle l_\varphi(u), v\rangle = \langle r_\varphi(v), u\rangle \quad u \in E, v \in F.$$

Proposition 29.3. *Given a bilinear map $\varphi\colon E \times F \to K$, if φ is nondegenerate and E and F are finite-dimensional, then $\dim(E) = \dim(F) = n$, and for every basis $(e_1, \ldots, e_n)$ of E, there is a basis $(f_1, \ldots, f_n)$ of F such that $\varphi(e_i, f_j) = \delta_{ij}$, for all $i, j = 1, \ldots, n$.*

Proof. Since φ is nondegenerate, by Proposition 29.1 we have $\dim(E) = \dim(F) = n$, and by Proposition 29.2, the linear map r_φ is bijective. Then, if $(e_1^*, \ldots, e_n^*)$ is the dual basis (in E^*) of the basis $(e_1, \ldots, e_n)$, the vectors $(f_1, \ldots, f_n)$ given by $f_i = r_\varphi^{-1}(e_i^*)$ form a basis of F, and we have

$$\varphi(e_i, f_j) = \langle r_\varphi(f_j), e_i\rangle = \langle e_i^*, e_j\rangle = \delta_{ij},$$

as claimed. ☐

If $E = F$ and φ is symmetric, then we have the following interesting result.

Theorem 29.4. *Given any bilinear form $\varphi\colon E \times E \to K$ with $\dim(E) = n$, if φ is symmetric (possibly degenerate) and K does not have characteristic 2, then there is a basis $(e_1, \ldots, e_n)$ of E such that $\varphi(e_i, e_j) = 0$, for all $i \neq j$.*

Proof. We proceed by induction on $n \geq 0$, following a proof due to Chevalley. The base case $n = 0$ is trivial. For the induction step, assume that $n \geq 1$ and that the induction hypothesis holds for all vector spaces of dimension $n-1$. If $\varphi(u, v) = 0$ for all $u, v \in E$, then the statement holds trivially. Otherwise, since K does not have characteristic 2, equation

$$2\varphi(u, v) = \varphi(u + v, u + v) - \varphi(u, u) - \varphi(v, v) \qquad (*)$$

show that there is some nonzero vector $e_1 \in E$ such that $\varphi(e_1, e_1) \neq 0$ since otherwise φ would vanish for all $u, v \in E$. We claim that the set

$$H = \{v \in E \mid \varphi(e_1, v) = 0\}$$

has dimension $n - 1$, and that $e_1 \notin H$.

This is because

$$H = \mathrm{Ker}\,(l_\varphi(e_1)),$$

where $l_\varphi(e_1)$ is the linear form in E^* determined by e_1. Since $\varphi(e_1, e_1) \neq 0$, we have $e_1 \notin H$, the linear form $l_\varphi(e_1)$ is not the zero form, and thus its kernel is a hyperplane H (a subspace of dimension $n - 1$). Since $\dim(H) = n - 1$ and $e_1 \notin H$, we have the direct sum

$$E = H \oplus K e_1.$$

By the induction hypothesis applied to H, we get a basis $(e_2, \ldots, e_n)$ of vectors in H such that $\varphi(e_i, e_j) = 0$, for all $i \neq j$ with $2 \leq i, j \leq n$. Since $\varphi(e_1, v) = 0$ for all $v \in H$ and since φ is symmetric, we also have $\varphi(v, e_1) = 0$ for all $v \in H$, so we obtain a basis $(e_1, \ldots, e_n)$ of E such that $\varphi(e_i, e_j) = 0$, for all $i \neq j$. $\square$

If E and F are finite-dimensional vector spaces and if $(e_1, \ldots, e_m)$ is a basis of E and $(f_1, \ldots, f_n)$ is a basis of F then the bilinearity of φ yields

$$\varphi\left(\sum_{i=1}^{m} x_i e_i, \sum_{j=1}^{n} y_j f_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i \varphi(e_i, f_j) y_j.$$

This shows that φ is completely determined by the $n \times m$ matrix $M = (m_{ij})$ with $m_{ij} = \varphi(e_j, f_i)$, and in matrix form, we have

$$\varphi(x, y) = x^\top M^\top y = y^\top M x,$$

where x and y are the column vectors associated with $(x_1, \ldots, x_m) \in K^m$ and $(y_1, \ldots, y_n) \in K^n$. As in Section 12.1, we are committing the slight abuse of notation of letting x denote both the vector $x = \sum_{i=1}^n x_i e_i$ and the column vector associated with $(x_1, \ldots, x_n)$ (and similarly for y).

Definition 29.6. If $(e_1, \ldots, e_m)$ is a basis of E and $(f_1, \ldots, f_n)$ is a basis of F, for any bilinear form $\varphi \colon E \times F \to K$, the $n \times m$ matrix $M = (m_{ij})$ given by $m_{ij} = \varphi(e_j, f_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ is called the *matrix of* φ *with respect to the bases* $(e_1, \ldots, e_m)$ *and* $(f_1, \ldots, f_n)$.

The following fact is easily proved.

Proposition 29.5. *If* $m = \dim(E) = \dim(F) = n$*, then* φ *is nondegenerate iff the matrix* M *is invertible iff* $\det(M) \neq 0$.

As we will see later, most bilinear forms that we will encounter are equivalent to one whose matrix is of the following form:

1. I_n, $-I_n$.
2. If $p + q = n$, with $p, q \geq 1$,
$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}$$
3. If $n = 2m$,
$$J_{m.m} = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}$$
4. If $n = 2m$,
$$A_{m,m} = I_{m.m} J_{m.m} = \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix}.$$

If we make changes of bases given by matrices P and Q, so that $x = Px'$ and $y = Qy'$, then the new matrix expressing φ is $P^\top M Q$. In particular, if $E = F$ and the same basis is used, then the new matrix is $P^\top M P$. This shows that if φ is nondegenerate, then the determinant of φ is determined up to a square element.

Observe that if φ is a symmetric bilinear form ($E = F$) and if K does not have characteristic 2, then by Theorem 29.4, there is a basis of E with respect to which the matrix M representing φ is a diagonal matrix. If $K = \mathbb{R}$ or $K = \mathbb{C}$, this allows us to classify completely the symmetric bilinear forms. Recall that $\Phi(u) = \varphi(u, u)$ for all $u \in E$.

Proposition 29.6. *Given any bilinear form $\varphi\colon E \times E \to K$ with $\dim(E) = n$, if φ is symmetric and K does not have characteristic 2, then there is a basis $(e_1, \ldots, e_n)$ of E such that*
$$\Phi\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^r \lambda_i x_i^2,$$
for some $\lambda_i \in K - \{0\}$ and with $r \leq n$. Furthermore, if $K = \mathbb{C}$, then there is a basis $(e_1, \ldots, e_n)$ of E such that
$$\Phi\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^r x_i^2,$$
and if $K = \mathbb{R}$, then there is a basis $(e_1, \ldots, e_n)$ of E such that
$$\Phi\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^p x_i^2 - \sum_{i=p+1}^{p+q} x_i^2,$$
with $0 \leq p, q$ and $p + q \leq n$.

Proof. The first statement is a direct consequence of Theorem 29.4. If $K = \mathbb{C}$, then every λ_i has a square root μ_i, and if replace e_i by e_i/μ_i, we obtained the desired form.

If $K = \mathbb{R}$, then there are two cases:

1. If $\lambda_i > 0$, let μ_i be a positive square root of λ_i and replace e_i by e_i/μ_i.

2. If $\lambda_i < 0$, et μ_i be a positive square root of $-\lambda_i$ and replace e_i by e_i/μ_i.

□

In the nondegenerate case, the matrices corresponding to the complex and the real case are, $I_n, -I_n$, and $I_{p,q}$. Observe that the second statement of Proposition 29.6 holds in any field in which every element has a square root. In the case $K = \mathbb{R}$, we can show that(p, q) only depends on φ.

Definition 29.7. Let $\varphi\colon E \times E \to \mathbb{R}$ be any symmetric real bilinear form. For any subspace U of E, we say that φ is *positive definite on* U iff $\varphi(u, u) > 0$ for all nonzero $u \in U$, and we say that φ is *negative definite on* U iff $\varphi(u, u) < 0$ for all nonzero $u \in U$. Then, let

$$r = \max\{\dim(U) \mid U \subseteq E,\ \varphi \text{ is positive definite on } U\}$$

and let

$$s = \max\{\dim(U) \mid U \subseteq E,\ \varphi \text{ is negative definite on } U\}$$

Proposition 29.7. *(Sylvester's inertia law) Given any symmetric bilinear form $\varphi\colon E\times E \to \mathbb{R}$ with $\dim(E) = n$, for any basis $(e_1, \ldots, e_n)$ of E such that*

$$\Phi\left(\sum_{i=1}^{n} x_i e_i\right) = \sum_{i=1}^{p} x_i^2 - \sum_{i=p+1}^{p+q} x_i^2,$$

with $0 \leq p, q$ and $p + q \leq n$, the integers p, q depend only on φ; in fact, $p = r$ and $q = s$, with r and s as defined above.

Proof. If we let U be the subspace spanned by $(e_1, \ldots, e_p)$, then φ is positive definite on U, so $r \geq p$. Similarly, if we let V be the subspace spanned by $(e_{p+1}, \ldots, e_{p+q})$, then φ is negative definite on V, so $s \geq q$.

Next, if W_1 is any subspace of maximum dimension such that φ is positive definite on W_1, and if we let V' be the subspace spanned by $(e_{p+1}, \ldots, e_n)$, then $\varphi(u, u) \leq 0$ on V', so $W_1 \cap V' = (0)$, which implies that $\dim(W_1) + \dim(V') \leq n$, and thus, $r + n - p \leq n$; that is, $r \leq p$. Similarly, if W_2 is any subspace of maximum dimension such that φ is negative definite on W_2, and if we let U' be the subspace spanned by $(e_1, \ldots, e_p, e_{p+q+1}, \ldots, e_n)$, then $\varphi(u, u) \geq 0$ on U', so $W_2 \cap U' = (0)$, which implies that $s + n - q \leq n$; that is, $s \leq q$. Therefore, $p = r$ and $q = s$, as claimed □

These last two results can be generalized to ordered fields. For example, see Snapper and Troyer [162], Artin [6], and Bourbaki [24].

29.2 Sesquilinear Forms

In order to accomodate Hermitian forms, we assume that some involutive automorphism, $\lambda \mapsto \overline{\lambda}$, of the field K is given. This automorphism of K satisfies the following properties:

$$\begin{aligned} \overline{(\lambda + \mu)} &= \overline{\lambda} + \overline{\mu} \\ \overline{(\lambda\mu)} &= \overline{\lambda}\,\overline{\mu} \\ \overline{\overline{\lambda}} &= \lambda. \end{aligned}$$

Since any field automorphism maps the multiplicative unit 1 to itself, we have $\overline{1} = 1$.

If the automorphism $\lambda \mapsto \overline{\lambda}$ is the identity, then we are in the standard situation of a bilinear form. When $K = \mathbb{C}$ (the complex numbers), then we usually pick the automorphism of $\mathbb{C}$ to be *conjugation*; namely, the map

$$a + ib \mapsto a - ib.$$

Definition 29.8. Given two vector spaces E and F over a field K with an involutive automorphism $\lambda \mapsto \overline{\lambda}$, a map $\varphi\colon E \times F \to K$ is a (right) *sesquilinear form* iff the following conditions hold: For all $u, u_1, u_2 \in E$, all $v, v_1, v_2 \in F$, for all $\lambda, \mu \in K$, we have

$$\begin{aligned} \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v) \\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2) \\ \varphi(\lambda u, v) &= \lambda\varphi(u, v) \\ \varphi(u, \mu v) &= \overline{\mu}\varphi(u, v). \end{aligned}$$

Again, $\varphi(0, v) = \varphi(u, 0) = 0$. If $E = F$, then we have

$$\begin{aligned} \varphi(\lambda u + \mu v, \lambda u + \mu v) &= \lambda\varphi(u, \lambda u + \mu v) + \mu\varphi(v, \lambda u + \mu v) \\ &= \lambda\overline{\lambda}\varphi(u, u) + \lambda\overline{\mu}\varphi(u, v) + \overline{\lambda}\mu\varphi(v, u) + \mu\overline{\mu}\varphi(v, v). \end{aligned}$$

If we let $\lambda = \mu = 1$ and then $\lambda = 1, \mu = -1$, we get

$$\begin{aligned} \varphi(u + v, u + v) &= \varphi(u, u) + \varphi(u, v) + \varphi(v, u) + \varphi(v, v) \\ \varphi(u - v, u - v) &= \varphi(u, u) - \varphi(u, v) - \varphi(v, u) + \varphi(v, v), \end{aligned}$$

so by subtraction, we get

$$2(\varphi(u, v) + \varphi(v, u)) = \varphi(u + v, u + v) - \varphi(u - v, u - v) \quad \text{for } u, v \in E.$$

If we replace v by λv (with $\lambda \neq 0$), we get

$$2(\overline{\lambda}\varphi(u, v) + \lambda\varphi(v, u)) = \varphi(u + \lambda v, u + \lambda v) - \varphi(u - \lambda v, u - \lambda v),$$

and by combining the above two equations, we get

$$\begin{aligned}2(\lambda-\overline{\lambda})\varphi(u,v) &= \lambda\varphi(u+v,u+v)-\lambda\varphi(u-v,u-v)\\ &\quad -\varphi(u+\lambda v,u+\lambda v)+\varphi(u-\lambda v,u-\lambda v).\end{aligned} \tag{$*$}$$

If the automorphism $\lambda \mapsto \overline{\lambda}$ is not the identity, then there is some $\lambda \in K$ such that $\lambda - \overline{\lambda} \neq 0$, and if K is not of characteristic 2, then we see that the sesquilinear form φ is completely determined by its restriction to the diagonal (that is, the set of values $\{\varphi(u,u) \mid u \in E\}$). In the special case where $K = \mathbb{C}$, we can pick $\lambda = i$, and we get

$$4\varphi(u,v) = \varphi(u+v,u+v)-\varphi(u-v,u-v)+i\varphi(u+\lambda v,u+\lambda v)-i\varphi(u-\lambda v,u-\lambda v).$$

Remark: If the automorphism $\lambda \mapsto \overline{\lambda}$ is the identity, then in general φ is not determined by its value on the diagonal, unless φ is symmetric.

In the sesquilinear setting, it turns out that the following two cases are of interest:

1. We have
$$\varphi(v,u) = \overline{\varphi(u,v)}, \quad \text{for all } u,v \in E,$$
in which case we say that φ is *Hermitian*. In the special case where $K = \mathbb{C}$ and the involutive automorphism is conjugation, we see that $\varphi(u,u) \in \mathbb{R}$, for $u \in E$.

2. We have
$$\varphi(v,u) = -\overline{\varphi(u,v)}, \quad \text{for all } u,v \in E,$$
in which case we say that φ is *skew-Hermitian*.

We observed that in characteristic different from 2, a sesquilinear form is determined by its restriction to the diagonal. For Hermitian and skew-Hermitian forms, we have the following kind of converse.

Proposition 29.8. *If φ is a nonzero Hermitian or skew-Hermitian form and if $\varphi(u,u) = 0$ for all $u \in E$, then K is of characteristic 2 and the automorphism $\lambda \mapsto \overline{\lambda}$ is the identity.*

Proof. We give the proof in the Hermitian case, the skew-Hermitian case being left as an exercise. Assume that φ is alternating. From the identity

$$\varphi(u+v,u+v) = \varphi(u,u)+\varphi(u,v)+\overline{\varphi(u,v)}+\varphi(v,v),$$

we get

$$\varphi(u,v) = -\overline{\varphi(u,v)} \quad \text{for all } u,v \in E.$$

Since φ is not the zero form, there exist some nonzero vectors $u,v \in E$ such that $\varphi(u,v) = 1$. For any $\lambda \in K$, we have

$$\lambda\varphi(u,v) = \varphi(\lambda u,v) = -\overline{\varphi(\lambda u,v)} = -\overline{\lambda}\,\overline{\varphi(u,v)},$$

and since $\varphi(u, v) = 1$, we get
$$\lambda = -\overline{\lambda} \quad \text{for all } \lambda \in K.$$
For $\lambda = 1$, we get $1 = -1$, which means that K has characterictic 2. But then
$$\lambda = -\overline{\lambda} = \overline{\lambda} \quad \text{for all } \lambda \in K,$$
so the automorphism $\lambda \mapsto \overline{\lambda}$ is the identity. $\square$

The definition of the linear maps l_φ and r_φ requires a small twist due to the automorphism $\lambda \mapsto \overline{\lambda}$.

Definition 29.9. Given a vector space E over a field K with an involutive automorphism $\lambda \mapsto \overline{\lambda}$, we define the K-vector space $\overline{E}$ as E with its abelian group structure, but with scalar multiplication given by
$$(\lambda, u) \mapsto \overline{\lambda}u.$$
Given two K-vector spaces E and F, a *semilinear map* $f\colon E \to F$ is a function, such that for all $u, v \in E$, for all $\lambda \in K$, we have
$$\begin{aligned} f(u + v) &= f(u) + f(v) \\ f(\lambda u) &= \overline{\lambda} f(u). \end{aligned}$$

Because $\overline{\overline{\lambda}} = \lambda$, observe that a function $f\colon E \to F$ is semilinear iff it is a linear map $f\colon \overline{E} \to F$. The K-vector spaces E and $\overline{E}$ are isomorphic, since any basis $(e_i)_{i\in I}$ of E is also a basis of $\overline{E}$.

The maps l_φ and r_φ are defined as follows:

For every $u \in E$, let $l_\varphi(u)$ be the linear form in F^* defined so that
$$l_\varphi(u)(y) = \overline{\varphi(u, y)} \quad \text{for all } y \in F,$$
and for every $v \in F$, let $r_\varphi(v)$ be the linear form in E^* defined so that
$$r_\varphi(v)(x) = \varphi(x, v) \quad \text{for all } x \in E.$$

The reader should check that because we used $\overline{\varphi(u, y)}$ in the definition of $l_\varphi(u)(y)$, the function $l_\varphi(u)$ is indeed a linear form in F^*. It is also easy to check that l_φ is a linear map $l_\varphi\colon \overline{E} \to F^*$, and that r_φ is a linear map $r_\varphi\colon \overline{F} \to E^*$ (equivalently, $l_\varphi\colon E \to F^*$ and $r_\varphi\colon F \to E^*$ are semilinear).

The notion of a nondegenerate sesquilinear form is identical to the notion for bilinear forms. For the convenience of the reader, we repeat the definition.

Definition 29.10. A sesquilinear map $\varphi\colon E \times F \to K$ is said to be *nondegenerate* iff the following conditions hold:

(1) For every $u \in E$, if $\varphi(u, v) = 0$ for all $v \in F$, then $u = 0$, and

(2) For every $v \in F$, if $\varphi(u, v) = 0$ for all $u \in E$, then $v = 0$.

Proposition 29.1 translates into the following proposition. The proof is left as an exercise.

Proposition 29.9. *Given a sesquilinear map $\varphi\colon E \times F \to K$, the following properties hold:*

(a) The map l_φ is injective iff Property (1) of Definition 29.10 holds.

(b) The map r_φ is injective iff Property (2) of Definition 29.10 holds.

(c) The sesquilinear form φ is nondegenerate and iff l_φ and r_φ are injective.

(d) If the sesquillinear form φ is nondegenerate and if E and F have finite dimensions, then $\dim(E) = \dim(F)$, and $l_\varphi\colon \overline{E} \to F^*$ and $r_\varphi\colon \overline{F} \to E^*$ are linear isomorphisms.

Propositions 29.2 and 29.3 also generalize to sesquilinear forms. We also have the following version of Theorem 29.4, whose proof is left as an exercise.

Theorem 29.10. *Given any sesquilinear form $\varphi\colon E \times E \to K$ with $\dim(E) = n$, if φ is Hermitian and K does not have characteristic 2, then there is a basis $(e_1, \ldots, e_n)$ of E such that $\varphi(e_i, e_j) = 0$, for all $i \neq j$.*

As in Section 29.1, if E and F are finite-dimensional vector spaces and if $(e_1, \ldots, e_m)$ is a basis of E and $(f_1, \ldots, f_n)$ is a basis of F then the sesquilinearity of φ yields

$$\varphi\biggl(\sum_{i=1}^m x_i e_i, \sum_{j=1}^n y_j f_j\biggr) = \sum_{i=1}^m \sum_{j=1}^n x_i \varphi(e_i, f_j) \overline{y}_j.$$

This shows that φ is completely determined by the $n \times m$ matrix $M = (m_{ij})$ with $m_{ij} = \varphi(e_j, f_i)$, and in matrix form, we have

$$\varphi(x, y) = x^\top M^\top \overline{y} = y^* M x,$$

where x and $\overline{y}$ are the column vectors associated with $(x_1, \ldots, x_m) \in K^m$ and $(\overline{y}_1, \ldots, \overline{y}_n) \in K^n$, and $y^* = \overline{y}^\top$. As earlier, we are committing the slight abuse of notation of letting x denote both the vector $x = \sum_{i=1}^n x_i e_i$ and the column vector associated with $(x_1, \ldots, x_n)$ (and similarly for y).

Definition 29.11. If $(e_1, \ldots, e_m)$ is a basis of E and $(f_1, \ldots, f_n)$ is a basis of F, for any sesquilinear form $\varphi\colon E \times F \to K$, the $n \times m$ matrix $M = (m_{ij})$ given by $m_{ij} = \varphi(e_j, f_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ is called the *matrix of φ with respect to the bases $(e_1, \ldots, e_m)$ and $(f_1, \ldots, f_n)$*.

Proposition 29.5 also holds for sesquilinear forms and their matrix representations.

Observe that if φ is a Hermitian form ($E = F$) and if K does not have characteristic 2, then by Theorem 29.10, there is a basis of E with respect to which the matrix M representing φ is a diagonal matrix. If $K = \mathbb{C}$, then these entries are real, and this allows us to classify completely the Hermitian forms.

Proposition 29.11. *Given any Hermitian form $\varphi\colon E \times E \to \mathbb{C}$ with $\dim(E) = n$, there is a basis $(e_1, \ldots, e_n)$ of E such that*

$$\Phi\left(\sum_{i=1}^{n} x_i e_i\right) = \sum_{i=1}^{p} x_i^2 - \sum_{i=p+1}^{p+q} x_i^2,$$

with $0 \leq p, q$ and $p + q \leq n$.

The proof of Proposition 29.11 is the same as the real case of Proposition 29.6. Sylvester's inertia law (Proposition 29.7) also holds for Hermitian forms: p and q only depend on φ.

29.3 Orthogonality

In this section we assume that we are dealing with a sesquilinear form $\varphi\colon E \times F \to K$. We allow the automorphism $\lambda \mapsto \overline{\lambda}$ to be the identity, in which case φ is a bilinear form. This way, we can deal with properties shared by bilinear forms and sesquilinear forms in a uniform fashion. Orthogonality is such a property.

Definition 29.12. Given a sesquilinear form $\varphi\colon E \times F \to K$, we say that two vectors $u \in E$ and $v \in F$ are *orthogonal* (or *conjugate*) if $\varphi(u, v) = 0$. Two subsets $E' \subseteq E$ and $F' \subseteq F$ are *orthogonal* if $\varphi(u, v) = 0$ for all $u \in E'$ and all $v \in F'$. Given a subspace U of E, the *right orthogonal space* of U, denoted $U^{\perp}$, is the subspace of F given by

$$U^{\perp} = \{v \in F \mid \varphi(u, v) = 0 \quad \text{for all } u \in U\},$$

and given a subspace V of F, the *left orthogonal space* of V, denoted $V^{\perp}$, is the subspace of E given by

$$V^{\perp} = \{u \in E \mid \varphi(u, v) = 0 \quad \text{for all } v \in V\}.$$

When E and F are distinct, there is little chance of confusing the right orthogonal subspace $U^{\perp}$ of a subspace U of E and the left orthogonal subspace $V^{\perp}$ of a subspace V of F. However, if $E = F$, then $\varphi(u, v) = 0$ *does not necessarily imply* that $\varphi(v, u) = 0$, that is, orthogonality is not necessarily symmetric. Thus, if both U and V are subsets of E, there is a notational ambiguity if $U = V$. In this case, we may write $U^{\perp_r}$ for the right orthogonal and $U^{\perp_l}$ for the left orthogonal.

The above discussion brings up the following point: When is orthogonality symmetric?

If φ is bilinear, it is shown in E. Artin [6] (and in Jacobson [98]) that orthogonality is symmetric iff either φ is symmetric or φ is alternating ($\varphi(u, u) = 0$ for all $u \in E$).

If φ is sesquilinear, the answer is more complicated. In addition to the previous two cases, there is a third possibility:

$$\varphi(u, v) = \epsilon\overline{\varphi(v, u)} \quad \text{for all } u, v \in E,$$

where ϵ is some nonzero element in K. We say that φ is *ϵ-Hermitian*. Observe that

$$\varphi(u, u) = \epsilon\overline{\epsilon}\varphi(u, u),$$

so if φ is not alternating, then $\varphi(u, u) \neq 0$ for some u, and we must have $\epsilon\overline{\epsilon} = 1$. The most common cases are

1. $\epsilon = 1$, in which case φ is *Hermitian*, and

2. $\epsilon = -1$, in which case φ is *skew-Hermitian*.

If φ is alternating and K is not of characteristic 2, then equation $(*)$ from Section 29.2 implies that the automorphism $\lambda \mapsto \overline{\lambda}$ must be the identity if φ is nonzero. If so, φ is skew-symmetric, so $\epsilon = -1$.

In summary, if φ is either symmetric, alternating, or ϵ-Hermitian, then orthogonality is symmetric, and it makes sense to talk about *the* orthogonal subspace $U^{\perp}$ of U.

Observe that if φ is ϵ-Hermitian, then

$$r_{\varphi} = \epsilon l_{\varphi}.$$

This is because

$$\begin{aligned} l_{\varphi}(u)(y) &= \overline{\varphi(u, y)} \\ r_{\varphi}(u)(y) &= \varphi(y, u) \\ &= \epsilon\overline{\varphi(u, y)}, \end{aligned}$$

so $r_{\varphi} = \epsilon l_{\varphi}$.

If E and F are finite-dimensional with bases $(e_1, \ldots, e_m)$ and $(f_1, \ldots, f_n)$, and if φ is represented by the $n \times m$ matrix M, then φ is ϵ-Hermitian iff

$$M = \epsilon M^*,$$

where $M^* = (\overline{M})^{\top}$ (as usual). This captures the following kinds of familiar matrices:

1. Symmetric matrices ($\epsilon = 1$)

2. Skew-symmetric matrices ($\epsilon = -1$)

3. Hermitian matrices ($\epsilon = 1$)

4. Skew-Hermitian matrices ($\epsilon = -1$).

Going back to a sesquilinear form $\varphi\colon E \times F \to K$, for any subspace U of E, it is easy to check that
$$U \subseteq (U^{\perp})^{\perp},$$
and that for any subspace V of F, we have
$$V \subseteq (V^{\perp})^{\perp}.$$
For simplicity of notation, we write $U^{\perp\perp}$ instead of $(U^{\perp})^{\perp}$ (and $V^{\perp\perp}$ instead of $(V^{\perp})^{\perp}$).

Given any two subspaces U_1 and U_2 of E, if $U_1 \subseteq U_2$, then $U_2^{\perp} \subseteq U_1^{\perp}$. Indeed, if $v \in U_2^{\perp}$ then $\varphi(u_2, v) = 0$ for all $u_2 \in U_2$, and since $U_1 \subseteq U_2$ this implies that $\varphi(u_1, v) = 0$ for all $u_1 \in U_1$, which shows that $v \in U_1^{\perp}$. Similarly for any two subspaces V_1, V_2 of F, if $V_1 \subseteq V_2$, then $V_2^{\perp} \subseteq V_1^{\perp}$. As a consequence,
$$U^{\perp} = U^{\perp\perp\perp}, \quad V^{\perp} = V^{\perp\perp\perp}.$$
First, we have $U^{\perp} \subseteq U^{\perp\perp\perp}$. Second, from $U \subseteq U^{\perp\perp}$, we get $U^{\perp\perp\perp} \subseteq U^{\perp}$, so $U^{\perp} = U^{\perp\perp\perp}$. The other equation is proved is a similar way.

Observe that φ is nondegenerate iff $E^{\perp} = \{0\}$ and $F^{\perp} = \{0\}$. Furthermore, since
$$\begin{aligned}\varphi(u + x, v) &= \varphi(u, v)\\ \varphi(u, v + y) &= \varphi(u, v)\end{aligned}$$
for any $x \in F^{\perp}$ and any $y \in E^{\perp}$, we see that we obtain by passing to the quotient a sesquilinear form
$$[\varphi]\colon (E/F^{\perp}) \times (F/E^{\perp}) \to K$$
which is nondegenerate.

Proposition 29.12. *For any sesquilinear form $\varphi\colon E \times F \to K$, the space $E/F^{\perp}$ is finite-dimensional iff the space $F/E^{\perp}$ is finite-dimensional; if so,* $\dim(E/F^{\perp}) = \dim(F/E^{\perp})$.

Proof. Since the sesquilinear form $[\varphi]\colon (E/F^{\perp}) \times (F/E^{\perp}) \to K$ is nondegenerate, the maps $l_{[\varphi]}\colon (E/F^{\perp}) \to \overline{(F/E^{\perp})^*}$ and $r_{[\varphi]}\colon (F/E^{\perp}) \to \overline{(E/F^{\perp})^*}$ are injective. If $\dim(E/F^{\perp}) = m$, then $\dim(E/F^{\perp}) = \dim((E/F^{\perp})^*)$, so by injectivity of $r_{[\varphi]}$, we have $\dim(F/E^{\perp}) = \dim(\overline{(F/E^{\perp})}) \leq m$. A similar reasoning using the injectivity of $l_{[\varphi]}$ applies if $\dim(F/E^{\perp}) = n$, and we get $\dim(E/F^{\perp}) = \dim(\overline{(E/F^{\perp})}) \leq n$. Therefore, $\dim(E/F^{\perp}) = m$ is finite iff $\dim(F/E^{\perp}) = n$ is finite, in which case $m = n$ by Proposition 29.1(d). $\square$

If U is a subspace of a space E, recall that the *codimension* of U is the dimension of E/U, which is also equal to the dimension of any subspace V such that E is a direct sum of U and V ($E = U \oplus V$).

Proposition 29.12 implies the following useful fact.

Proposition 29.13. *Let $\varphi\colon E \times F \to K$ be any nondegenerate sesquilinear form. A subspace U of E has finite dimension iff $U^\perp$ has finite codimension in F. If $\dim(U)$ is finite, then $\mathrm{codim}(U^\perp) = \dim(U)$, and $U^{\perp\perp} = U$.*

Proof. Since φ is nondegenerate $E^\perp = \{0\}$ and $F^\perp = \{0\}$, so Proposition 29.12 applied to the restriction of φ to $U \times F$ implies that a subspace U of E has finite dimension iff $U^\perp$ has finite codimension in F, and that if $\dim(U)$ is finite, then $\mathrm{codim}(U^\perp) = \dim(U)$. Since $U^\perp$ and $U^{\perp\perp}$ are orthogonal, and since $\mathrm{codim}(U^\perp)$ is finite, $\dim(U^{\perp\perp})$ is finite and we have $\dim(U^{\perp\perp}) = \mathrm{codim}(U^{\perp\perp\perp}) = \mathrm{codim}(U^\perp) = \dim(U)$. Since $U \subseteq U^{\perp\perp}$, we must have $U = U^{\perp\perp}$. $\square$

Proposition 29.14. *Let $\varphi\colon E \times F \to K$ be any sesquilinear form. Given any two subspaces U and V of E, we have*

$$(U + V)^\perp = U^\perp \cap V^\perp.$$

Furthermore, if φ is nondegenerate and if U and V are finite-dimensional, then

$$(U \cap V)^\perp = U^\perp + V^\perp.$$

Proof. If $w \in (U + V)^\perp$, then $\varphi(u + v, w) = 0$ for all $u \in U$ and all $v \in V$. In particular, with $v = 0$, we have $\varphi(u, w) = 0$ for all $u \in U$, and with $u = 0$, we have $\varphi(v, w) = 0$ for all $v \in V$, so $w \in U^\perp \cap V^\perp$. Conversely, if $w \in U^\perp \cap V^\perp$, then $\varphi(u, w) = 0$ for all $u \in U$ and $\varphi(v, w) = 0$ for all $v \in V$. By bilinearity, $\varphi(u + v, w) = \varphi(u, w) + \varphi(v, w) = 0$, which shows that $w \in (U + V)^\perp$. Therefore, the first identity holds.

Now, assume that φ is nondegenerate and that U and V are finite-dimensional, and let $W = U^\perp + V^\perp$. Using the equation that we just established and the fact that U and V are finite-dimensional, by Proposition 29.13, we get

$$W^\perp = U^{\perp\perp} \cap V^{\perp\perp} = U \cap V.$$

We can apply Proposition 29.12 to the restriction of φ to $U \times W$ (since $U^\perp \subseteq W$ and $W^\perp \subseteq U$), and we get

$$\dim(U/W^\perp) = \dim(U/(U \cap V)) = \dim(W/U^\perp).$$

If T is a supplement of $U^\perp$ in W so that $W = U^\perp \oplus T$ and if S is a supplement of W in E so that $E = W \oplus S$, then $\mathrm{codim}(W) = \dim(S)$, $\dim(T) = \dim(W/U^\perp)$, and we have the direct sum

$$E = U^\perp \oplus T \oplus S$$

which implies that

$$\dim(T) = \operatorname{codim}(U^{\perp}) - \dim(S) = \operatorname{codim}(U^{\perp}) - \operatorname{codim}(W)$$

so

$$\dim(U/(U \cap V)) = \dim(W/U^{\perp}) = \operatorname{codim}(U^{\perp}) - \operatorname{codim}(W),$$

and since $\operatorname{codim}(U^{\perp}) = \dim(U)$, we deduce that

$$\dim(U \cap V) = \operatorname{codim}(W).$$

However, by Proposition 29.13, we have $\dim(U \cap V) = \operatorname{codim}((U \cap V)^{\perp})$, so $\operatorname{codim}(W) = \operatorname{codim}((U \cap V)^{\perp})$, and since $W \subseteq W^{\perp\perp} = (U \cap V)^{\perp}$, we must have $W = (U \cap V)^{\perp}$, as claimed. ☐

In view of Proposition 29.12, we can make the following definition.

Definition 29.13. Let $\varphi\colon E \times F \to K$ be any sesquilinear form. If $E/F^{\perp}$ and $F/E^{\perp}$ are finite-dimensional, then their common dimension is called the *rank* of the form φ. If $E/F^{\perp}$ and $F/E^{\perp}$ have infinite dimension, we say that φ has infinite rank.

Not surprisingly, the rank of φ is related to the ranks of l_{φ} and r_{φ}.

Proposition 29.15. *Let $\varphi\colon E \times F \to K$ be any sesquilinear form. If φ has finite rank r, then l_{φ} and r_{φ} have the same rank, which is equal to r.*

Proof. Because for every $u \in E$,

$$l_{\varphi}(u)(y) = \overline{\varphi(u, y)} \quad \text{for all } y \in F,$$

and for every $v \in F$,

$$r_{\varphi}(v)(x) = \varphi(x, v) \quad \text{for all } x \in E,$$

it is clear that the kernel of $l_{\varphi}\colon \overline{E} \to F^{*}$ is equal to $F^{\perp}$ and that, the kernel of $r_{\varphi}\colon \overline{F} \to E^{*}$ is equal to $E^{\perp}$. Therefore, $\operatorname{rank}(l_{\varphi}) = \dim(\operatorname{Im} l_{\varphi}) = \dim(E/F^{\perp}) = r$, and similarly $\operatorname{rank}(r_{\varphi}) = \dim(F/E^{\perp}) = r$. ☐

Remark: If the sesquilinear form φ is represented by the matrix $n \times m$ matrix M with respect to the bases $(e_1, \ldots, e_m)$ in E and $(f_1, \ldots, f_n)$ in F, it can be shown that the matrix representing l_{φ} with respect to the bases $(e_1, \ldots, e_m)$ and $(f_1^{*}, \ldots, f_n^{*})$ is $\overline{M}$, and that the matrix representing r_{φ} with respect to the bases $(f_1, \ldots, f_n)$ and $(e_1^{*}, \ldots, e_m^{*})$ is $M^{\top}$. It follows that the rank of φ is equal to the rank of M.

29.4 Adjoint of a Linear Map

Let E_1 and E_2 be two K-vector spaces, and let $\varphi_1 \colon E_1 \times E_1 \to K$ be a sesquilinear form on E_1 and $\varphi_2 \colon E_2 \times E_2 \to K$ be a sesquilinear form on E_2. It is also possible to deal with the more general situation where we have four vector spaces E_1, F_1, E_2, F_2 and two sesquilinear forms $\varphi_1 \colon E_1 \times F_1 \to K$ and $\varphi_2 \colon E_2 \times F_2 \to K$, but we will leave this generalization as an exercise. We also assume that l_{φ_1} and r_{φ_1} are bijective, which implies that that φ_1 is nondegenerate. This is automatic if the space E_1 is finite dimensional and φ_1 is nondegenerate.

Given any linear map $f \colon E_1 \to E_2$, for any fixed $u \in E_2$, we can consider the linear form in E_1^* given by

$$x \mapsto \varphi_2(f(x), u), \quad x \in E_1.$$

Since $r_{\varphi_1} \colon \overline{E_1} \to E_1^*$ is bijective, there is a unique $y \in E_1$ (because the vector spaces E_1 and $\overline{E_1}$ only differ by scalar multiplication), so that

$$\varphi_2(f(x), u) = \varphi_1(x, y), \quad \text{for all } x \in E_1.$$

If we denote this unique $y \in E_1$ by $f^{*l}(u)$, then we have

$$\varphi_2(f(x), u) = \varphi_1(x, f^{*l}(u)), \quad \text{for all } x \in E_1, \text{ and all } u \in E_2.$$

Thus, we get a function $f^{*l} \colon E_2 \to E_1$. We claim that this function is a linear map. For any $v_1, v_2 \in E_2$, we have

$$\begin{aligned}
\varphi_2(f(x), v_1 + v_2) &= \varphi_2(f(x), v_1) + \varphi_2(f(x), v_2) \\
&= \varphi_1(x, f^{*l}(v_1)) + \varphi_1(x, f^{*l}(v_2)) \\
&= \varphi_1(x, f^{*l}(v_1) + f^{*l}(v_2)) \\
&= \varphi_1(x, f^{*l}(v_1 + v_2)),
\end{aligned}$$

for all $x \in E_1$. Since r_{φ_1} is injective, we conclude that

$$f^{*l}(v_1 + v_2) = f^{*l}(v_1) + f^{*l}(v_2).$$

For any $\lambda \in K$, we have

$$\begin{aligned}
\varphi_2(f(x), \lambda v) &= \overline{\lambda}\varphi_2(f(x), v) \\
&= \overline{\lambda}\varphi_1(x, f^{*l}(v)) \\
&= \varphi_1(x, \lambda f^{*l}(v)) \\
&= \varphi_1(x, f^{*l}(\lambda v)),
\end{aligned}$$

for all $x \in E_1$. Since r_{φ_1} is injective, we conclude that

$$f^{*l}(\lambda v) = \lambda f^{*l}(v).$$

Therefore, f^{*l} is linear. We call it the *left adjoint* of f.

Now, for any fixed $u \in E_2$, we can consider the linear form in E_1^* given by

$$x \mapsto \overline{\varphi_2(u, f(x))} \quad x \in E_1.$$

Since $l_{\varphi_1} \colon \overline{E_1} \to E_1^*$ is bijective, there is a unique $y \in E_1$ so that

$$\overline{\varphi_2(u, f(x))} = \overline{\varphi_1(y, x)}, \quad \text{for all } x \in E_1.$$

If we denote this unique $y \in E_1$ by $f^{*r}(u)$, then we have

$$\varphi_2(u, f(x)) = \varphi_1(f^{*r}(u), x), \quad \text{for all } x \in E_1, \text{ and all } u \in E_2.$$

Thus, we get a function $f^{*r} \colon E_2 \to E_1$. As in the previous situation, it easy to check that f^{*r} is linear. We call it the *right adjoint* of f. In summary, we make the following definition.

Definition 29.14. Let E_1 and E_2 be two K-vector spaces, and let $\varphi_1 \colon E_1 \times E_1 \to K$ and $\varphi_2 \colon E_2 \times E_2 \to K$ be two sesquilinear forms. Assume that l_{φ_1} and r_{φ_1} are bijective, so that φ_1 is nondegnerate. For every linear map $f \colon E_1 \to E_2$, there exist unique linear maps $f^{*l} \colon E_2 \to E_1$ and $f^{*r} \colon E_2 \to E_1$, such that

$$\begin{aligned} \varphi_2(f(x), u) &= \varphi_1(x, f^{*l}(u)), \quad \text{for all } x \in E_1, \text{ and all } u \in E_2 \\ \varphi_2(u, f(x)) &= \varphi_1(f^{*r}(u), x), \quad \text{for all } x \in E_1, \text{ and all } u \in E_2. \end{aligned}$$

The map f^{*l} is called the *left adjoint* of f, and the map f^{*r} is called the *right adjoint* of f.

If E_1 and E_2 are finite-dimensional with bases $(e_1, \ldots, e_m)$ and $(f_1, \ldots, f_n)$, then we can work out the matrices A^{*l} and A^{*r} corresponding to the left adjoint f^{*l} and the right adjoint f^{*r} of f. Assumine that f is represented by the $n \times m$ matrix A, φ_1 is represented by the $m \times m$ matrix M_1, and φ_2 is represented by the $n \times n$ matrix M_2. Since

$$\begin{aligned} \varphi_1(x, f^{*l}(u)) &= (A^{*l}u)^* M_1 x = u^* (A^{*l})^* M_1 x \\ \varphi_2(f(x), u) &= u^* M_2 A x \end{aligned}$$

we find that $(A^{*l})^* M_1 = M_2 A$, that is $(A^{*l})^* = M_2 A M_1^{-1}$, and similarly

$$\begin{aligned} \varphi_1(f^{*r}(u), x) &= x^* M_1 A^{*r} u \\ \varphi_2(u, f(x)) &= (Ax)^* M_2 u = x^* A^* M_2 u, \end{aligned}$$

we have $M_1 A^{*r} = A^* M_2$, that is $A^{*r} = (M_1)^{-1} A^* M_2$. Thus, we obtain

$$\begin{aligned} A^{*l} &= (M_1^*)^{-1} A^* M_2^* \\ A^{*r} &= (M_1)^{-1} A^* M_2. \end{aligned}$$

If φ_1 and φ_2 are symmetric bilinear forms, then $f^{*l} = f^{*r}$. This also holds if φ is ϵ-Hermitian. Indeed, since

$$\varphi_2(u, f(x)) = \varphi_1(f^{*r}(u), x),$$

we get

$$\epsilon\overline{\varphi_2(f(x), u)} = \epsilon\overline{\varphi_1(x, f^{*r}(u))},$$

and since $\lambda \mapsto \overline{\lambda}$ is an involution, we get

$$\varphi_2(f(x), u) = \varphi_1(x, f^{*r}(u)).$$

Since we also have

$$\varphi_2(f(x), u) = \varphi_1(x, f^{*l}(u)),$$

we obtain

$$\varphi_1(x, f^{*r}(u)) = \varphi_1(x, f^{*l}(u)) \quad \text{for all } x \in E_1, \text{ and all } u \in E_2,$$

and since φ_1 is nondegenerate, we conclude that $f^{*l} = f^{*r}$. Whenever $f^{*l} = f^{*r}$, we use the simpler notation f^*.

If $f\colon E_1 \to E_2$ and $g\colon E_1 \to E_2$ are two linear maps, we have the following properties:

$$\begin{aligned}(f+g)^{*l} &= f^{*l} + g^{*l}\\ \mathrm{id}^{*l} &= \mathrm{id}\\ (\lambda f)^{*l} &= \overline{\lambda} f^{*l},\end{aligned}$$

and similarly for right adjoints. If E_3 is another space, φ_3 is a sesquilinear form on E_3, and if l_{φ_2} and r_{φ_2} are bijective, then for any linear maps $f\colon E_1 \to E_2$ and $g\colon E_2 \to E_3$, we have

$$(g \circ f)^{*l} = f^{*l} \circ g^{*l},$$

and similarly for right adjoints. Furthermore, if $E_1 = E_2 = E$ and $\varphi\colon E \times E \to K$ is ϵ-Hermitian, for any linear map $f\colon E \to E$ (recall that in this case $f^{*l} = f^{*r} = f^*$), we have

$$f^{**} = \epsilon\overline{\epsilon} f.$$

29.5 Isometries Associated with Sesquilinear Forms

The notion of adjoint is a good tool to investigate the notion of isometry between spaces equipped with sesquilinear forms. First, we define metric maps and isometries.

Definition 29.15. If (E_1, φ_1) and (E_2, φ_2) are two pairs of spaces and sesquilinear maps $\varphi_1\colon E_1 \times E_1 \to K$ and $\varphi_2\colon E_2 \times E_2 \to K$, a *metric map* from (E_1, φ_1) to (E_2, φ_2) is a linear map $f\colon E_1 \to E_2$ such that

$$\varphi_1(u, v) = \varphi_2(f(u), f(v)) \quad \text{for all } u, v \in E_1.$$

We say that φ_1 and φ_2 are *equivalent* iff there is a metric map $f\colon E_1 \to E_2$ which is bijective. Such a metric map is called an *isometry*.

The problem of classifying sesquilinear forms up to equivalence is an important but very difficult problem. Solving this problem depends intimately on properties of the field K, and a complete answer is only known in a few cases. The problem is easily solved for $K = \mathbb{R}$, $K = \mathbb{C}$. It is also solved for finite fields and for $K = \mathbb{Q}$ (the rationals), but the solution is surprisingly involved!

It is hard to say anything interesting if φ_1 is degenerate and if the linear map f does not have adjoints. The next few propositions make use of natural conditions on φ_1 that yield a useful criterion for being a metric map.

Proposition 29.16. *With the same assumptions as in Definition 29.14 (which imply that φ_1 is nondegenerate), if $f\colon E_1 \to E_2$ is a bijective linear map, then we have*

$$\begin{aligned}&\varphi_1(x, y) = \varphi_2(f(x), f(y)) \quad \text{for all } x, y \in E_1 \text{ iff}\\ &f^{-1} = f^{*l} = f^{*r}.\end{aligned}$$

Proof. We have

$$\varphi_1(x, y) = \varphi_2(f(x), f(y))$$

iff

$$\varphi_1(x, y) = \varphi_2(f(x), f(y)) = \varphi_1(x, f^{*l}(f(y))$$

iff

$$\varphi_1(x, (\mathrm{id} - f^{*l} \circ f)(y)) = 0 \quad \text{for all } x \in E_1 \text{ and all } y \in E_2.$$

Since φ_1 is nondegenerate, we must have

$$f^{*l} \circ f = \mathrm{id},$$

which implies that $f^{-1} = f^{*l}$. Similarly,

$$\varphi_1(x, y) = \varphi_2(f(x), f(y))$$

iff

$$\varphi_1(x, y) = \varphi_2(f(x), f(y)) = \varphi_1(f^{*r}(f(x)), y)$$

iff

$$\varphi_1((\mathrm{id} - f^{*r} \circ f)(x), y) = 0 \quad \text{for all } x \in E_1 \text{ and all } y \in E_2.$$

Since φ_1 is nondegenerate, we must have

$$f^{*r} \circ f = \mathrm{id},$$

which implies that $f^{-1} = f^{*r}$. Therefore, $f^{-1} = f^{*l} = f^{*r}$. For the converse, do the computations in reverse. ∎

As a corollary, we get the following important proposition.

Proposition 29.17. *If $\varphi\colon E \times E \to K$ is a sesquilinear map, and if l_φ and r_φ are bijective, for every bijective linear map $f\colon E \to E$, then we have*

$$\begin{aligned}&\varphi(f(x), f(y)) = \varphi(x, y) \quad \textit{for all } x, y \in E \textit{ iff}\\ &f^{-1} = f^{*l} = f^{*r}.\end{aligned}$$

We also have the following facts.

Proposition 29.18. *(1) If $\varphi\colon E \times E \to K$ is a sesquilinear map and if l_φ is injective, then for every linear map $f\colon E \to E$, if*

$$\varphi(f(x), f(y)) = \varphi(x, y) \quad \textit{for all } x, y \in E, \tag{$*$}$$

then f is injective.

(2) If E is finite-dimensional and if φ is nondegenerate, then the linear maps $f\colon E \to E$ satisfying $(*)$ form a group. The inverse of f is given by $f^{-1} = f^*$.

Proof. (1) If $f(x) = 0$, then

$$\varphi(x, y) = \varphi(f(x), f(y)) = \varphi(0, f(y)) = 0 \quad \text{for all } y \in E.$$

Since l_φ is injective, we must have $x = 0$, and thus f is injective.

(2) If E is finite-dimensional, since a linear map satisfying $(*)$ is injective, it is a bijection. By Proposition 29.17, we have $f^{-1} = f^*$. We also have

$$\varphi(f(x), f(y)) = \varphi((f^* \circ f)(x), y) = \varphi(x, y) = \varphi((f \circ f^*)(x), y) = \varphi(f^*(x), f^*(y)),$$

which shows that f^* satisfies $(*)$. If $\varphi(f(x), f(y)) = \varphi(x, y)$ for all $x, y \in E$ and $\varphi(g(x), g(y)) = \varphi(x, y)$ for all $x, y \in E$, then we have

$$\varphi((g \circ f)(x), (g \circ f)(y)) = \varphi(f(x), f(y)) = \varphi(x, y) \quad \text{for all } x, y \in E.$$

Obviously, the identity map id_E satisfies $(*)$. Therefore, the set of linear maps satisfying $(*)$ is a group. $\square$

The above considerations motivate the following definition.

Definition 29.16. Let $\varphi\colon E \times E \to K$ be a sesquilinear map, and assume that E is finite-dimensional and that φ is nondegenerate. A linear map $f\colon E \to E$ is an *isometry* of E (with respect to φ) iff

$$\varphi(f(x), f(y)) = \varphi(x, y) \quad \text{for all } x, y \in E.$$

The set of all isometries of E is a group denoted by $\mathbf{Isom}(\varphi)$.

If φ is symmetric, then the group $\mathbf{Isom}(\varphi)$ is denoted $\mathbf{O}(\varphi)$ and called the *orthogonal group* of φ. If φ is alternating, then the group $\mathbf{Isom}(\varphi)$ is denoted $\mathbf{Sp}(\varphi)$ and called the *symplectic group* of φ. If φ is ϵ-Hermitian, then the group $\mathbf{Isom}(\varphi)$ is denoted $\mathbf{U}_\epsilon(\varphi)$ and called the *ϵ-unitary group* of φ. When $\epsilon = 1$, we drop ϵ and just say *unitary group*.

If $(e_1, \ldots, e_n)$ is a basis of E, φ is the represented by the $n \times n$ matrix M, and f is represented by the $n \times n$ matrix A, since $A^{-1} = A^{*l} = A^{*r} = M^{-1}A^*M$, then we find that $f \in \mathbf{Isom}(\varphi)$ iff

$$A^*MA = M,$$

and A^{-1} is given by $A^{-1} = M^{-1}A^*M$.

More specifically, we define the following groups, using the matrices $I_{p,q}$, $J_{m,m}$ and $A_{m,m}$ defined at the end of Section 29.1.

(1) $K = \mathbb{R}$. We have

$$\begin{aligned}
\mathbf{O}(n) &= \{A \in \mathrm{M}_n(\mathbb{R}) \mid A^\top A = I_n\}\\
\mathbf{O}(p,q) &= \{A \in \mathrm{M}_{p+q}(\mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}\}\\
\mathbf{Sp}(2n,\mathbb{R}) &= \{A \in \mathrm{M}_{2n}(\mathbb{R}) \mid A^\top J_{n,n} A = J_{n,n}\}\\
\mathbf{SO}(n) &= \{A \in \mathrm{M}_n(\mathbb{R}) \mid A^\top A = I_n,\ \det(A) = 1\}\\
\mathbf{SO}(p,q) &= \{A \in \mathrm{M}_{p+q}(\mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q},\ \det(A) = 1\}.
\end{aligned}$$

The group $\mathbf{O}(n)$ is the *orthogonal group*, $\mathbf{Sp}(2n,\mathbb{R})$ is the *real symplectic group*, and $\mathbf{SO}(n)$ is the *special orthogonal group*. We can define the group

$$\{A \in \mathrm{M}_{2n}(\mathbb{R}) \mid A^\top A_{n,n} A = A_{n,n}\},$$

but it is isomorphic to $\mathbf{O}(n,n)$.

(2) $K = \mathbb{C}$. We have

$$\begin{aligned}
\mathbf{U}(n) &= \{A \in \mathrm{M}_n(\mathbb{C}) \mid A^* A = I_n\}\\
\mathbf{U}(p,q) &= \{A \in \mathrm{M}_{p+q}(\mathbb{C}) \mid A^* I_{p,q} A = I_{p,q}\}\\
\mathbf{Sp}(2n,\mathbb{C}) &= \{A \in \mathrm{M}_{2n}(\mathbb{C}) \mid A^\top J_{n,n} A = J_{n,n}\}\\
\mathbf{SU}(n) &= \{A \in \mathrm{M}_n(\mathbb{C}) \mid A^* A = I_n,\ \det(A) = 1\}\\
\mathbf{SU}(p,q) &= \{A \in \mathrm{M}_{p+q}(\mathbb{C}) \mid A^* I_{p,q} A = I_{p,q},\ \det(A) = 1\}.
\end{aligned}$$

The group $\mathbf{U}(n)$ is the *unitary group*, $\mathbf{Sp}(2n,\mathbb{C})$ is the *complex symplectic group*, and $\mathbf{SU}(n)$ is the *special unitary group*.

It can be shown that if $A \in \mathbf{Sp}(2n,\mathbb{R})$ or if $A \in \mathbf{Sp}(2n,\mathbb{C})$, then $\det(A) = 1$.

29.6 Totally Isotropic Subspaces

In this section, we deal with ϵ-Hermitian forms, $\varphi\colon E \times E \to K$. In general, E may have subspaces U such that $U \cap U^{\perp} \neq (0)$, or worse, such that $U \subseteq U^{\perp}$ (that is, φ is zero on U). We will see that such subspaces play a crucial in the decomposition of E into orthogonal subspaces.

Definition 29.17. Given an ϵ-Hermitian forms $\varphi\colon E \times E \to K$, a nonzero vector $u \in E$ is said to be *isotropic* if $\varphi(u, u) = 0$. It is convenient to consider 0 to be isotropic. Given any subspace U of E, the subspace $\mathrm{rad}(U) = U \cap U^{\perp}$ is called the *radical* of U. We say that

(i) U is *degenerate* if $\mathrm{rad}(U) \neq (0)$ (equivalently if there is some nonzero vector $u \in U$ such that $x \in U^{\perp}$). Otherwise, we say that U is *nondegenerate*.

(ii) U is *totally isotropic* if $U \subseteq U^{\perp}$ (equivalently if the restriction of φ to U is zero).

By definition, the trivial subspace $U = (0)$ $(= \{0\})$ is nondegenerate. Observe that a subspace U is nondegenerate iff the restriction of φ to U is nondegenerate. A degenerate subspace is sometimes called an *isotropic* subspace. Other authors say that a subspace U is *isotropic* if it contains some (nonzero) isotropic vector. A subspace which has no nonzero isotropic vector is often called *anisotropic*. The space of all isotropic vectors is a cone often called the *light cone* (a terminology coming from the theory of relativity). This is not to be confused with the cone of silence (from Get Smart)! It should also be noted that some authors (such as Serre) use the term *isotropic* instead of *totally isotropic*. The apparent lack of standard terminology is almost as bad as in graph theory!

It is clear that any direct sum of pairwise orthogonal totally isotropic subspaces is totally isotropic. Thus, every totally isotropic subspace is contained in some maximal totally isotropic subspace. Here is another fact that we will use all the time: if V is a totally isotropic subspace and if U is a subspace of V, then U is totally isotropic.

This is because by definition V is isotropic if $V \subseteq V^{\perp}$, and since $U \subseteq V$ we get $V^{\perp} \subseteq U^{\perp}$, so $U \subseteq V \subseteq V^{\perp} \subseteq U^{\perp}$, which shows that U is totally isotropic.

First, let us show that in order to sudy an ϵ-Hermitian form on a space E, it suffices to restrict our attention to nondegenerate forms.

Proposition 29.19. *Given an ϵ-Hermitian form $\varphi\colon E \times E \to K$ on E, we have:*

(a) If U and V are any two orthogonal subspaces of E, then

$$\mathrm{rad}(U + V) = \mathrm{rad}(U) + \mathrm{rad}(V).$$

(b) $\mathrm{rad}(\mathrm{rad}(E)) = \mathrm{rad}(E)$.

(c) *If* U *is any subspace supplementary to* $\mathrm{rad}(E)$*, so that*

$$E = \mathrm{rad}(E) \oplus U,$$

then U *is nondegenerate, and* $\mathrm{rad}(E)$ *and* U *are orthogonal.*

Proof. (a) If U and V are orthogonal, then $U \subseteq V^\perp$ and $V \subseteq U^\perp$. We get

$$\begin{aligned}
\mathrm{rad}(U + V) &= (U + V) \cap (U + V)^\perp \\
&= (U + V) \cap U^\perp \cap V^\perp \\
&= U \cap U^\perp \cap V^\perp + V \cap U^\perp \cap V^\perp \\
&= U \cap U^\perp + V \cap V^\perp \\
&= \mathrm{rad}(U) + \mathrm{rad}(V).
\end{aligned}$$

(b) By definition, $\mathrm{rad}(E) = E^\perp$, and obviously $E = E^{\perp\perp}$, so we get

$$\mathrm{rad}(\mathrm{rad}(E)) = E^\perp \cap E^{\perp\perp} = E^\perp \cap E = E^\perp = \mathrm{rad}(E).$$

(c) If $E = \mathrm{rad}(E) \oplus U$, by definition of $\mathrm{rad}(E)$, the subspaces $\mathrm{rad}(E)$ and U are orthogonal. From (a) and (b), we get

$$\mathrm{rad}(E) = \mathrm{rad}(E) + \mathrm{rad}(U).$$

Since $\mathrm{rad}(U) = U \cap U^\perp \subseteq U$ and since $\mathrm{rad}(E) \oplus U$ is a direct sum, we have a direct sum

$$\mathrm{rad}(E) = \mathrm{rad}(E) \oplus \mathrm{rad}(U),$$

which implies that $\mathrm{rad}(U) = (0)$; that is, U is nondegenerate. $\square$

Proposition 29.19(c) shows that the restriction of φ to any supplement U of $\mathrm{rad}(E)$ is nondegenerate and φ is zero on $\mathrm{rad}(U)$, so we may restrict our attention to nondegenerate forms.

The following is also a key result.

Proposition 29.20. *Given an* ϵ*-Hermitian form* $\varphi\colon E \times E \to K$ *on* E*, if* U *is a finite-dimensional nondegenerate subspace of* E*, then* $E = U \oplus U^\perp$*.*

Proof. By hypothesis, the restriction φ_U of φ to U is nondegenerate, so the semilinear map $r_{\varphi_U}\colon U \to U^*$ is injective. Since U is finite-dimensional, r_{φ_U} is actually bijective, so for every $v \in E$, if we consider the linear form in U^* given by $u \mapsto \varphi(u, v)$ ($u \in U$), there is a unique $v_0 \in U$ such that

$$\varphi(u, v_0) = \varphi(u, v) \quad \text{for all } u \in U;$$

that is, $\varphi(u, v - v_0) = 0$ for all $u \in U$, so $v - v_0 \in U^\perp$. It follows that $v = v_0 + v - v_0$, with $v_0 \in U$ and $v_0 - v \in U^\perp$, and since U is nondegenerate $U \cap U^\perp = (0)$, and $E = U \oplus U^\perp$. $\square$

As a corollary of Proposition 29.20, we get the following result.

Proposition 29.21. *Given an ϵ-Hermitian form $\varphi\colon E\times E \to K$ on E, if φ is nondegenerate and if U is a finite-dimensional subspace of E, then* $\mathrm{rad}(U) = \mathrm{rad}(U^{\perp})$, *and the following conditions are equivalent:*

(i) U *is nondegenerate.*

(ii) $U^{\perp}$ *is nondegenerate.*

(iii) $E = U \oplus U^{\perp}$.

Proof. By definition, $\mathrm{rad}(U^{\perp}) = U^{\perp} \cap U^{\perp\perp}$, and since φ is nondegenerate and U is finite-dimensional, $U^{\perp\perp} = U$, so $\mathrm{rad}(U^{\perp}) = U^{\perp} \cap U^{\perp\perp} = U \cap U^{\perp} = \mathrm{rad}(U)$.

By Proposition 29.20, (i) implies (iii). If $E = U \oplus U^{\perp}$, then $\mathrm{rad}(U) = U \cap U^{\perp} = (0)$, so U is nondegenerate and (iii) implies (i). Since $\mathrm{rad}(U^{\perp}) = \mathrm{rad}(U)$, (iii) also implies (ii). Now, if $U^{\perp}$ is nondegenerate, we have $U^{\perp} \cap U^{\perp\perp} = (0)$, and since $U \subseteq U^{\perp\perp}$, we get

$$U \cap U^{\perp} \subseteq U^{\perp\perp} \cap U^{\perp} = (0),$$

which shows that U is nondegenerate, proving the implication (ii) $\Longrightarrow$ (i). $\square$

If E is finite-dimensional, we have the following results.

Proposition 29.22. *Given an ϵ-Hermitian form $\varphi\colon E \times E \to K$ on a finite-dimensional space E, if φ is nondegenerate, then for every subspace U of E we have*

(i) $\dim(U) + \dim(U^{\perp}) = \dim(E)$.

(ii) $U^{\perp\perp} = U$.

Proof. (i) Since φ is nondegenerate and E is finite-dimensional, the semilinear map $l_{\varphi}\colon E \to E^*$ is bijective. By transposition, the inclusion $i\colon U \to E$ yields a surjection $r\colon E^* \to U^*$ (with $r(f) = f \circ i$ for every $f \in E^*$; the map $f \circ i$ is the restriction of the linear form f to U). It follows that the semilinear map $r \circ l_{\varphi}\colon E \to U^*$ given by

$$(r \circ l_{\varphi})(x)(u) = \overline{\varphi(x, u)} \quad x \in E, u \in U$$

is surjective, and its kernel is $U^{\perp}$. Thus, we have

$$\dim(U^*) + \dim(U^{\perp}) = \dim(E),$$

and since $\dim(U) = \dim(U^*)$ because U is finite-dimensional, we get

$$\dim(U) + \dim(U^{\perp}) = \dim(U^*) + \dim(U^{\perp}) = \dim(E).$$

(ii) Applying the above formula to $U^{\perp}$, we deduce that $\dim(U) = \dim(U^{\perp\perp})$. Since $U \subseteq U^{\perp\perp}$, we must have $U^{\perp\perp} = U$. $\square$

Remark: We already proved in Proposition 29.13 that if U is finite-dimensional, then $\mathrm{codim}(U^{\perp}) = \dim(U)$ and $U^{\perp\perp} = U$, but it doesn't hurt to give another proof. Observe that (i) implies that
$$\dim(U) + \dim(\mathrm{rad}(U)) \leq \dim(E).$$

We can now proceed with the Witt decomposition, but before that, we quickly take care of the structure theorem for alternating bilinear forms (the case where $\varphi(u, u) = 0$ for all $u \in E$). For an alternating bilinear form, the space E is totally isotropic. For example in dimension 2, the matrix
$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
defines the alternating form given by
$$\varphi((x_1, y_1), (x_2, y_2)) = x_1 y_2 - x_2 y_1.$$

This case is surprisingly general.

Proposition 29.23. *Let $\varphi\colon E \times E \to K$ be an alternating bilinear form on E. If $u, v \in E$ are two (nonzero) vectors such that $\varphi(u, v) = \lambda \neq 0$, then u and v are linearly independent. If we let $u_1 = \lambda^{-1} u$ and $v_1 = v$, then $\varphi(u_1, v_1) = 1$, and the restriction of φ to the plane spanned by u_1 and v_1 is represented by the matrix*
$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Proof. If u and v were linearly dependent, as $u, v \neq 0$, we could write $v = \mu u$ for some $\mu \neq 0$, but then, since φ is alternating, we would have
$$\lambda = \varphi(u, v) = \varphi(u, \mu u) = \mu \varphi(u, u) = 0,$$
contradicting the fact that $\lambda \neq 0$. The rest is obvious. $\square$

Proposition 29.23 yields a plane spanned by two vectors u_1, v_1 such that $\varphi(u_1, u_1) = \varphi(v_1, v_1) = 0$ and $\varphi(u_1, v_1) = 1$. Such a plane is called a *hyperbolic plane*. If E is finite-dimensional, we obtain the following theorem.

Theorem 29.24. *Let $\varphi\colon E \times E \to K$ be an alternating bilinear form on a space E of finite dimension n. Then, there is a direct sum decomposition of E into pairwise orthogonal subspaces*
$$E = W_1 \oplus \cdots \oplus W_r \oplus \mathrm{rad}(E),$$
where each W_i is a hyperbolic plane and $\mathrm{rad}(E) = E^{\perp}$. Therefore, there is a basis of E of the form
$$(u_1, v_1, \ldots, u_r, v_r, w_1, \ldots, w_{n-2r}),$$

with respect to which the matrix representing φ is a block diagonal matrix M of the form

$$M = \begin{pmatrix} J & & & & 0 \\ & J & & & \\ & & \ddots & & \\ & & & J & \\ 0 & & & & 0_{n-2r} \end{pmatrix},$$

with

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Proof. If $\varphi = 0$, then $E = E^{\perp}$ and we are done. Otherwise, there are two nonzero vectors $u, v \in E$ such that $\varphi(u, v) \neq 0$, so by Proposition 29.23, we obtain a hyperbolic plane W_2 spanned by two vectors u_1, v_1 such that $\varphi(u_1, v_1) = 1$. The subspace W_1 is nondegenerate (for example, $\det(J) = -1$), so by Proposition 29.21, we get a direct sum

$$E = W_1 \oplus W_1^{\perp}.$$

By Proposition 29.14, we also have

$$E^{\perp} = (W_1 \oplus W_1^{\perp}) = W_1^{\perp} \cap W_1^{\perp\perp} = \mathrm{rad}(W_1^{\perp}).$$

By the induction hypothesis applied to $W_1^{\perp}$, we obtain our theorem. $\square$

The following corollary follows immediately.

Proposition 29.25. *Let $\varphi \colon E \times E \to K$ be an alternating bilinear form on a space E of finite dimension n.*

(1) The rank of φ is even.

(2) If φ is nondegenerate, then $\dim(E) = n$ is even.

(3) Two alternating bilinear forms $\varphi_1 \colon E_1 \times E_1 \to K$ and $\varphi_2 \colon E_2 \times E_2 \to K$ are equivalent iff $\dim(E_1) = \dim(E_2)$ and φ_1 and φ_2 have the same rank.

The only part that requires a proof is part (3), which is left as an easy exercise.

If φ is nondegenerate, then $n = 2r$, and a basis of E as in Theorem 29.24 is called a *symplectic basis*. The space E is called a *hyperbolic space* (or *symplectic space*).

Observe that if we reorder the vectors in the basis

$$(u_1, v_1, \ldots, u_r, v_r, w_1, \ldots, w_{n-2r})$$

to obtain the basis

$$(u_1, \ldots, u_r, v_1, \ldots v_r, w_1, \ldots, w_{n-2r}),$$

then the matrix representing φ becomes

$$\begin{pmatrix} 0 & I_r & 0 \\ -I_r & 0 & 0 \\ 0 & 0 & 0_{n-2r} \end{pmatrix}.$$

This particularly simple matrix is often preferable, especially when dealing with the matrices (symplectic matrices) representing the isometries of φ (in which case $n = 2r$).

As a warm up for Proposition 29.29 of the next section, we prove an analog of Proposition 29.23 in the case of a symmetric bilinear form.

Proposition 29.26. *Let $\varphi\colon E \times E \to K$ be a nondegenerate symmetric bilinear form with K a field of characteristic different from 2. For any nonzero isotropic vector u, there is another nonzero isotropic vector v such that $\varphi(u, v) = 2$, and u and v are linearly independent. In the basis $(u, v/2)$, the restriction of φ to the plane spanned by u and $v/2$ is of the form*

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Proof. Since φ is nondegenerate, there is some nonzero vector z such that (rescaling z if necessary) $\varphi(u, z) = 1$. If

$$v = 2z - \varphi(z, z)u,$$

then since $\varphi(u, u) = 0$ and $\varphi(u, z) = 1$, note that

$$\varphi(u, v) = \varphi(u, 2z - \varphi(z, z)u) = 2\varphi(u, z) - \varphi(z, z)\varphi(u, u) = 2,$$

and

$$\begin{aligned} \varphi(v, v) &= \varphi(2z - \varphi(z, z)u, 2z - \varphi(z, z)u) \\ &= 4\varphi(z, z) - 4\varphi(z, z)\varphi(u, z) + \varphi(z, z)^2\varphi(u, u) \\ &= 4\varphi(z, z) - 4\varphi(z, z) = 0. \end{aligned}$$

If u and z were linearly dependent, as $u, z \neq 0$, we could write $z = \mu u$ for some $\mu \neq 0$, but then, we would have

$$\varphi(u, z) = \varphi(u, \mu u) = \mu\varphi(u, u) = 0,$$

contradicting the fact that $\varphi(u, z) \neq 0$. Then u and $v = 2z - \varphi(z, z)u$ are also linearly independent, since otherwise z could be expressed as a multiple of u. The rest is obvious. $\square$

Proposition 29.26 yields a plane spanned by two vectors u_1, v_1 such that $\varphi(u_1, u_1) = \varphi(v_1, v_1) = 0$ and $\varphi(u_1, v_1) = 1$. Such a plane is called an *Artinian plane*. Proposition 29.26 also shows that nonzero isotropic vectors come in pair.

Proposition 29.26 has the following corollary which has applications in number theory; see Serre [157], Chapter IV.

Proposition 29.27. *If Φ is any nondegenerate quadratic form (over a field of characteristic $\neq 2$) such that there is some nonzero vector $x \in E$ with $\Phi(x) = 0$, then for every $\alpha \in K$, there is some $y \in E$ such that $\Phi(y) = \alpha$.*

Proof. Since by hypothesis there is some nonzero vector $u \in E$ with $\Phi(u) = 0$, by Proposition 29.26 there is another isotropic vector v such that u and v are linearly independent and such that (after rescaling) $\varphi(u, v) = 1$. Then for any $\alpha \in K$, check that

$$\Phi\left(u + \frac{\alpha}{2}v\right) = \alpha,$$

as desired. $\square$

Remark: Some authors refer to the above plane as a *hyperbolic plane*. Berger (and others) point out that this terminology is undesirable because the notion of hyperbolic plane already exists in differential geometry and refers to a very different object.

We leave it as an exercice to figure out that the group of isometries of the Artinian plane, the set of all 2×2 matrices A such that

$$A^{\top} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

consists of all matrices of the form

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & \lambda \\ \lambda^{-1} & 0 \end{pmatrix}, \quad \lambda \in K - \{0\}.$$

In particular, if $K = \mathbb{R}$, then this group denoted $\mathbf{O}(1, 1)$ has four connected components.

We now turn to the Witt decomposition.

29.7 Witt Decomposition

From now on, $\varphi\colon E \times E \to K$ is an ϵ-Hermitian form. The following assumption will be needed:

Property (T). For every $u \in E$, there is some $\alpha \in K$ such that $\varphi(u, u) = \alpha + \epsilon\overline{\alpha}$.

Property (T) is always satisfied if φ is alternating, or if K is of characteristic $\neq 2$ and $\epsilon = \pm 1$, with $\alpha = \frac{1}{2}\varphi(u, u)$.

The following (bizarre) technical lemma will be needed.

Lemma 29.28. *Let φ be an ϵ-Hermitian form on E and assume that φ satisfies property (T). For any totally isotropic subspace $U \neq (0)$ of E, for every $x \in E$ not orthogonal to U, and for every $\alpha \in K$, there is some $y \in U$ so that*

$$\varphi(x + y, x + y) = \alpha + \epsilon\overline{\alpha}.$$

Proof. By property (T), we have $\varphi(x, x) = \beta + \epsilon\overline{\beta}$ for some $\beta \in K$. For any $y \in U$, since φ is ϵ-Hermitian, $\varphi(y, x) = \epsilon\overline{\varphi(x, y)}$, and since U is totally isotropic $\varphi(y, y) = 0$, so we have

$$\begin{aligned}\varphi(x+y, x+y) &= \varphi(x, x) + \varphi(x, y) + \varphi(y, x) + \varphi(y, y) \\ &= \beta + \epsilon\overline{\beta} + \varphi(x, y) + \epsilon\overline{\varphi(x, y)} \\ &= \beta + \varphi(x, y) + \epsilon\overline{(\beta + \varphi(x, y)}.\end{aligned}$$

Since x is not orthogonal to U, the function $y \mapsto \varphi(x, y) + \beta$ is not the constant function. Consequently, this function takes the value α for some $y \in U$, which proves the lemma. $\square$

Definition 29.18. Let φ be an ϵ-Hermitian form on E. A *weak Witt decomposition* of E is a triple (U, U', W), such that

(i) $E = U \oplus U' \oplus W$ (a direct sum).

(ii) U and U' are totally isotropic.

(iii) W is nondegenerate and orthogonal to $U \oplus U'$.

We say that a weak Witt decomposition (U, U', W) is *nontrivial* if $U \neq (0)$ and $U' \neq (0)$. Furthermore, if E is finite-dimensional, then $\dim(U) = \dim(U')$ and in a suitable basis, the matrix representing φ is of the form

$$\begin{pmatrix} 0 & A & 0 \\ \epsilon\overline{A} & 0 & 0 \\ 0 & 0 & B \end{pmatrix}$$

We say that φ is a *neutral form* if it is nondegenerate, E is finite-dimensional, and if $W = (0)$. In this case, the matrix B is missing.

A Witt decomposition for which W has no nonzero isotropic vectors (W is anisotropic) is called a *Witt decomposition*.

Observe that if Φ is nondegenerate, then we have the trivial weak Witt decomposition obtained by letting $U = U' = (0)$ and $W = E$. Thus a weak Witt decomposition is informative only if E is not anisotropic (there is some nonzero isotropic vector, *i.e.* some $u \neq 0$ such that $\Phi(u) = 0$), in which case the most informative nontrivial weak Witt decompositions are those for which W is anisotropic and U and U' are as big as possible.

Sometimes, we use the notation $U_1 \overset{\perp}{\oplus} U_2$ to indicate that in a direct sum $U_1 \oplus U_2$, the subspaces U_1 and U_2 are orthogonal. Then, in Definition 29.18, we can write that $E = (U \oplus U') \overset{\perp}{\oplus} W$.

The first step in showing the existence of a Witt decomposition is this.

Proposition 29.29. *Let φ be an ϵ-Hermitian form on E, assume that φ is nondegenerate and satisfies property (T), and let U be any totally isotropic subspace of E of finite dimension* $\dim(U) = r \geq 1$.

(1) *If U' is any totally isotropic subspace of dimension r and if $U' \cap U^{\perp} = (0)$, then $U \oplus U'$ is nondegenerate, and for any basis $(u_1, \ldots, u_r)$ of U, there is a basis $(u'_1, \ldots, u'_r)$ of U' such that $\varphi(u_i, u'_j) = \delta_{ij}$, for all $i, j = 1, \ldots, r$.*

(2) *If W is any totally isotropic subspace of dimension at most r and if $W \cap U^{\perp} = (0)$, then there exists a totally isotropic subspace U' with $\dim(U') = r$ such that $W \subseteq U'$ and $U' \cap U^{\perp} = (0)$.*

Proof. (1) Let φ' be the restriction of φ to $U \times U'$. Since $U' \cap U^{\perp} = (0)$, for any $v \in U'$, if $\varphi(u, v) = 0$ for all $u \in U$, then $v = 0$. Thus, φ' is nondegenerate (we only have to check on the left since φ is ϵ-Hermitian). Then, the assertion about bases follows from the version of Proposition 29.3 for sesquilinear forms. Since U is totally isotropic, $U \subseteq U^{\perp}$, and since $U' \cap U^{\perp} = (0)$, we must have $U' \cap U = (0)$, which show that we have a direct sum $U \oplus U'$.

It remains to prove that $U + U'$ is nondegenerate. Observe that

$$H = (U + U') \cap (U + U')^{\perp} = (U + U') \cap U^{\perp} \cap U'^{\perp}.$$

Since U is totally isotropic, $U \subseteq U^{\perp}$, and since $U' \cap U^{\perp} = (0)$, we have

$$(U + U') \cap U^{\perp} = (U \cap U^{\perp}) + (U' \cap U^{\perp}) = U + (0) = U,$$

thus $H = U \cap U'^{\perp}$. Since φ' is nondegenerate, $U \cap U'^{\perp} = (0)$, so $H = (0)$ and $U + U'$ is nondegenerate.

(2) We proceed by descending induction on $s = \dim(W)$. The base case $s = r$ is trivial. For the induction step, it suffices to prove that if $s < r$, then there is a totally isotropic subspace W' containing W such that $\dim(W') = s + 1$ and $W' \cap U^{\perp} = (0)$.

Since $s = \dim(W) < \dim(U)$, the restriction of φ to $U \times W$ is degenerate. Since $W \cap U^{\perp} = (0)$, we must have $U \cap W^{\perp} \neq (0)$. We claim that

$$W^{\perp} \not\subseteq W + U^{\perp}.$$

If we had

$$W^{\perp} \subseteq W + U^{\perp},$$

then because U and W are finite-dimensional and φ is nondegenerate, by Proposition 29.13, $U^{\perp\perp} = U$ and $W^{\perp\perp} = W$, so by taking orthogonals, $W^{\perp} \subseteq W + U^{\perp}$ would yield

$$(W + U^{\perp})^{\perp} \subseteq W^{\perp\perp},$$

that is,

$$W^{\perp} \cap U \subseteq W,$$

thus $W^\perp \cap U \subseteq W \cap U$, and since U is totally isotropic, $U \subseteq U^\perp$, which yields

$$W^\perp \cap U \subseteq W \cap U \subseteq W \cap U^\perp = (0),$$

contradicting the fact that $U \cap W^\perp \neq (0)$.

Therefore, there is some $u \in W^\perp$ such that $u \notin W + U^\perp$. Since $U \subseteq U^\perp$, we can add to u any vector $z \in W^\perp \cap U \subseteq U^\perp$ so that $u + z \in W^\perp$ and $u + z \notin W + U^\perp$ (if $u + z \in W + U^\perp$, since $z \in U^\perp$, then $u \in W + U^\perp$, a contradiction). Since $W^\perp \cap U \neq (0)$ is totally isotropic and $u \notin W + U^\perp = (W^\perp \cap U)^\perp$, we can invoke Lemma 29.28 to find a $z \in W^\perp \cap U$ such that $\varphi(u + z, u + z) = 0$. See Figure 29.1. If we write $x = u + z$, then $x \notin W + U^\perp$, so $W' = W + Kx$ is a totally isotropic subspace of dimension $s + 1$. Furthermore, we claim that $W' \cap U^\perp = 0$.

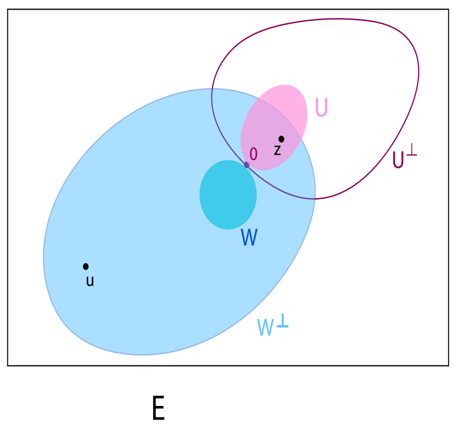


Figure 29.1: A schematic illustration of W and $x = u + z$

Otherwise, we would have $y = w + \lambda x \in U^\perp$, for some $w \in W$ and some $\lambda \in K$, and then we would have $\lambda x = -w + y \in W + U^\perp$. If $\lambda \neq 0$, then $x \in W + U^\perp$, a contradiction. Therefore, $\lambda = 0$, $y = w$, and since $y \in U^\perp$ and $w \in W$, we have $y \in W \cap U^\perp = (0)$, which means that $y = 0$. Therefore, W' is the required subspace and this completes the proof. $\square$

Here are some consequences of Proposition 29.29. If we set $W = (0)$ in Proposition 29.29(2), then we get the following theorem showing that if E is not anisotropic (there is some nonzero isotropic vector) then weak nontrivial Witt decompositions exist.

Theorem 29.30. *Let φ be an ϵ-Hermitian form on E which is nondegenerate and satisfies property (T). For any totally isotropic subspace U of E of finite dimension $r \geq 1$, there exists a totally isotropic subspace U' of dimension r such that $U \cap U' = (0)$ and $U \oplus U'$ is nondegenerate. As a consequence, if E is not anisotropic, then $(U, U', (U \oplus U')^\perp)$ is a weak nontrivial Witt decomposition for E. Furthermore, by Proposition 29.29(1), the block A in the matrix of φ is the identity matrix.*

Proposition 29.31. *Any two ϵ-Hermitian neutral forms satisfying property (T) defined on spaces of the same dimension are equivalent.*

The following proposition shows that every subspace U of E can be embedded into a nondegenerate subspace. It is needed to prove a version of the Witt extension theorem (Theorem 29.48).

Proposition 29.32. *Let φ be an ϵ-Hermitian form on E which is nondegenerate and satisfies property (T). For any subspace U of E of finite dimension, if we write*

$$U = V \overset{\perp}{\oplus} W,$$

for some orthogonal complement W of $V = \mathrm{rad}(U)$, and if we let $r = \dim(\mathrm{rad}(U))$, then there exists a totally isotropic subspace V' of dimension r such that $V \cap V' = (0)$, and $(V \oplus V') \overset{\perp}{\oplus} W = V' \oplus U$ is nondegenerate. Furthermore, any isometry f from U into another space (E', φ') where φ' is an ϵ-Hermitian form satisfying the same assumptions as φ can be extended to an isometry on $(V \oplus V') \overset{\perp}{\oplus} W$.

Proof. Since W is nondegenerate, $W^{\perp}$ is also nondegenerate, and $V \subseteq W^{\perp}$. Therefore, we can apply Theorem 29.30 to the restriction of φ to $W^{\perp}$ and to V to obtain the required V'. We know that $V \oplus V'$ is nondegenerate and orthogonal to W, which is also nondegenerate, so $(V \oplus V') \overset{\perp}{\oplus} W = V' \oplus U$ is nondegenerate.

We leave the second statement about extending f as an exercise (use the fact that $f(U) = f(V) \overset{\perp}{\oplus} f(W)$, where $V_1 = f(V)$ is totally isotropic of dimension r, to find another totally isotropic susbpace V_1' of dimension r such that $V_1 \cap V_1' = (0)$ and $V_1 \oplus V_1'$ is orthogonal to $f(W)$). ∎

The subspace $(V \oplus V') \overset{\perp}{\oplus} W = V' \oplus U$ is often called a *nondegenerate completion* of U. The subspace $V \oplus V'$ is called an *Artinian space*. Proposition 29.29 show that $V \oplus V'$ has a basis $(u_1, v_1, \ldots, u_r, v_r)$ consisting of vectors $u_i \in V$ and $v_j \in V'$ such that $\varphi(u_i, u_j) = \delta_{ij}$. The subspace spanned by (u_i, v_i) is an Artinian plane, so $V \oplus V'$ is the orthogonal direct sum of r Artinian planes. Such a space is often denoted by Ar_{2r}.

In order to obtain the stronger version of the Witt decomposition when φ has some nonzero isotropic vector and W is anisotropic we now sharpen Proposition 29.29

Theorem 29.33. *Let φ be an ϵ-Hermitian form on E which is nondegenerate and satisfies property (T). Let U_1 and U_2 be two totally isotropic maximal subspaces of E, with U_1 or U_2 of finite dimension ≥ 1. Write $U = U_1 \cap U_2$, let S_1 be a supplement of U in U_1 and S_2 be a supplement of U in U_2 (so that $U_1 = U \oplus S_1$, $U_2 = U \oplus S_2$), and let $S = S_1 + S_2$. Then, there exist two subspaces W and D of E such that:*

(a) The subspaces S, $U + W$, and D are nondegenerate and pairwise orthogonal.

(b) We have a direct sum $E = S \overset{\perp}{\oplus} (U \oplus W) \overset{\perp}{\oplus} D$.

(c) The subspace D *contains no nonzero isotropic vector* (D *is anisotropic*).

(d) The subspace W *is totally isotropic.*

Furthermore, U_1 *and* U_2 *are both finite dimensional, and we have* $\dim(U_1) = \dim(U_2)$, $\dim(W) = \dim(U)$, $\dim(S_1) = \dim(S_2)$, *and* $\mathrm{codim}(D) = 2\dim(F_1)$.

Proof. First observe that if X is a totally isotropic maximal subspace of E, then any isotropic vector $x \in E$ orthogonal to X must belong to X, since otherwise, $X + Kx$ would be a totally isotropic subspace strictly containing X, contradicting the maximality of X. As a consequence, if x_i is any isotropic vector such that $x_i \in U_i^{\perp}$ (for $i = 1, 2$), then $x_i \in U_i$.

We claim that

$$S_1 \cap S_2^{\perp} = (0) \quad \text{and} \quad S_2 \cap S_1^{\perp} = (0).$$

Assume that $y \in S_1$ is orthogonal to S_2. Since $U_1 = U \oplus S_1$ and U_1 is totally isotropic, y is orthogonal to U_1, and thus orthogonal to U, so that y is orthogonal to $U_2 = U \oplus S_2$. Since $S_1 \subseteq U_1$ and U_1 is totally isotropic, y is an isotropic vector orthogonal to U_2, which by a previous remark implies that $y \in U_2$. Then, since $S_1 \subseteq U_1$ and $U \oplus S_1$ is a direct sum, we have

$$y \in S_1 \cap U_2 = S_1 \cap U_1 \cap U_2 = S_1 \cap U = (0).$$

Therefore $S_1 \cap S_2^{\perp} = (0)$. A similar proof show that $S_2 \cap S_1^{\perp} = (0)$. If U_1 is finite-dimensional (the case where U_2 is finite-dimensional is similar), then S_1 is finite-dimensional, so by Proposition 29.13, $S_1^{\perp}$ has finite codimension. Since $S_2 \cap S_1^{\perp} = (0)$, and since any supplement of $S_1^{\perp}$ has finite dimension, we must have

$$\dim(S_2) \leq \mathrm{codim}(S_1^{\perp}) = \dim(S_1).$$

By a similar argument, $\dim(S_1) \leq \dim(S_2)$, so we have

$$\dim(S_1) = \dim(S_2).$$

By Proposition 29.29(1), we conclude that $S = S_1 + S_2$ is nondegenerate.

By Proposition 29.21, the subspace $N = S^{\perp} = (S_1 + S_2)^{\perp}$ is nondegenerate. Since $U_1 = U \oplus S_1$, $U_2 = U \oplus S_2$, and U_1, U_2 are totally isotropic, U is orthogonal to S_1 and to S_2, so $U \subseteq N$. Since U is totally isotropic, by Proposition 29.30 applied to N, there is a totally isotropic subspace W of N such that $\dim(W) = \dim(U)$, $U \cap W = (0)$, and $U + W$ is nondegenerate. Consequently, (d) is satisfied by W.

To satisfy (a) and (b), we pick D to be the orthogonal of $U \oplus W$ in N. Then, $N = (U \oplus W) \overset{\perp}{\oplus} D$ and $E = S \overset{\perp}{\oplus} N$, so $E = S \overset{\perp}{\oplus} (U \oplus W) \overset{\perp}{\oplus} D$.

As to (c), since D is orthogonal $U \oplus W$, D is orthogonal to U, and since $D \subseteq N$ and N is orthogonal to $S_1 + S_2$, D is orthogonal to S_1, so D is orthogonal to $U_1 = U \oplus S_1$. If $y \in D$

is any isotropic vector, since $y \in U_1^{\perp}$, by a previous remark, $y \in U_1$, so $y \in D \cap U_1$. But, $D \subseteq N$ with $N \cap (S_1 + S_2) = (0)$, and $D \cap (U + W) = (0)$, so $D \cap (U + S_1) = D \cap U_1 = (0)$, which yields $y = 0$. The statements about dimensions are easily obtained. $\square$

Finally, Theorem 29.33 yields the strong form of the Witt decomposition in which W is anistropic. Given any matrix $A \in \mathrm{M}_n(K)$, we say that A is *definite* if $x^{\top} A x \neq 0$ for all $x \in K^n$.

Theorem 29.34. *Let φ be an ϵ-Hermitian form on E which is nondegenerate and satisfies property (T).*

(1) Any two totally isotropic maximal spaces of finite dimension have the same dimension.

(2) For any totally isotropic maximal subspace U of finite dimension $r \geq 1$, there is another totally isotropic maximal subspace U' of dimension r such that $U \cap U' = (0)$, and $U \oplus U'$ is nondegenerate. Furthermore, if $D = (U \oplus U')^{\perp}$, then (U, U', D) is a Witt decomposition of E; that is, there are no nonzero isotropic vectors in D (D is anisotropic).

(3) If E has finite dimension $n \geq 1$ and there is some nonzero isotropic vector for φ (E is not anisotropic), then E has a nontrivial Witt decomposition (U, U', D) as in (2). There is a basis of E such that

(a) if φ is alternating ($\epsilon = -1$ and $\lambda = \overline{\lambda}$ for all $\lambda \in K$), then $n = 2m$ and φ is represented by a matrix of the form

$$\begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}$$

(b) if φ is symmetric ($\epsilon = +1$ and $\lambda = \overline{\lambda}$ for all $\lambda \in K$), then φ is represented by a matrix of the form

$$\begin{pmatrix} 0 & I_r & 0 \\ I_r & 0 & 0 \\ 0 & 0 & P \end{pmatrix},$$

where either $n = 2r$ and P does not occur, or $n > 2r$ and P is a definite symmetric matrix.

(c) if φ is ϵ-Hermitian (the involutive automorphism $\lambda \mapsto \overline{\lambda}$ is not the identity), then φ is represented by a matrix of the form

$$\begin{pmatrix} 0 & I_r & 0 \\ \epsilon I_r & 0 & 0 \\ 0 & 0 & P \end{pmatrix},$$

where either $n = 2r$ and P does not occur, or $n > 2r$ and P is a definite matrix such that $P^* = \epsilon P$.

Proof. Part (1) follows from Theorem 29.33. By Proposition 29.30, we obtain a totally isotropic subspace U' of dimension r such that $U \cap U' = (0)$. By applying Theorem 29.33 to $U_1 = U$ and $U_2 = U'$, we get $U = W = (0)$, which proves (2). Part (3) is an immediate consequence of (2). $\square$

As a consequence of Theorem 29.34, we make the following definition.

Definition 29.19. Let E be a vector space of finite dimension n, and let φ be an ϵ-Hermitian form on E which is nondegenerate and satisfies property (T). The *index* (or *Witt index*) ν of φ, is the common dimension of all totally isotropic maximal subspaces of E. We have $2\nu \leq n$.

Neutral forms only exist if n is even, in which case, $\nu = n/2$. Forms of index $\nu = 0$ have no nonzero isotropic vectors. When $K = \mathbb{R}$, this is satisfied by positive definite or negative definite symmetric forms. When $K = \mathbb{C}$, this is satisfied by positive definite or negative definite Hermitian forms. The vector space of a neutral Hermitian form ($\epsilon = +1$) is an Artinian space, and the vector space of a neutral alternating form is a hyperbolic space.

If the field K is algebraically closed, we can describe all nondegenerate quadratic forms.

Proposition 29.35. *If K is algebraically closed and E has dimension n, then for every nondegenerate quadratic form Φ, there is a basis $(e_1, \ldots, e_n)$ such that Φ is given by*

$$\Phi\left(\sum_{i-1}^{n} x_i e_i\right) = \begin{cases} \sum_{i=1}^{m} x_i x_{m+i} & \text{if } n = 2m \\ \sum_{i=1}^{m} x_i x_{m+i} + x_{2m+1}^2 & \text{if } n = 2m+1. \end{cases}$$

Proof. We work with the polar form φ of Φ. Let U_1 and U_2 be some totally isotropic subspaces such that $U_1 \cap U_2 = (0)$ given by Theorem 29.34, and let q be their common dimension. Then, $W = U = (0)$. Since we can pick bases $(e_1, \ldots e_q)$ in U_1 and $(e_{q+1}, \ldots, e_{2q})$ in U_2 such that $\varphi(e_i, e_{i+q}) = 0$, for $i, j = 1, \ldots, q$, it suffices to proves that $\dim(D) \leq 1$. If $x, y \in D$ with $x \neq 0$, from the identity

$$\Phi(y - \lambda x) = \Phi(y) - \lambda\varphi(x, y) + \lambda^2\Phi(x)$$

and the fact that $\Phi(x) \neq 0$ since $x \in D$ and $x \neq 0$, we see that the equation $\Phi(y - \lambda y) = 0$ has at least one solution. Since $\Phi(z) \neq 0$ for every nonzero $z \in D$, we get $y = \lambda x$, and thus $\dim(D) \leq 1$, as claimed. $\square$

Proposition 29.35 shows that for every nondegenerate quadratic form Φ over an algebraically closed field, if $\dim(E) = 2m$ or $\dim(E) = 2m + 1$ with $m \geq 1$, then Φ has some nonzero isotropic vector.

29.8 Symplectic Groups

In this section, we are dealing with a nondegenerate alternating form φ on a vector space E of dimension n. As we saw earlier, n must be even, say $n = 2m$. By Theorem 29.24, there is a direct sum decomposition of E into pairwise orthogonal subspaces

$$E = W_1 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} W_m,$$

where each W_i is a hyperbolic plane. Each W_i has a basis (u_i, v_i), with $\varphi(u_i, u_i) = \varphi(v_i, v_i) = 0$ and $\varphi(u_i, v_i) = 1$, for $i = 1, \ldots, m$. In the basis

$$(u_1, \ldots, u_m, v_1, \ldots, v_m),$$

φ is represented by the matrix

$$J_{m,m} = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}.$$

The symplectic group $\mathbf{Sp}(2m, K)$ is the group of isometries of φ. The maps in $\mathbf{Sp}(2m, K)$ are called *symplectic* maps. With respect to the above basis, $\mathbf{Sp}(2m, K)$ is the group of $2m \times 2m$ matrices A such that

$$A^\top J_{m,m} A = J_{m,m}.$$

Matrices satisfying the above identity are called *symplectic* matrices. In this section, we show that $\mathbf{Sp}(2m, K)$ is a subgroup of $\mathbf{SL}(2m, K)$ (that is, $\det(A) = +1$ for all $A \in \mathbf{Sp}(2m, K)$), and we show that $\mathbf{Sp}(2m, K)$ is generated by special linear maps called *symplectic transvections*.

First, we leave it as an easy exercise to show that $\mathbf{Sp}(2, K) = \mathbf{SL}(2, K)$. The reader should also prove that $\mathbf{Sp}(2m, K)$ has a subgroup isomorphic to $\mathbf{GL}(m, K)$.

Next we characterize the symplectic maps f that leave fixed every vector in some given hyperplane H, that is,

$$f(v) = v \quad \text{for all } v \in H.$$

Since φ is nondegenerate, by Proposition 29.22, the orthogonal $H^\perp$ of H is a line (that is, $\dim(H^\perp) = 1$). For every $u \in E$ and every $v \in H$, since f is an isometry and $f(v) = v$ for all $v \in H$, we have

$$\begin{aligned} \varphi(f(u) - u, v) &= \varphi(f(u), v) - \varphi(u, v) \\ &= \varphi(f(u), v) - \varphi(f(u), f(v)) \\ &= \varphi(f(u), v - f(v))) \\ &= \varphi(f(u), 0) = 0, \end{aligned}$$

which shows that $f(u) - u \in H^\perp$ for all $u \in E$. Therefore, $f - \mathrm{id}$ is a linear map from E into the line $H^\perp$ whose kernel contains H, which means that there is some nonzero vector $w \in H^\perp$ and some linear form ψ such that

$$f(u) = u + \psi(u)w, \quad u \in E.$$

Since f is an isometry, we must have $\varphi(f(u), f(v)) = \varphi(u, v)$ for all $u, v \in E$, which means that

$$\begin{aligned}
\varphi(u, v) &= \varphi(f(u), f(v)) \\
&= \varphi(u + \psi(u)w, v + \psi(v)w) \\
&= \varphi(u, v) + \psi(u)\varphi(w, v) + \psi(v)\varphi(u, w) + \psi(u)\psi(v)\varphi(w, w) \\
&= \varphi(u, v) + \psi(u)\varphi(w, v) - \psi(v)\varphi(w, u),
\end{aligned}$$

which yields

$$\psi(u)\varphi(w, v) = \psi(v)\varphi(w, u) \quad \text{for all } u, v \in E.$$

Since φ is nondegenerate, we can pick some v_0 such that $\varphi(w, v_0) \neq 0$, and we get $\psi(u)\varphi(w, v_0) = \psi(v_0)\varphi(w, u)$ for all $u \in E$; that is,

$$\psi(u) = \lambda\varphi(w, u) \quad \text{for all } u \in E,$$

for some $\lambda \in K$. Therefore, f is of the form

$$f(u) = u + \lambda\varphi(w, u)w, \quad \text{for all } u \in E.$$

It is also clear that every f of the above form is a symplectic map. If $\lambda = 0$, then $f = \mathrm{id}$. Otherwise, if $\lambda \neq 0$, then $f(u) = u$ iff $\varphi(w, u) = 0$ iff $u \in (Kw)^{\perp} = H$, where H is a hyperplane. Thus, f fixes every vector in the hyperplane H. Note that since φ is alternating, $\varphi(w, w) = 0$, which means that $w \in H$.

In summary, we have characterized all the symplectic maps that leave every vector in some hyperplane fixed, and we make the following definition.

Definition 29.20. Given a nondegenerate alternating form φ on a space E, a *symplectic transvection (of direction w)* is a linear map f of the form

$$f(u) = u + \lambda\varphi(w, u)w, \quad \text{for all } u \in E,$$

for some nonzero $w \in E$ and some $\lambda \in K$. If $\lambda \neq 0$, the subspace of vectors left fixed by f is the hyperplane $H = (Kw)^{\perp}$. The map f is also denoted $\tau_{w,\lambda}$.

Observe that

$$\tau_{w,\lambda} \circ \tau_{w,\mu} = \tau_{w,\lambda+\mu}$$

and $\tau_{w,\lambda} = \mathrm{id}$ iff $\lambda = 0$. The above shows that $\det(\tau_{w,\lambda}) = 1$, since when $\lambda \neq 0$, we have $\tau_{w,\lambda} = (\tau_{w,\lambda/2})^2$.

Our next goal is to show that if u and v are any two nonzero vectors in E, then there is a simple symplectic map f such that $f(u) = v$.

Proposition 29.36. *Given any two nonzero vectors $u, v \in E$, there is a symplectic map f such that $f(u) = v$, and f is either a symplectic transvection, or the composition of two symplectic transvections.*

Proof. There are two cases.

Case 1. $\varphi(u, v) \neq 0$.

In this case, $u \neq v$, since $\varphi(u, u) = 0$. Let us look for a symplectic transvection of the form $\tau_{v-u,\lambda}$. We want

$$v = u + \lambda\varphi(v - u, u)(v - u) = u + \lambda\varphi(v, u)(v - u),$$

which yields

$$(\lambda\varphi(v, u) - 1)(v - u) = 0.$$

Since $\varphi(u, v) \neq 0$ and $\varphi(v, u) = -\varphi(u, v)$, we can pick $\lambda = \varphi(v, u)^{-1}$ and $\tau_{v-u,\lambda}$ maps u to v.

Case 2. $\varphi(u, v) = 0$.

If $u = v$, use $\tau_{u,0} = \mathrm{id}$. Now, assume $u \neq v$. We claim that it is possible to pick some $w \in E$ such that $\varphi(u, w) \neq 0$ and $\varphi(v, w) \neq 0$. Indeed, if $(Ku)^\perp = (Kv)^\perp$, then pick any nonzero vector w not in the hyperplane $(Ku)^\perp$. Othwerwise, $(Ku)^\perp$ and $(Kv)^\perp$ are two distinct hyperplanes, so neither is contained in the other (they have the same dimension), so pick any nonzero vector w_1 such that $w_1 \in (Ku)^\perp$ and $w_1 \notin (Kv)^\perp$, and pick any nonzero vector w_2 such that $w_2 \in (Kv)^\perp$ and $w_2 \notin (Ku)^\perp$. If we let $w = w_1 + w_2$, then $\varphi(u, w) = \varphi(u, w_2) \neq 0$, and $\varphi(v, w) = \varphi(v, w_1) \neq 0$. From case 1, we have some symplectic transvection τ_{w-u,λ_1} such that $\tau_{w-u,\lambda_1}(u) = w$, and some symplectic transvection τ_{v-w,λ_2} such that $\tau_{v-w,\lambda_2}(w) = v$, so the composition $\tau_{v-w,\lambda_2} \circ \tau_{w-u,\lambda_1}$ maps u to v. $\square$

Next, we would like to extend Proposition 29.36 to two hyperbolic planes W_1 and W_2.

Proposition 29.37. *Given any two hyperbolic planes W_1 and W_2 given by bases (u_1, v_1) and (u_2, v_2) (with $\varphi(u_i, u_i) = \varphi(v_i, v_i) = 0$ and $\varphi(u_i, v_i) = 1$, for $i = 1, 2$), there is a symplectic map f such that $f(u_1) = u_2$, $f(v_1) = v_2$, and f is the composition of at most four symplectic transvections.*

Proof. From Proposition 29.36, we can map u_1 to u_2, using a map f which is the composition of at most two symplectic transvections. Say $v_3 = f(v_1)$. We claim that there is a map g such that $g(u_2) = u_2$ and $g(v_3) = v_2$, and g is the composition of at most two symplectic transvections. If so, $g \circ f$ maps the pair (u_1, v_1) to the pair (u_2, v_2), and $g \circ f$ consists of at most four symplectic transvections. Thus, we need to prove the following claim:

Claim. If (u, v) and (u, v') are hyperbolic bases determining two hyperbolic planes, then there is a symplectic map g such that $g(u) = u$, $g(v) = v'$, and g is the composition of at most two symplectic transvections. There are two case.

Case 1. $\varphi(v, v') \neq 0$.

In this case, there is a symplectic transvection $\tau_{v'-v,\lambda}$ such that $\tau_{v'-v,\lambda}(v) = v'$. We also have

$$\varphi(u, v' - v) = \varphi(u, v') - \varphi(u, v) = 1 - 1 = 0.$$

Therefore, $\tau_{v'-v,\lambda}(u) = u$, and $g = \tau_{v'-v,\lambda}$ does the job.

Case 2. $\varphi(v, v') = 0$.

First, check that $(u, u+v)$ is a also hyperbolic basis. Furthermore,

$$\varphi(v, u+v) = \varphi(v, u) + \varphi(v, v) = \varphi(v, u) = -1 \neq 0.$$

Thus, there is a symplectic transvection τ_{v,λ_1} such that $\tau_{u,\lambda_1}(v) = u + v$ and $\tau_{u,\lambda_1}(u) = u$. We also have

$$\varphi(u+v, v') = \varphi(u, v') + \varphi(v, v') = \varphi(u, v') = 1 \neq 0,$$

so there is a symplectic transvection $\tau_{v'-u-v,\lambda_2}$ such that $\tau_{v'-u-v,\lambda_2}(u+v) = v'$. Since

$$\varphi(u, v'-u-v) = \varphi(u, v') - \varphi(u, u) - \varphi(u, v) = 1 - 0 - 1 = 0,$$

we have $\tau_{v'-u-v,\lambda_2}(u) = u$. Then, the composition $g = \tau_{v'-u-v,\lambda_2} \circ \tau_{u,\lambda_1}$ is such that $g(u) = u$ and $g(v) = v'$. $\square$

We will use Proposition 29.37 in an inductive argument to prove that the symplectic transvections generate the symplectic group. First, make the following observation: If U is a nondegenerate subspace of E, so that

$$E = U \overset{\perp}{\oplus} U^{\perp},$$

and if τ is a transvection of $H^{\perp}$, then we can form the linear map $\mathrm{id}_U \overset{\perp}{\oplus} \tau$ whose restriction to U is the identity and whose restriction to $U^{\perp}$ is τ, and $\mathrm{id}_U \overset{\perp}{\oplus} \tau$ is a transvection of E.

Theorem 29.38. *The symplectic group* $\mathbf{Sp}(2m, K)$ *is generated by the symplectic transvections. For every transvection* $f \in \mathbf{Sp}(2m, K)$*, we have* $\det(f) = 1$.

Proof. Let G be the subgroup of $\mathbf{Sp}(2m, K)$ generated by the transvections. We need to prove that $G = \mathbf{Sp}(2m, K)$. Let $(u_1, v_1, \ldots, u_m, v_m)$ be a symplectic basis of E, and let $f \in \mathbf{Sp}(2m, K)$ be any symplectic map. Then, f maps $(u_1, v_1, \ldots, u_m, v_m)$ to another symplectic basis $(u'_1, v'_1, \ldots, u'_m, v'_m)$. If we prove that there is some $g \in G$ such that $g(u_i) = u'_i$ and $g(v_i) = v'_i$ for $i = 1, \ldots, m$, then $f = g$ and $G = \mathbf{Sp}(2m, K)$.

We use induction on i to prove that there is some $g_i \in G$ so that g_i maps $(u_1, v_1, \ldots, u_i, v_i)$ to $(u'_1, v'_1, \ldots, u'_i, v'_i)$.

The base case $i = 1$ follows from Proposition 29.37.

For the induction step, assume that we have some $g_i \in G$ mapping $(u_1, v_1, \ldots, u_i, v_i)$ to $(u'_1, v'_1, \ldots, u'_i, v'_i)$, and let $(u''_{i+1}, v''_{i+1}, \ldots, u''_m, v''_m)$ be the image of $(u_{i+1}, v_{i+1}, \ldots, u_m, v_m)$ by g_i. If U is the subspace spanned by $(u'_1, v'_1, \ldots, u'_m, v'_m)$, then each hyperbolic plane W'_{i+k} given by (u'_{i+k}, v'_{i+k}) and each hyperbolic plane W''_{i+k} given by (u''_{i+k}, v''_{i+k}) belongs to

$U^\perp$. Using the remark before the theorem and Proposition 29.37, we can find a transvection τ mapping W''_{i+1} onto W'_{i+1} and leaving every vector in U fixed. Then, $\tau \circ g_i$ maps $(u_1, v_1, \ldots, u_{i+1}, v_{i+1})$ to $(u'_1, v'_1, \ldots, u'_{i+1}, v'_{i+1})$, establishing the induction step.

For the second statement, since we already proved that every transvection has a determinant equal to $+1$, this also holds for any composition of transvections in G, and since $G = \mathbf{Sp}(2m, K)$, we are done. $\square$

It can also be shown that the center of $\mathbf{Sp}(2m, K)$ is reduced to the subgroup $\{\mathrm{id}, -\mathrm{id}\}$. The *projective symplectic group* $\mathbf{PSp}(2m, K)$ is the quotient group $\mathbf{PSp}(2m, K)/\{\mathrm{id}, -\mathrm{id}\}$. All symplectic projective groups are simple, except $\mathbf{PSp}(2, \mathbb{F}_2)$, $\mathbf{PSp}(2, \mathbb{F}_3)$, and $\mathbf{PSp}(4, \mathbb{F}_2)$, see Grove [83].

The orders of the symplectic groups over finite fields can be determined. For details, see Artin [6], Jacobson [98] and Grove [83].

An interesting property of symplectic spaces is that the determinant of a skew-symmetric matrix B is the square of some polynomial $\mathrm{Pf}(B)$ called the *Pfaffian*; see Jacobson [98] and Artin [6]. We leave considerations of the Pfaffian to the exercises.

We now take a look at the orthogonal groups.

29.9 Orthogonal Groups and the Cartan–Dieudonné Theorem

In this section we are dealing with a nondegenerate symmetric bilinear from φ over a finite-dimensional vector space E of dimension n over a field of characateristic not equal to 2. Recall that the orthogonal group $\mathbf{O}(\varphi)$ is the group of isometries of φ; that is, the group of linear maps $f\colon E \to E$ such that

$$\varphi(f(u), f(v)) = \varphi(u, v) \quad \text{for all } u, v \in E.$$

The elements of $\mathbf{O}(\varphi)$ are also called *orthogonal transformations*. If M is the matrix of φ in any basis, then a matrix A represents an orthogonal transformation iff

$$A^\top M A = M.$$

Since φ is nondegenerate, M is invertible, so we see that $\det(A) = \pm 1$. The subgroup

$$\mathbf{SO}(\varphi) = \{f \in \mathbf{O}(\varphi) \mid \det(f) = 1\}$$

is called the *special orthogonal group* (of φ), and its members are called *rotations* (or *proper orthogonal transformations*). Isometries $f \in \mathbf{O}(\varphi)$ such that $\det(f) = -1$ are called *improper orthogonal transformations*, or sometimes *reversions*.

If H is any nondegenerate hyperplane in E, then $D = H^{\perp}$ is a nondegenerate line and we have

$$E = H \overset{\perp}{\oplus} H^{\perp}.$$

For any nonzero vector $u \in D = H^{\perp}$ Consider the map τ_u given by

$$\tau_u(v) = v - 2\frac{\varphi(v, u)}{\varphi(u, u)}u \quad \text{for all } v \in E.$$

If we replace u by λu with $\lambda \neq 0$, we have

$$\tau_{\lambda u}(v) = v - 2\frac{\varphi(v, \lambda u)}{\varphi(\lambda u, \lambda u)}\lambda u = v - 2\frac{\lambda\varphi(v, u)}{\lambda^2\varphi(u, u)}\lambda u = v - 2\frac{\varphi(v, u)}{\varphi(u, u)}u,$$

which shows that τ_u depends only on the line D, and thus only the hyperplane H. Therefore, denote by τ_H the linear map τ_u determined as above by any nonzero vector $u \in H^{\perp}$. Note that if $v \in H$, then

$$\tau_H(v) = v,$$

and if $v \in D$, then

$$\tau_H(v) = -v.$$

A simple computation shows that

$$\varphi(\tau_H(u), \tau_H(v)) = \varphi(u, v) \quad \text{for all } u, v \in E,$$

so $\tau_H \in \mathbf{O}(\varphi)$, and by picking a basis consisting of u and vectors in H, that $\det(\tau_H) = -1$. It is also clear that $\tau_H^2 = \mathrm{id}$.

Definition 29.21. If H is any nondegenerate hyperplane in E, for any nonzero vector $u \in H^{\perp}$, the linear map τ_H given by

$$\tau_H(v) = v - 2\frac{\varphi(v, u)}{\varphi(u, u)}u \quad \text{for all } v \in E$$

is an involutive isometry of E called the *reflection through (or about) the hyperplane H.*

Remarks:

1. It can be shown that if $f \in \mathbf{O}(\varphi)$ leaves every vector in some hyperplane H fixed, then either $f = \mathrm{id}$ or $f = \tau_H$; see Taylor [174] (Chapter 11). Thus, there is no analog to symplectic transvections in the orthogonal group.

2. If $K = \mathbb{R}$ and φ is the usual Euclidean inner product, the matrices corresponding to hyperplane reflections are called *Householder matrices*.

Our goal is to prove that $\mathbf{O}(\varphi)$ is generated by the hyperplane reflections. The following proposition is needed.

Proposition 29.39. *Let φ be a nondegenerate symmetric bilinear form on a vector space E. For any two nonzero vectors $u, v \in E$, if $\varphi(u, u) = \varphi(v, v)$ and $v - u$ is nonisotropic, then the hyperplane reflection $\tau_H = \tau_{v-u}$ maps u to v, with $H = (K(v - u))^{\perp}$.*

Proof. Since $v - u$ is not isotropic, $\varphi(v - u, v - u) \neq 0$, and we have

$$\begin{aligned}
\tau_{v-u}(u) &= u - 2\frac{\varphi(u, v-u)}{\varphi(v-u, v-u)}(v-u) \\
&= u - 2\frac{\varphi(u, v) - \varphi(u, u)}{\varphi(v, v) - 2\varphi(u, v) + \varphi(u, u)}(v-u) \\
&= u - \frac{2(\varphi(u, v) - \varphi(u, u))}{2(\varphi(u, u) - 2\varphi(u, v))}(v-u) \\
&= v,
\end{aligned}$$

which proves the proposition. ☐

We can now obtain a cheap version of the Cartan–Dieudonné theorem.

Theorem 29.40. *(Cartan–Dieudonné, weak form) Let φ be a nondegenerate symmetric bilinear form on a K-vector space E of dimension n ($\mathrm{char}(K) \neq 2$). Then, every isometry $f \in \mathbf{O}(\varphi)$ with $f \neq \mathrm{id}$ is the composition of at most $2n - 1$ hyperplane reflections.*

Proof. We proceed by induction on n. For $n = 0$, this is trivial (since $\mathbf{O}(\varphi) = \{\mathrm{id}\}$).

Next, assume that $n \geq 1$. Since φ is nondegenerate, we know that there is some nonisotropic vector $u \in E$. There are three cases.

Case 1. $f(u) = u$.

Since φ is nondegenrate and u is nonisotropic, the hyperplane $H = (Ku)^{\perp}$ is nondegenerate, $E = H \overset{\perp}{\oplus} Ku$, and since $f(u) = u$, we must have $f(H) = H$. The restriction f' of of f to H is an isometry of H. By the induction hypothesis, we can write

$$f' = \tau_k' \circ \cdots \circ \tau_1',$$

where τ_i is some hyperplane reflection about a hyperplane L_i in H, with $k \leq 2n - 3$. We can extend each τ_i' to a reflection τ_i about the hyperplane $L_i \overset{\perp}{\oplus} Ku$ so that $\tau_i(u) = u$, and clearly,

$$f = \tau_k \circ \cdots \circ \tau_1.$$

Case 2. $f(u) = -u$.

If τ is the hyperplane reflection about the hyperplane $H = (Ku)^{\perp}$, then $g = \tau \circ f$ is an isometry of E such that $g(u) = u$, and we are back to Case (1). Since $\tau^2 = 1$ We obtain

$$f = \tau \circ \tau_k \circ \cdots \circ \tau_1$$

where τ and the τ_i are hyperplane reflections, with $k \geq 2n-3$, and we get a total of $2n-2$ hyperplane reflections.

Case 3. $f(u) \neq u$ and $f(u) \neq -u$.

Note that $f(u) - u$ and $f(u) + u$ are orthogonal, since

$$\begin{aligned}\varphi(f(u) - u, f(u) + u) &= \varphi(f(u), f(u)) + \varphi(f(u), u) - \varphi(u, f(u)) - \varphi(u, u) \\ &= \varphi(u, u) - \varphi(u, u) = 0.\end{aligned}$$

We also have

$$\begin{aligned}\varphi(u, u) &= \varphi((f(u) + u - (f(u) - u))/2, (f(u) + u - (f(u) - u))/2) \\ &= \frac{1}{4}\varphi(f(u) + u, f(u) + u) + \frac{1}{4}\varphi(f(u) - u, f(u) - u),\end{aligned}$$

so $f(u) + u$ and $f(u) - u$ cannot be both isotropic, since u is not isotropic.

If $f(u) - u$ is not isotropic, then the reflection $\tau_{f(u)-u}$ is such that

$$\tau_{f(u)-u}(u) = f(u),$$

and since $\tau^2_{f(u)-u} = \mathrm{id}$, if $g = \tau_{f(u)-u} \circ f$, then $g(u) = u$, and we are back to case (1). We obtain

$$f = \tau_{f(u)-u} \circ \tau_k \circ \cdots \circ \tau_1$$

where $\tau_{f(u)-u}$ and the τ_i are hyperplane reflections, with $k \geq 2n-3$, and we get a total of $2n-2$ hyperplane reflections.

If $f(u) + u$ is not isotropic, then the reflection $\tau_{f(u)+u}$ is such that

$$\tau_{f(u)+u}(u) = -f(u),$$

and since $\tau^2_{f(u)+u} = \mathrm{id}$, if $g = \tau_{f(u)+u} \circ f$, then $g(u) = -u$, and we are back to case (2). We obtain

$$f = \tau_{f(u)-u} \circ \tau \circ \tau_k \circ \cdots \circ \tau_1$$

where $\tau, \tau_{f(u)-u}$ and the τ_i are hyperplane reflections, with $k \geq 2n-3$, and we get a total of $2n-1$ hyperplane reflections. This proves the induction step. $\square$

The bound $2n-1$ is not optimal. The strong version of the Cartan–Dieudonné theorem says that at most n reflections are needed, but the proof is harder. Here is a neat proof due to E. Artin (see [6], Chapter III, Section 4).

Case 1 remains unchanged. Case 2 is slightly different: $f(u) - u \neq 0$ is not isotropic. Since $\varphi(f(u) + u, f(u) - u) = 0$, as in the first subcase of Case (3), $g = \tau_{f(u)-u} \circ f$ is such that $g(u) = u$ and we are back to Case 1. This only costs one more reflection.

The new (bad) case is:

Case 3'. $f(u) - u$ is nonzero and isotropic for all nonisotropic $u \in E$. In this case, what saves us is that E must be an Artinian space of dimension $n = 2m$ and that f must be a rotation ($f \in \mathbf{SO}(\varphi)$).

If we acccept this fact proved in Proposition 29.43 then pick any hyperplane reflection τ. Then, since f is a rotation, $g = \tau \circ f$ is *not* a rotation because $\det(g) = \det(\tau)\det(f) = (-1)(+1) = -1$, so $g(u) - u$ is either 0 or not isotropic for some nonisotropic $u \in E$ (otherwise, g would be a rotation), we are back to either Case 1 or Case 2, and using the induction hypothesis, we get

$$\tau \circ f = \tau_k \circ \ldots, \tau_1,$$

where each τ_i is a hyperplane reflection, and $k \leq 2m$. Since $\tau \circ f$ is not a rotation, actually $k \leq 2m - 1$, and then $f = \tau \circ \tau_k \circ \ldots, \tau_1$, the composition of at most $k + 1 \leq 2m$ hyperplane reflections.

Therefore, except for the fact that in Case 3', E must be an Artinian space of dimension $n = 2m$ and that f must be a rotation, which has not been proven yet, we proved the following theorem.

Theorem 29.41. *(Cartan–Dieudonné, strong form) Let φ be a nondegenerate symmetric bilinear form on a K-vector space E of dimension n ($\mathrm{char}(K) \neq 2$). Then, every isometry $f \in \mathbf{O}(\varphi)$ with $f \neq \mathrm{id}$ is the composition of at most n hyperplane reflections.*

To fill in the gap, we need two propositions.

Proposition 29.42. *Let (E, φ) be an Artinian space of dimension $2m$, and let U be a totally isotropic subspace of dimension m. For any isometry $f \in \mathbf{O}(\varphi)$, if $f(U) = U$, then $\det(f) = 1$ (f is a rotation).*

Proof. We know that we can find a basis $(u_1, \ldots, u_m, v_1, \ldots, v_m)$ of E such $(u_1, \ldots, u_m)$ is a basis of U and φ is represented by the matrix

$$\begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix}.$$

Since $f(U) = U$, the matrix representing f is of the form

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}.$$

The condition $A^\top A_{m,m} A = A_{m,m}$ translates as

$$\begin{pmatrix} B^\top & 0 \\ C^\top & D^\top \end{pmatrix} \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix} \begin{pmatrix} B & C \\ 0 & D \end{pmatrix} = \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix}$$

that is,

$$\begin{pmatrix} B^\top & 0 \\ C^\top & D^\top \end{pmatrix} \begin{pmatrix} 0 & D \\ B & C \end{pmatrix} = \begin{pmatrix} 0 & B^\top D \\ D^\top B & C^\top D + D^\top C \end{pmatrix} = \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix},$$

which implies that $B^\top D = I$, and so

$$\det(A) = \det(B)\det(D) = \det(B^\top)\det(D) = \det(B^\top D) = \det(I) = 1,$$

as claimed □

Proposition 29.43. *Let φ be a nondegenerate symmetric bilinear form on a space E of dimension n, and let f be any isometry $f \in \mathbf{O}(\varphi)$ such that $f(u) - u$ is nonzero and isotropic for every nonisotropic vector $u \in E$. Then, E is an Artinian space of dimension $n = 2m$, and f is a rotation ($f \in \mathbf{SO}(\varphi)$).*

Proof. We follow E. Artin's proof (see [6], Chapter III, Section 4). First, consider the case $n = 2$. Since we are assuming that E has some nonzero isotropic vector, by Proposition 29.26, E is an Artinian plane and there is a basis in which φ is represented by the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

we have $\varphi((x_1, x_2), (x_1, x_2)) = 2x_1x_2$, and the matrices representing isometries are of the form

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & \lambda \\ \lambda^{-1} & 0 \end{pmatrix}, \quad \lambda \in K - \{0\}.$$

In the second case,

$$\begin{pmatrix} 0 & \lambda \\ \lambda^{-1} & 0 \end{pmatrix}\begin{pmatrix} \lambda \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda \\ 1 \end{pmatrix},$$

but $u = (\lambda, 1)$ is a nonisotropic vector such that $f(u) - u = 0$. Therefore, we must be in the first case, and $\det(f) = +1$.

Let us now assume that $n \geq 3$. We are going to prove that $f(y) - y$ is isotropic for all nonzero isotropic vectors y. Let y be any nonzero isotropic vector. Since $n \geq 3$, the orthogonal space $(Ky)^\perp$ has dimension at least 2, and we know that $\mathrm{rad}(Ky) = \mathrm{rad}((Ky)^\perp)$, a space of dimension at most 1, which implies that $(Ky)^\perp$ contains some nonisotropic vector, say x. We have $\varphi(x, y) = 0$, so $\varphi(x + \epsilon y, x + \epsilon y) = \varphi(x, x) \neq 0$, for $\epsilon = \pm 1$. Then, by hypothesis, the vectors $f(x) - x, f(x + y) - (x + y) = f(x) - x + (f(y) - y)$, and $f(x - y) - (x - y) = f(x) - x - (f(y) - y)$ are isotropic. The last two vectors can be written as $f(x) - x + \epsilon(f(y) - y)$ with $\epsilon = \pm 1$, so we have

$$\begin{aligned} 0 = \varphi(f(x) - x) + \epsilon(f(y) - y), f(x) - x) + \epsilon(f(y) - y)) \\ = 2\epsilon\varphi(f(x) - x, f(y) - y)) + \epsilon^2\varphi(f(y) - y, f(y) - y). \end{aligned}$$

If we write the two equations corresponding to $\epsilon = \pm 1$, and then add them up, we get

$$\varphi(f(y) - y, f(y) - y) = 0.$$

This proves that $f(y) - y$ is isotropic for any nonzero isotropic vector y. Since by hypothesis $f(u) - u$ is isotropic for every nonisotropic vector u, we proved that $f(u) - u$ is isotropic for *every* $u \in E$. If we let $W = \mathrm{Im}(f - \mathrm{id})$, then every vector in W is isotropic, and thus W is totally isotropic (recall that we assumed that $\mathrm{char}(K) \neq 2$, so φ is determined by Φ). For any $u \in E$ and any $v \in W^{\perp}$, since W is totally isotropic, we have

$$\varphi(f(u) - u, f(v) - v) = 0,$$

and since $f(u) - u \in W$ and $v \in W^{\perp}$, we have $\varphi(f(u) - u, v) = 0$, and so

$$\begin{aligned}
0 &= \varphi(f(u) - u, f(v) - v) \\
&= \varphi(f(u), f(v)) - \varphi(u, f(v)) - \varphi(f(u) - u, v) \\
&= \varphi(u, v) - \varphi(u, f(v)) \\
&= \varphi(u, v - f(v)),
\end{aligned}$$

for all $u \in E$. Since φ is nonsingular, this means that $f(v) = v$, for all $v \in W^{\perp}$. However, by hypothesis, no nonisotropic vector is left fixed, which implies that $W^{\perp}$ is also totally isotropic. In summary, we proved that $W \subseteq W^{\perp}$ and $W^{\perp} \subseteq W^{\perp\perp} = W$, that is,

$$W = W^{\perp}.$$

Since, $\dim(W) + \dim(W^{\perp}) = n$, we conclude that W is a totally isotropic subspace of E such that

$$\dim(W) = n/2.$$

By Proposition 29.29, the space E is an Artinian space of dimension $n = 2m$. Since $W = W^{\perp}$ and $f(W) = W$, by Proposition 29.42, the isometry f is a rotation. $\square$

Remarks:

1. Another way to finish the proof of Proposition 29.43 is to prove that if f is an isometry, then

$$\mathrm{Ker}\,(f - \mathrm{id}) = (\mathrm{Im}(f - \mathrm{id}))^{\perp}.$$

   After having proved that $W = \mathrm{Im}(f - \mathrm{id})$ is totally isotropic, we get

$$\mathrm{Ker}\,(f - \mathrm{id}) = \mathrm{Im}(f - \mathrm{id}),$$

   which implies that $(f - \mathrm{id})^2 = 0$. From this, we deduce that $\det(f) = 1$. For details, see Jacobson [98] (Chapter 6, Section 6).

2. If $f = \tau_{H_k} \circ \cdots \circ \tau_{H_1}$, where the H_i are hyperplanes, then it can be shown that

$$\dim(H_1 \cap H_2 \cap \cdots \cap H_s) \geq n - s.$$

   Now, since each H_i is left fixed by τ_{H_i}, we see that every vector in $H_1 \cap \cdots \cap H_s$ is left fixed by f. In particular, if $s < n$, then f has some nonzero fixed point. As a consequence, an isometry without fixed points requires n hyperplane reflections.

29.10 Witt's Theorem

Witt's theorem was referred to as a "scandal" by Emil Artin. What he meant by this is that one had to wait until 1936 (Witt [190]) to formulate and prove a theorem at once so simple in its statement and underlying concepts, and so useful in various domains (geometry, arithmetic of quadratic forms).[1]

Besides Witt's original proof (Witt [190]), Chevalley's proof [37] seems to be the "best" proof that applies to the symmetric as well as the skew-symmetric case. The proof in Bourbaki [24] is based on Chevalley's proof, and so are a number of other proofs. This is the one we follow (slightly reorganized). In the symmetric case, Serre's exposition is hard to beat (see Serre [157], Chapter IV).

The following observation is one of the key ingredients in the proof of Theorem 29.45.

Proposition 29.44. *Given a finite-dimensional space E equipped with an ϵ-Hermitan form φ, if U_1 and U_2 are two subspaces of E such that $U_1 \cap U_2 = (0)$ and if we have metric linear maps $f_1 \colon U_1 \to E$ and $f_2 \colon U_2 \to E$ such that*

$$\varphi(f_1(u_1), f_2(u_2)) = \varphi(u_1, u_2) \quad \textit{for } u_i \in U_i \ (i = 1, 2), \tag{$*$}$$

then the linear map $f \colon U_1 \oplus U_2 \to E$ given by $f(u_1 + u_2) = f_1(u_1) + f_2(u_2)$ extends f_1 and f_2 and is metric. Furthermore, if f_1 and f_2 are injective, then so if f.

Proof. Indeed, since f_1 and f_2 are metric and using $(*)$, we have

$$\begin{aligned}
\varphi(f_1(u_1) + f_2(u_2), f_1(v_1) + f_2(v_2)) &= \varphi(f_1(u_1), f_1(v_1)) + \varphi(f_1(u_1), f_2(v_2)) \\
&\quad + \varphi(f_2(u_2), f_1(v_1)) + \varphi(f_2(u_2), f_2(v_2)) \\
&= \varphi(u_1, v_1) + \varphi(u_1, v_2) + \varphi(u_2, v_1) + \varphi(u_2, v_2) \\
&= \varphi(u_1 + u_2, v_2 + v_2).
\end{aligned}$$

Thus f is a metric map extending f_1 and f_2. ☐

Theorem 29.45. *(Witt, 1936) Let E and E' be two finite-dimensional spaces respectively equipped with two nondegenerate ϵ-Hermitan forms φ and φ' satisfying condition (T), and assume that there is an isometry between (E, φ) and (E', φ'). For any subspace U of E, every injective metric linear map f from U into E' extends to an isometry from E to E'.*

Proof. Since (E, φ) and (E', φ') are isometric, we may assume that $E' = E$ and $\varphi' = \varphi$ (if $h \colon E \to E'$ is an isometry, then $h^{-1} \circ f$ is an injective metric map from U into E. The details are left to the reader).

[1] Curiously, some references to Witt's paper claim its date of publication to be 1936, but others say 1937. The answer to this mystery is that Volume 176 of *Crelle Journal* was published in four issues. The cover page of volume 176 mentions the year 1937, but Witt's paper is dated May 1936. This is not the only paper of Witt appearing in this volume!

We proceed by induction on the dimension r of U. Since the proof is quite intricate, we spell out the general plan of attack. For the induction step, we first show that we can reduce the situation to what we call *Case (H)*, namely that the subspace of U left fixed by f is a hyperplane H in U. Then, the set $D = \{f(u) - u \mid u \in U\}$ is a line in U and it turns out that $D^{\perp}$ is a hyperplane in E. We now introduce *Hypothesis (V)*, which says we can find a nontrivial subspace V of E orthogonal to D and such that $V \cap U = V \cap f(U) = (0)$. We show that if Hypothesis (V) holds, then f can be extended to an isometry of $U \oplus V$. It is then possible to further extend f to an isometry of E.

To prove that Hypothesis (V) holds we consider two cases. In Case (a), we obtain some V such that $E = U \oplus V$ and we are done. In Case (b), we obtain some V such that $D^{\perp} = U \oplus V$. We are then reduced to the situation where $U = D^{\perp}$ is a hyperplane in E and f is an isometry of U. To finish the proof we pick any $v \notin U$, so that $E = U \oplus Kv$, and we find some $v_1 \in E$ such that

$$\begin{aligned} \varphi(f(u), v_1) &= \varphi(u, v) \quad \text{for all } u \in U \\ \varphi(v_1, v_1) &= \varphi(v, v). \end{aligned}$$

Then, by Proposition 29.44, we can extend f to a metric map g of $U + Kv = E$ such that $g(v) = v_1$. The argument used to find v_1 makes use of (†) (see below) and is bit tricky. We also makes use of Property (T) in the form of Lemma 29.28.

We now go back to the proof. The case $r = 0$ is trivial. For the induction step, $r \geq 1$ so $U \neq (0)$, and let H be any hyperplane in U. Let $f\colon U \to E$ be an injective metric linear map. By the induction hypothesis, the restriction f_0 of f to H extends to an isometry g_0 of E. If g_0 extends f, we are done. Otherwise, H is the subspace of elements of U left fixed by $g_0^{-1} \circ f$. If the theorem holds in this situation, namely the subspace of U left fixed by $g_0^{-1} \circ f$ is a hyperplane H in U, then we have an isometry g_1 of E extending $g_0^{-1} \circ f$, and $g_0 \circ g_1$ is an isometry of E extending f. Therefore, we are reduced to the following situation:

Case (H). The subspace of U left fixed by f is a hyperplane H in U.

In this case, the set $D = \{f(u) - u \mid u \in U\}$ is a line in U (a one-dimensional subspace). For all $u, v \in U$, we have

$$\varphi(f(u), f(v) - v) = \varphi(f(u), f(v)) - \varphi(f(u), v) = \varphi(u, v) - \varphi(f(u), v) = \varphi(u - f(u), v),$$

that is

$$\varphi(f(u), f(v) - v) = \varphi(u - f(u), v) \quad \text{for all } u, v \in U, \tag{$**$}$$

and if $u \in H$, which means that $f(u) = u$, we get $u \in D^{\perp}$. Therefore, $H \subseteq D^{\perp}$. Since φ is nondegenerate, we have $\dim(D) + \dim(D^{\perp}) = \dim(E)$, and since $\dim(D) = 1$, the subspace $D^{\perp}$ is a hyperplane in E.

Hypothesis (V). We can find a nontrivial subspace V of E orthogonal to D and such that $V \cap U = V \cap f(U) = (0)$.

Claim. Hypothesis (V) implies that f can be extended to an isometry of $U \oplus V$.

Proof of Claim. If Hypothesis (V) holds, then we have

$$\varphi(f(u), v) = \varphi(u, v) \quad \text{for all } u \in U \text{ and all } v \in V,$$

since $\varphi(f(u), v) - \varphi(u, v) = \varphi(f(u) - u, v) = 0$, with $f(u) - u \in D$ and $v \in V$ orthogonal to D. By Proposition 29.44 with $f_1 = f$ and f_2 the inclusion of V into E, we can extend f to an injective metric map on $U \oplus V$ leaving all vectors in V fixed. In this case, the set $\{f(w) - w \mid w \in U \oplus V\}$ is still the line D. ☐

We show below that the fact that f can be extended to $U \oplus V$ implies that f can be extended to the whole of E. There are two cases. In Case (a), $E = U \oplus V$ and we are done. In case (b), $D^\perp = U \oplus V$ where $D^\perp$ is a hyperplane in E and f is an isometry of $D^\perp$. By a subtle argument, we will show that f can be extended to an isometry of E.

We are reduced to proving that a subspace V as above exists. We distinguish between two cases.

Case (a). $U \not\subseteq D^\perp$.

Proof of Case (a). In this case, formula $(**)$ show that $f(U)$ is not contained in $D^\perp$ (check this!). Consequently,

$$U \cap D^\perp = f(U) \cap D^\perp = H.$$

We can pick V to be any supplement of H in $D^\perp$, and the above formula shows that $V \cap U = V \cap f(U) = (0)$. Since $U \oplus V$ contains the hyperplane $D^\perp$ (since $D^\perp = H \oplus V$ and $H \subseteq U$), and $U \oplus V \neq D^\perp$ (since U is not contained in $D^\perp$ and $V \subseteq D^\perp$), we must have $E = U \oplus V$, and as we showed as a consequence of hypothesis (V), f can be extended to an isometry of $U \oplus V = E$. ☐

Case (b). $U \subseteq D^\perp$.

Proof of Case (b). In this case, formula $(**)$ shows that $f(U) \subseteq D^\perp$ so $U + f(U) \subseteq D^\perp$, and since $D = \{f(u) - u \mid u \in U\}$, we have $D \subseteq D^\perp$; that is, the line D is isotropic.

We show that there exists a subspace V of $D^\perp$, such that

$$D^\perp = U \oplus V = f(U) \oplus V.$$

Thus, case (b) shows that we are reduced to the situation where $U = D^\perp$ and f is an isometry of U.

If $U = f(U)$ we pick V to be a supplement of U in $D^\perp$. Otherwise, let $x \in U$ with $x \notin H$, and let $y \in f(U)$ with $y \notin H$. Since $f(H) = H$ (pointwise), f is injective, and H is a hyperplane in U, we have

$$U = H \oplus Kx, \quad f(U) = H \oplus Ky.$$

We claim that $x + y \notin U$. Otherwise, since $y = x + y - x$, with $x + y, x \in U$ and since $y \in f(U)$, we would have $y \in U \cap f(U) = H$, a contradiction. Similarly, $x + y \notin f(U)$. It follows that

$$U + f(U) = U \oplus K(x + y) = f(U) \oplus K(x + y).$$

Now, pick W to be any supplement of $U + f(U)$ in $D^\perp$ so that $D^\perp = (U + f(U)) \oplus W$, and let

$$V = K(x + y) + W.$$

Then, since $x \in U, y \in f(U)$, $W \subseteq D^\perp$, and $U + f(U) \subseteq D^\perp$, we have $V \subseteq D^\perp$. We also have

$$U \oplus V = U \oplus K(x + y) \oplus W = (U + f(U)) \oplus W = D^\perp$$

and

$$f(U) \oplus V = f(U) \oplus K(x + y) \oplus W = (U + f(U)) \oplus W = D^\perp,$$

so as we showed as a consequence of hypothesis (V), f can be extended to an isometry of the hyperplane $D^\perp = U \oplus V$, and D is still the line $\{f(w) - w \mid w \in U \oplus V\}$. $\square$

The argument in the proof of Case (b) shows that we are reduced to the situation where $U = D^\perp$ is a hyperplane in E and f is an isometry of U. If we pick any $v \notin U$, then $E = U \oplus Kv$, so suppose we can find some $v_1 \in E$ such that

$$\begin{aligned} \varphi(f(u), v_1) &= \varphi(u, v) \quad \text{for all } u \in U \\ \varphi(v_1, v_1) &= \varphi(v, v). \end{aligned}$$

The first condition is condition $(*)$ of Proposition 29.44, and the second condition asserts that the map $\lambda v \mapsto \lambda v_2$ from the line Kv to the line Kv_1 is a metric map. Then, by Proposition 29.44, we can extend f to a metric map g of $U + Kv = E$ such that $g(v) = v_1$.

To find v_1, let us prove that for every $v \in E$, there is some $v' \in E$ such that

$$\varphi(f(u), v') = \varphi(u, v) \quad \text{for all } u \in U. \tag{$\dagger$}$$

This is because the linear form $u \mapsto \varphi(f^{-1}(u), v)$ ($u \in U$) is the restriction of a linear form $\psi \in E^*$, and since φ is nondegenerate, there is some (unique) $v' \in E$, such that

$$\psi(x) = \varphi(x, v') \quad \text{for all } x \in E,$$

which implies that

$$\varphi(u, v') = \varphi(f^{-1}(u), v) \quad \text{for all } u \in U,$$

and since f is an automorphism of U, that ($\dagger$) holds. Furthermore, observe that formula ($\dagger$) still holds if we add to v' any vector y in D, since $f(U) = U = D^\perp$. Therefore, for any $v_1 = v' + y$ with $y \in D$, if we extend f to a linear map of E by setting $g(v) = v_1$, then by ($\dagger$) we have

$$\varphi(g(u), g(v)) = \varphi(u, v) \quad \text{for all } u \in U.$$

We still need to pick $y \in D$ so that $v_1 = v' + y$ satisfies $\varphi(v_1, v_1) = \varphi(v, v)$. However, since $v \notin U = D^{\perp}$, the vector v is not orthogonal D, and by Lemma 29.28, there is some $y_0 \in D$ such that

$$\varphi(v' + y_0, v' + y_0) = \varphi(v, v).$$

Then, if we let $v_1 = v' + y_0$, by Proposition 29.44, we can extend f to a metric map g of $U + Kv = E$ by setting $g(v) = v_1$. Since φ is nondegenerate, g is an isometry. $\square$

The first corollary of Witt's theorem is sometimes called the Witt's cancellation theorem.

Theorem 29.46. *(Witt Cancellation Theorem) Let (E_1, φ_1) and (E_2, φ_2) be two pairs of finite-dimensional spaces and nondegenerate ϵ-Hermitian forms satisfying condition (T), and assume that (E_1, φ_1) and (E_2, φ_2) are isometric. For any subspace U of E_1 and any subspace V of E_2, if there is an isometry $f \colon U \to V$, then there is an isometry $g \colon U^{\perp} \to V^{\perp}$.*

Proof. If $f \colon U \to V$ is an isometry between U and V, by Witt's theorem (Theorem 29.46), the linear map f extends to an isometry g between E_1 and E_2. We claim that g maps $U^{\perp}$ into $V^{\perp}$. This is because if $v \in U^{\perp}$, we have $\varphi_1(u, v) = 0$ for all $u \in U$, so

$$\varphi_2(g(u), g(v)) = \varphi_1(u, v) = 0 \quad \text{for all } u \in U,$$

and since g is a bijection between U and V, we have $g(U) = V$, so we see that $g(v)$ is orthogonal to V for every $v \in U^{\perp}$; that is, $g(U^{\perp}) \subseteq V^{\perp}$. Since g is a metric map and since φ_1 is nondegenerate, the restriction of g to $U^{\perp}$ is an isometry from $U^{\perp}$ to $V^{\perp}$. $\square$

A pair (E, φ) where E is finite-dimensional and φ is a nondegenerate ϵ-Hermitian form is often called an *ϵ-Hermitian space*. When $\epsilon = 1$ and φ is symmetric, we use the term *Euclidean* space or *quadratic* space. When $\epsilon = -1$ and φ is alternating, we use the term *symplectic* space. When $\epsilon = 1$ and the automorphism $\lambda \mapsto \overline{\lambda}$ is not the identity we use the term *Hermitian* space, and when $\epsilon = -1$, we use the term *skew-Hermitian* space.

We also have the following result showing that the group of isometries of an ϵ-Hermitian space is transitive on totally isotropic subspaces of the same dimension.

Theorem 29.47. *Let E be a finite-dimensional vector space and let φ be a nondegenerate ϵ-Hermitian form on E satisfying condition (T). Then for any two totally isotropic subspaces U and V of the same dimension, there is an isometry $f \in \mathbf{Isom}(\varphi)$ such that $f(U) = V$. Furthermore, every linear automorphism of U is induced by an isometry of E.*

Remark: Witt's cancelation theorem can be used to define an equivalence relation on ϵ-Hermitian spaces and to define a group structure on these equivalence classes. This way, we obtain the *Witt group*, but we will not discuss it here.

Witt's Theorem can be sharpened to isometries in $\mathbf{SO}(\varphi)$, but some condition on U is needed.

Theorem 29.48. *(Witt–Sharpened Version) Let E be a finite-dimensional space equipped with a nondegenerate symmetric bilinear forms φ. For any subspace U of E, every linear injective metric map f from U into E extends to an isometry g of E with a prescribed value ± 1 of $\det(g)$ iff*

$$\dim(U) + \dim(\mathrm{rad}(U)) < \dim(E) = n.$$

If

$$\dim(U) + \dim(\mathrm{rad}(U)) = \dim(E) = n,$$

and $\det(f) = -1$, then there is no $g \in \mathbf{SO}(\varphi)$ extending f.

Proof. If g_1 and g_2 are two extensions of f such that $\det(g_1)\det(g_2) = -1$, then $h = g_1^{-1} \circ g_2$ is an isometry such that $\det(h) = -1$, and h leaves every vector of U fixed. Conversely, if h is an isometry such that $\det(h) = -1$, and $h(u) = u$ for all $u \in U$, then for any extesnion g_1 of f, the map $g_2 = h \circ g_1$ is another extension of f such that $\det(g_2) = -\det(g_1)$. Therefore, we need to show that a map h as above exists.

If $\dim(U) + \dim(\mathrm{rad}(U)) < \dim(E)$, consider the nondegenerate completion $\overline{U}$ of U given by Proposition 29.32. We know that $\dim(\overline{U}) = \dim(U) + \dim(\mathrm{rad}(U)) < n$, and since $\overline{U}$ is nondegenerate, we have

$$E = \overline{U} \overset{\perp}{\oplus} \overline{U}^{\perp},$$

with $\overline{U}^{\perp} \neq (0)$. Pick any isometry τ of $\overline{U}^{\perp}$ such that $\det(\tau) = -1$, and extend it to an isometry h of E whose restriction to $\overline{U}$ is the identity.

If $\dim(U) + \dim(\mathrm{rad}(U)) = \dim(E) = n$, then $U = V \overset{\perp}{\oplus} W$ with $V = \mathrm{rad}(U)$ and since $\dim(\overline{U}) = \dim(U) + \dim(\mathrm{rad}(U)) = n$, we have

$$E = \overline{U} = (V \oplus V') \overset{\perp}{\oplus} W,$$

where $V \oplus V' = \mathrm{Ar}_{2r} = W^{\perp}$ is an Artinian space. Any isometry h of E which is the identity on U and with $\det(h) = -1$ is the identity on W, and thus it must map $W^{\perp} = \mathrm{Ar}_{2r} = V \oplus V'$ into itself, and the restriction h' of h to Ar_{2r} has $\det(h') = -1$. However, h' is the identity on $V = \mathrm{rad}(U)$, a totally isotropic subspace of Ar_{2r} of dimension r, and by Proposition 29.42, we have $\det(h') = +1$, a contradiction. □

It can be shown that the center of $\mathbf{O}(\varphi)$ is $\{\mathrm{id}, -\mathrm{id}\}$. For further properties of orthogonal groups, see Grove [83], Jacobson [98], Taylor [174], and Artin [6].

Part IV

Algebra: PID's, UFD's, Noetherian Rings, Tensors, Modules over a PID, Normal Forms

Chapter 30

Polynomials, Ideals and PID's

30.1 Multisets

This chapter contains a review of polynomials and their basic properties. First, multisets are defined. Polynomials in one variable are defined next. The notion of a polynomial function in one argument is defined. Polynomials in several variable are defined, and so is the notion of a polynomial function in several arguments. The Euclidean division algorithm is presented, and the main consequences of its existence are derived. Ideals are defined, and the characterization of greatest common divisors of polynomials in one variables (gcd's) in terms of ideals is shown. We also prove the Bezout identity. Next, we consider the factorization of polynomials in one variables into irreducible factors. The unique factorization of polynomials in one variable into irreducible factors is shown. Roots of polynomials and their multiplicity are defined. It is shown that a nonnull polynomial in one variable and of degree m over an integral domain has at most m roots. The chapter ends with a brief treatment of polynomial interpolation: Lagrange, Newton, and Hermite interpolants are introduced.

In this chapter, it is assumed that all rings considered are commutative. Recall that a (commutative) ring A is an *integral domain* (or an *entire ring*) if $1 \neq 0$, and if $ab = 0$, then either $a = 0$ or $b = 0$, for all $a, b \in A$. This second condition is equivalent to saying that if $a \neq 0$ and $b \neq 0$, then $ab \neq 0$. Also, recall that $a \neq 0$ is *not* a zero divisor if $ab \neq 0$ whenever $b \neq 0$. Observe that a field is an integral domain.

Our goal is to define polynomials in one or more indeterminates (or variables) $X_1, \ldots, X_n$, with coefficients in a ring A. This can be done in several ways, and we choose a definition that has the advantage of extending immediately from one to several variables. First, we need to review the notion of a (finite) multiset.

Definition 30.1. Given a set I, a *(finite) multiset over* I is any function $M\colon I \to \mathbb{N}$ such that $M(i) \neq 0$ for finitely many $i \in I$. The multiset M such that $M(i) = 0$ for all $i \in I$ is the *empty multiset*, and it is denoted by 0. If $M(i) = k \neq 0$, we say that i *is a member of* M *of multiplicity* k. The *union* $M_1 + M_2$ of two multisets M_1 and M_2 is defined such that $(M_1 + M_2)(i) = M_1(i) + M_2(i)$, for every $i \in I$. If I is finite, say $I = \{1, \ldots, n\}$, the multiset

M such that $M(i) = k_i$ for every i, $1 \leq i \leq n$, is denoted by $k_1 \cdot 1 + \cdots + k_n \cdot n$, or more simply, by $(k_1, \ldots, k_n)$, and $\deg(k_1 \cdot 1 + \cdots + k_n \cdot n) = k_1 + \cdots + k_n$ is the *size* or *degree* of M. The set of all multisets over I is denoted by $\mathbb{N}^{(I)}$, and when $I = \{1, \ldots, n\}$, by $\mathbb{N}^{(n)}$.

Intuitively, the order of the elements of a multiset is irrelevant, but the multiplicity of each element is relevant, contrary to sets. Every $i \in I$ is identified with the multiset M_i such that $M_i(i) = 1$ and $M_i(j) = 0$ for $j \neq i$. When $I = \{1\}$, the set $\mathbb{N}^{(1)}$ of multisets $k \cdot 1$ can be identified with $\mathbb{N}$ and $\{1\}^*$. We will denote $k \cdot 1$ simply by k.

However, beware that when $n \geq 2$, the set $\mathbb{N}^{(n)}$ of multisets cannot be identified with the set of strings in $\{1, \ldots, n\}^*$, because multiset union is commutative, but concatenation of strings in $\{1, \ldots, n\}^*$ is not commutative when $n \geq 2$. This is because in a multiset $k_1 \cdot 1 + \cdots + k_n \cdot n$, the order is irrelevant, whereas in a string, the order is relevant. For example, $2 \cdot 1 + 3 \cdot 2 = 3 \cdot 2 + 2 \cdot 1$, but $11222 \neq 22211$, as strings over $\{1, 2\}$.

Nevertheless, $\mathbb{N}^{(n)}$ and the set $\mathbb{N}^n$ of ordered n-tuples under component-wise addition are isomorphic under the map

$$k_1 \cdot 1 + \cdots + k_n \cdot n \mapsto (k_1, \ldots, k_n).$$

Thus, since the notation $(k_1, \ldots, k_n)$ is less cumbersome that $k_1 \cdot 1 + \cdots + k_n \cdot n$, it will be preferred. We just have to remember that the order of the k_i is really irrelevant.

But when I is infinite, beware that $\mathbb{N}^{(I)}$ and the set $\mathbb{N}^I$ of ordered I-tuples are not isomorphic.

We are now ready to define polynomials.

30.2 Polynomials

We begin with polynomials in one variable.

Definition 30.2. Given a ring A, we define the set $\mathcal{P}_A(1)$ of *polynomials over A in one variable* as the set of functions $P \colon \mathbb{N} \to A$ such that $P(k) \neq 0$ for finitely many $k \in \mathbb{N}$. The polynomial such that $P(k) = 0$ for all $k \in \mathbb{N}$ is the *null (or zero) polynomial* and it is denoted by 0. We define addition of polynomials, multiplication by a scalar, and multiplication of polynomials, as follows: Given any three polynomials $P, Q, R \in \mathcal{P}_A(1)$, letting $a_k = P(k)$, $b_k = Q(k)$, and $c_k = R(k)$, for every $k \in \mathbb{N}$, we define $R = P + Q$ such that

$$c_k = a_k + b_k,$$

$R = \lambda P$ such that

$$c_k = \lambda a_k,$$

where $\lambda \in A$,

and $R = PQ$ such that

$$c_k = \sum_{i+j=k} a_i b_j.$$

We define the polynomial e_k such that $e_k(k) = 1$ and $e_k(i) = 0$ for $i \neq k$. We also denote e_0 by 1 when $k = 0$. Given a polynomial P, the $a_k = P(k) \in A$ are called the *coefficients of* P. If P is not the null polynomial, there is a greatest $n \geq 0$ such that $a_n \neq 0$ (and thus, $a_k = 0$ for all $k > n$) called the *degree of* P and denoted by $\deg(P)$. Then, P is written uniquely as

$$P = a_0e_0 + a_1e_1 + \cdots + a_ne_n.$$

When P is the null polynomial, we let $\deg(P) = -\infty$.

There is an injection of A into $\mathcal{P}_A(1)$ given by the map $a \mapsto a1$ (recall that 1 denotes e_0). There is also an injection of $\mathbb{N}$ into $\mathcal{P}_A(1)$ given by the map $k \mapsto e_k$. Observe that $e_k = e_1^k$ (with $e_1^0 = e_0 = 1$). In order to alleviate the notation, we often denote e_1 by X, and we call X a *variable (or indeterminate)*. Then, $e_k = e_1^k$ is denoted by X^k. Adopting this notation, given a nonnull polynomial P of degree n, if $P(k) = a_k$, P is denoted by

$$P = a_0 + a_1X + \cdots + a_nX^n,$$

or by

$$P = a_nX^n + a_{n-1}X^{n-1} + \cdots + a_0,$$

if this is more convenient (the order of the terms does not matter anyway). Sometimes, it will also be convenient to write a polynomial as

$$P = a_0X^n + a_1X^{n-1} + \cdots + a_n.$$

The set $\mathcal{P}_A(1)$ is also denoted by $A[X]$ and a polynomial P may be denoted by $P(X)$. In denoting polynomials, we will use both upper-case and lower-case letters, usually, P, Q, R, S, p, q, r, s, but also f, g, h, etc., if needed (as long as no ambiguities arise).

Given a nonnull polynomial P of degree n, the nonnull coefficient a_n is called the *leading coefficient of* P. The coefficient a_0 is called the *constant term* of P. A polynomial of the form a_kX^k is called a *monomial*. We say that a_kX^k *occurs* in P if $a_k \neq 0$. A nonzero polynomial P of degree n is called a *monic polynomial (or unitary polynomial, or monic)* if $a_n = 1$, where a_n is its leading coefficient, and such a polynomial can be written as

$$P = X^n + a_{n-1}X^{n-1} + \cdots + a_0 \qquad \text{or} \qquad P = X^n + a_1X^{n-1} + \cdots + a_n.$$

The choice of the variable X to denote e_1 is standard practice, but there is nothing special about X. We could have chosen Y, Z, or any other symbol, as long as no ambiguities arise.

Formally, the definition of $\mathcal{P}_A(1)$ has nothing to do with X. The reason for using X is simply convenience. Indeed, it is more convenient to write a polynomial as $P = a_0 + a_1X + \cdots + a_nX^n$ rather than as $P = a_0e_0 + a_1e_1 + \cdots + a_ne_n$.

We have the following simple but crucial proposition.

Proposition 30.1. *Given two nonnull polynomials $P(X) = a_0+a_1X+\cdots+a_mX^m$ of degree m and $Q(X) = b_0 + b_1X + \cdots + b_nX^n$ of degree n, if either a_m or b_n is not a zero divisor, then $a_mb_n \neq 0$, and thus, $PQ \neq 0$ and*

$$\deg(PQ) = \deg(P) + \deg(Q).$$

In particular, if A is an integral domain, then $A[X]$ is an integral domain.

Proof. Since the coefficient of X^{m+n} in PQ is a_mb_n, and since we assumed that either a_m or a_n is not a zero divisor, we have $a_mb_n \neq 0$, and thus, $PQ \neq 0$ and

$$\deg(PQ) = \deg(P) + \deg(Q).$$

Then, it is obvious that $A[X]$ is an integral domain. $\square$

It is easily verified that $A[X]$ is a commutative ring, with multiplicative identity $1X^0 = 1$. It is also easily verified that $A[X]$ satisfies all the conditions of Definition 3.1, but $A[X]$ is not a vector space, since A is not necessarily a field.

A structure satisfying the axioms of Definition 3.1 when K is a ring (and not necessarily a field) is called a *module*. Modules fail to have some of the nice properties that vector spaces have, and thus, they are harder to study. For example, there are modules that do not have a basis. We postpone the study of modules until Chapter 35.

However, when the ring A is a field, $A[X]$ is a vector space. But even when A is just a ring, the family of polynomials $(X^k)_{k\in\mathbb{N}}$ is a basis of $A[X]$, since every polynomial $P(X)$ can be written in a unique way as $P(X) = a_0 + a_1X + \cdots + a_nX^n$ (with $P(X) = 0$ when $P(X)$ is the null polynomial). Thus, $A[X]$ is a free module.

Next, we want to define the notion of evaluating a polynomial $P(X)$ at some $\alpha \in A$. For this, we need a proposition.

Proposition 30.2. *Let A, B be two rings and let $h\colon A \to B$ be a ring homomorphism. For any $\beta \in B$, there is a unique ring homomorphism $\varphi\colon A[X] \to B$ extending h such that $\varphi(X) = \beta$, as in the following diagram (where we denote by $h+\beta$ the map $h+\beta\colon A\cup\{X\} \to B$ such that $(h+\beta)(a) = h(a)$ for all $a \in A$ and $(h+\beta)(X) = \beta$):*

$$\begin{tikzcd}
A \cup \{X\} \arrow[r, "\iota"] \arrow[dr, "h+\beta"'] & A[X] \arrow[d, "\varphi"] \\
 & B
\end{tikzcd}$$

Proof. Let $\varphi(0) = 0$, and for every nonnull polynomial $P(X) = a_0 + a_1X + \cdots + a_nX^n$, let
$$\varphi(P(X)) = h(a_0) + h(a_1)\beta + \cdots + h(a_n)\beta^n.$$
It is easily verified that φ is the unique homomorphism $\varphi\colon A[X] \to B$ extending h such that $\varphi(X) = \beta$. $\square$

Taking $A = B$ in Proposition 30.2 and $h\colon A \to A$ the identity, for every $\beta \in A$, there is a unique homomorphism $\varphi_\beta\colon A[X] \to A$ such that $\varphi_\beta(X) = \beta$, and for every polynomial $P(X)$, we write $\varphi_\beta(P(X))$ as $P(\beta)$ and we call $P(\beta)$ the *value of* $P(X)$ *at* $X = \beta$. Thus, we can define a function $P_A\colon A \to A$ such that $P_A(\beta) = P(\beta)$, for all $\beta \in A$. This function is called the *polynomial function induced by* P.

More generally, P_B can be defined for any (commutative) ring B such that $A \subseteq B$. In general, it is possible that $P_A = Q_A$ for distinct polynomials P, Q. We will see shortly conditions for which the map $P \mapsto P_A$ is injective. In particular, this is true for $A = \mathbb{R}$ (in general, any infinite integral domain). We now define polynomials in n variables.

Definition 30.3. Given $n \geq 1$ and a ring A, the set $\mathcal{P}_A(n)$ of *polynomials over* A *in* n *variables* is the set of functions $P\colon \mathbb{N}^{(n)} \to A$ such that $P(k_1, \ldots, k_n) \neq 0$ for finitely many $(k_1, \ldots, k_n) \in \mathbb{N}^{(n)}$. The polynomial such that $P(k_1, \ldots, k_n) = 0$ for all $(k_1, \ldots, k_n)$ is the *null (or zero) polynomial* and it is denoted by 0. We define addition of polynomials, multiplication by a scalar, and multiplication of polynomials, as follows: Given any three polynomials $P, Q, R \in \mathcal{P}_A(n)$, letting $a_{(k_1,\ldots,k_n)} = P(k_1, \ldots, k_n)$, $b_{(k_1,\ldots,k_n)} = Q(k_1, \ldots, k_n)$, $c_{(k_1,\ldots,k_n)} = R(k_1, \ldots, k_n)$, for every $(k_1, \ldots, k_n) \in \mathbb{N}^{(n)}$, we define $R = P + Q$ such that
$$c_{(k_1,\ldots,k_n)} = a_{(k_1,\ldots,k_n)} + b_{(k_1,\ldots,k_n)},$$
$R = \lambda P$, where $\lambda \in A$, such that
$$c_{(k_1,\ldots,k_n)} = \lambda a_{(k_1,\ldots,k_n)},$$
and $R = PQ$, such that
$$c_{(k_1,\ldots,k_n)} = \sum_{(i_1,\ldots,i_n)+(j_1,\ldots,j_n)=(k_1,\ldots,k_n)} a_{(i_1,\ldots,i_n)} b_{(j_1,\ldots,j_n)}.$$

For every $(k_1, \ldots, k_n) \in \mathbb{N}^{(n)}$, we let $e_{(k_1,\ldots,k_n)}$ be the polynomial such that
$$e_{(k_1,\ldots,k_n)}(k_1, \ldots, k_n) = 1 \quad \text{and} \quad e_{(k_1,\ldots,k_n)}(h_1, \ldots, h_n) = 0,$$
for $(h_1, \ldots, h_n) \neq (k_1, \ldots, k_n)$. We also denote $e_{(0,\ldots,0)}$ by 1. Given a polynomial P, the $a_{(k_1,\ldots,k_n)} = P(k_1, \ldots, k_n) \in A$, are called the *coefficients of* P. If P is not the null polynomial, there is a greatest $d \geq 0$ such that $a_{(k_1,\ldots,k_n)} \neq 0$ for some $(k_1, \ldots, k_n) \in \mathbb{N}^{(n)}$, with $d = k_1 + \cdots + k_n$, called the *total degree of* P and denoted by $\deg(P)$. Then, P is written uniquely as
$$P = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} e_{(k_1,\ldots,k_n)}.$$
When P is the null polynomial, we let $\deg(P) = -\infty$.

There is an injection of A into $\mathcal{P}_A(n)$ given by the map $a \mapsto a1$ (where 1 denotes $e_{(0,\ldots,0)}$). There is also an injection of $\mathbb{N}^{(n)}$ into $\mathcal{P}_A(n)$ given by the map $(h_1, \ldots, h_n) \mapsto e_{(h_1,\ldots,h_n)}$. Note that $e_{(h_1,\ldots,h_n)}e_{(k_1,\ldots,k_n)} = e_{(h_1+k_1,\ldots,h_n+k_n)}$. In order to alleviate the notation, let $X_1, \ldots, X_n$ be n distinct variables and denote $e_{(0,\ldots,0,1,0\ldots,0)}$, where 1 occurs in the position i, by X_i (where $1 \leq i \leq n$). With this convention, in view of $e_{(h_1,\ldots,h_n)}e_{(k_1,\ldots,k_n)} = e_{(h_1+k_1,\ldots,h_n+k_n)}$, the polynomial $e_{(k_1,\ldots,k_n)}$ is denoted by $X_1^{k_1} \cdots X_n^{k_n}$ (with $e_{(0,\ldots,0)} = X_1^0 \cdots X_n^0 = 1$) and it is called a *primitive monomial*. Then, P is also written as

$$P = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}.$$

We also denote $\mathcal{P}_A(n)$ by $A[X_1, \ldots, X_n]$. A polynomial $P \in A[X_1, \ldots, X_n]$ is also denoted by $P(X_1, \ldots, X_n)$.

As in the case $n = 1$, there is nothing special about the choice of $X_1, \ldots, X_n$ as variables (or indeterminates). It is just a convenience. After all, the construction of $\mathcal{P}_A(n)$ has nothing to do with $X_1, \ldots, X_n$.

Given a nonnull polynomial P of degree d, the nonnull coefficients $a_{(k_1,\ldots,k_n)} \neq 0$ such that $d = k_1 + \cdots + k_n$ are called the *leading coefficients of* P. A polynomial of the form $a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}$ is called a *monomial*. Note that $\deg(a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}) = k_1+\cdots+k_n$. Given a polynomial

$$P = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n},$$

a monomial $a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}$ *occurs in the polynomial* P if $a_{(k_1,\ldots,k_n)} \neq 0$.

A polynomial

$$P = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}$$

is *homogeneous of degree* d if

$$\deg(X_1^{k_1} \cdots X_n^{k_n}) = d,$$

for every monomial $a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}$ occurring in P. If P is a polynomial of total degree d, it is clear that P can be written uniquely as

$$P = P^{(0)} + P^{(1)} + \cdots + P^{(d)},$$

where $P^{(i)}$ is the sum of all monomials of degree i occurring in P, where $0 \leq i \leq d$.

It is easily verified that $A[X_1, \ldots, X_n]$ is a commutative ring, with multiplicative identity $1X_1^0 \cdots X_n^0 = 1$. It is also easily verified that $A[X]$ is a module. When A is a field, $A[X]$ is a vector space.

Even when A is just a ring, the family of polynomials

$$(X_1^{k_1} \cdots X_n^{k_n})_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}}$$

is a basis of $A[X_1, \ldots, X_n]$, since every polynomial $P(X_1, \ldots, X_n)$ can be written in a unique way as

$$P(X_1, \ldots, X_n) = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}.$$

Thus, $A[X_1, \ldots, X_n]$ is a free module.

Remark: The construction of Definition 30.3 can be immediately extended to an arbitrary set I, and not just $I = \{1, \ldots, n\}$. It can also be applied to monoids more general that $\mathbb{N}^{(I)}$.

Proposition 30.2 is generalized as follows.

Proposition 30.3. *Let A, B be two rings and let $h\colon A \to B$ be a ring homomorphism. For any $\beta = (\beta_1, \ldots, \beta_n) \in B^n$, there is a unique ring homomorphism $\varphi\colon A[X_1, \ldots, X_n] \to B$ extending h such that $\varphi(X_i) = \beta_i$, $1 \le i \le n$, as in the following diagram (where we denote by $h + \beta$ the map $h + \beta\colon A \cup \{X_1, \ldots, X_n\} \to B$ such that $(h + \beta)(a) = h(a)$ for all $a \in A$ and $(h + \beta)(X_i) = \beta_i$, $1 \le i \le n$):*

$$\begin{tikzcd} A \cup \{X_1, \ldots, X_n\} \arrow[r, "\iota"] \arrow[dr, "h+\beta"'] & A[X_1, \ldots, X_n] \arrow[d, "\varphi"] \\ & B \end{tikzcd}$$

Proof. Let $\varphi(0) = 0$, and for every nonull polynomial

$$P(X_1, \ldots, X_n) = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n},$$

let

$$\varphi(P(X_1, \ldots, X_n)) = \sum h(a_{(k_1,\ldots,k_n)}) \beta_1^{k_1} \cdots \beta_n^{k_n}.$$

It is easily verified that φ is the unique homomorphism $\varphi\colon A[X_1, \ldots, X_n] \to B$ extending h such that $\varphi(X_i) = \beta_i$. ☐

Taking $A = B$ in Proposition 30.3 and $h\colon A \to A$ the identity, for every $\beta_1, \ldots, \beta_n \in A$, there is a unique homomorphism $\varphi\colon A[X_1, \ldots, X_n] \to A$ such that $\varphi(X_i) = \beta_i$, and for every polynomial $P(X_1, \ldots, X_n)$, we write $\varphi(P(X_1, \ldots, X_n))$ as $P(\beta_1, \ldots, \beta_n)$ and we call $P(\beta_1, \ldots, \beta_n)$ the *value of $P(X_1, \ldots, X_n)$ at $X_1 = \beta_1, \ldots, X_n = \beta_n$*. Thus, we can define a function $P_A\colon A^n \to A$ such that $P_A(\beta_1, \ldots, \beta_n) = P(\beta_1, \ldots, \beta_n)$, for all $\beta_1, \ldots, \beta_n \in A$. This function is called the *polynomial function induced by P*.

More generally, P_B can be defined for any (commutative) ring B such that $A \subseteq B$. As in the case of a single variable, it is possible that $P_A = Q_A$ for distinct polynomials P, Q. We will see shortly that the map $P \mapsto P_A$ is injective when $A = \mathbb{R}$ (in general, any infinite integral domain).

Given any nonnull polynomial $P(X_1, \ldots, X_n) = \sum_{(k_1,\ldots,k_n)\in\mathbb{N}^{(n)}} a_{(k_1,\ldots,k_n)} X_1^{k_1} \cdots X_n^{k_n}$ in $A[X_1, \ldots, X_n]$, where $n \geq 2$, $P(X_1, \ldots, X_n)$ can be uniquely written as

$$P(X_1, \ldots, X_n) = \sum Q_{k_n}(X_1, \ldots, X_{n-1}) X_n^{k_n},$$

where each polynomial $Q_{k_n}(X_1, \ldots, X_{n-1})$ is in $A[X_1, \ldots, X_{n-1}]$. Even if A is a field, $A[X_1, \ldots, X_{n-1}]$ is not a field, which confirms that it is useful (and necessary!) to consider polynomials over rings that are not necessarily fields.

It is not difficult to show that $A[X_1, \ldots, X_n]$ and $A[X_1, \ldots, X_{n-1}][X_n]$ are isomorphic rings. This way, it is often possible to prove properties of polynomials in several variables $X_1, \ldots, X_n$, by induction on the number n of variables. For example, given two nonnull polynomials $P(X_1, \ldots, X_n)$ of total degree p and $Q(X_1, \ldots, X_n)$ of total degree q, since we assumed that A is an integral domain, we can prove that

$$\deg(PQ) = \deg(P) + \deg(Q),$$

and that $A[X_1, \ldots, X_n]$ is an integral domain.

Next, we will consider the division of polynomials (in one variable).

30.3 Euclidean Division of Polynomials

We know that every natural number $n \geq 2$ can be written uniquely as a product of powers of prime numbers and that prime numbers play a very important role in arithmetic. It would be nice if every polynomial could be expressed (uniquely) as a product of "irreducible" factors. This is indeed the case for polynomials over a field. The fact that there is a division algorithm for the natural numbers is essential for obtaining many of the arithmetical properties of the natural numbers. As we shall see next, there is also a division algorithm for polynomials in $A[X]$, when A is a field.

Proposition 30.4. *Let A be a ring, let $f(X), g(X) \in A[X]$ be two polynomials of degree $m = \deg(f)$ and $n = \deg(g)$ with $f(X) \neq 0$, and assume that the leading coefficient a_m of $f(X)$ is invertible. Then, there exist unique polynomials $q(X)$ and $r(X)$ in $A[X]$ such that*

$$g = fq + r \quad \textit{and} \quad \deg(r) < \deg(f) = m.$$

Proof. We first prove the existence of q and r. Let

$$f = a_m X^m + a_{m-1} X^{m-1} + \cdots + a_0,$$

and

$$g = b_n X^n + b_{n-1} X^{n-1} + \cdots + b_0.$$

If $n < m$, then let $q = 0$ and $r = g$. Since $\deg(g) < \deg(f)$ and $r = g$, we have $\deg(r) < \deg(f)$.

If $n \geq m$, we proceed by induction on n. If $n = 0$, then $g = b_0$, $m = 0$, $f = a_0 \neq 0$, and we let $q = a_0^{-1} b_0$ and $r = 0$. Since $\deg(r) = \deg(0) = -\infty$ and $\deg(f) = \deg(a_0) = 0$ because $a_0 \neq 0$, we have $\deg(r) < \deg(f)$.

If $n \geq 1$, since $n \geq m$, note that

$$\begin{aligned} g_1(X) &= g(X) - b_n a_m^{-1} X^{n-m} f(X) \\ &= b_n X^n + b_{n-1} X^{n-1} + \cdots + b_0 - b_n a_m^{-1} X^{n-m} (a_m X^m + a_{m-1} X^{m-1} + \cdots + a_0) \end{aligned}$$

is a polynomial of degree $\deg(g_1) < n$, since the terms $b_n X^n$ and $b_n a_m^{-1} X^{n-m} a_m X^m$ of degree n cancel out. Now, since $\deg(g_1) < n$, by the induction hypothesis, we can find q_1 and r such that

$$g_1 = f q_1 + r \quad \text{and} \quad \deg(r) < \deg(f) = m,$$

and thus,

$$g_1(X) = g(X) - b_n a_m^{-1} X^{n-m} f(X) = f(X) q_1(X) + r(X),$$

from which, letting $q(X) = b_n a_m^{-1} X^{n-m} + q_1(X)$, we get

$$g = fq + r \quad \text{and} \quad \deg(r) < m = \deg(f).$$

We now prove uniqueness. If

$$g = f q_1 + r_1 = f q_2 + r_2,$$

with $\deg(r_1) < \deg(f)$ and $\deg(r_2) < \deg(f)$, we get

$$f(q_1 - q_2) = r_2 - r_1.$$

If $q_2 - q_1 \neq 0$, since the leading coefficient a_m of f is invertible, by Proposition 30.1, we have

$$\deg(r_2 - r_1) = \deg(f(q_1 - q_2)) = \deg(f) + \deg(q_2 - q_1),$$

and so, $\deg(r_2 - r_1) \geq \deg(f)$, which contradicts the fact that $\deg(r_1) < \deg(f)$ and $\deg(r_2) < \deg(f)$. Thus, $q_1 = q_2$, and then also $r_1 = r_2$. $\square$

It should be noted that the proof of Proposition 30.4 actually provides an algorithm for finding the *quotient* q and the *remainder* r of the division of g by f. This algorithm is called the *Euclidean algorithm, or division algorithm*. Note that the division of g by f is always possible when f is a monic polynomial, since 1 is invertible. Also, when A is a field, $a_m \neq 0$ is always invertible, and thus, the division can always be performed. We say that f *divides* g when $r = 0$ in the result of the division $g = fq + r$. We now draw some important consequences of the existence of the Euclidean algorithm.

30.4 Ideals, PID's, and Greatest Common Divisors

First, we introduce the fundamental concept of an ideal.

Definition 30.4. Given a ring A, an *ideal of* A is any nonempty subset $\mathfrak{I}$ of A satisfying the following two properties:

(ID1) If $a, b \in \mathfrak{I}$, then $b - a \in \mathfrak{I}$.

(ID2) If $a \in \mathfrak{I}$, then $ax \in \mathfrak{I}$ for every $x \in A$.

An ideal $\mathfrak{I}$ is a *principal ideal* if there is some $a \in \mathfrak{I}$, called a *generator*, such that

$$\mathfrak{I} = \{ax \mid x \in A\}.$$

The equality $\mathfrak{I} = \{ax \mid x \in A\}$ is also written as $\mathfrak{I} = aA$ or as $\mathfrak{I} = (a)$. The ideal $\mathfrak{I} = (0) = \{0\}$ is called the *null ideal* (or *zero ideal*).

An ideal $\mathfrak{I}$ is a *maximal ideal* if $\mathfrak{I} \neq A$ and for every ideal $\mathfrak{J} \neq A$, if $\mathfrak{I} \subseteq \mathfrak{J}$, then $\mathfrak{J} = \mathfrak{I}$. An ideal $\mathfrak{I}$ is a *prime ideal* if $\mathfrak{I} \neq A$ and if $ab \in \mathfrak{I}$, then $a \in \mathfrak{I}$ or $b \in \mathfrak{I}$, for all $a, b \in A$. Equivalently, $\mathfrak{I}$ is a prime ideal if $\mathfrak{I} \neq A$ and if $a, b \in A - \mathfrak{I}$, then $ab \in A - \mathfrak{I}$, for all $a, b \in A$. In other words, $A - \mathfrak{I}$ is closed under multiplication and $1 \in A - \mathfrak{I}$.

Note that if $\mathfrak{I}$ is an ideal, then $\mathfrak{I} = A$ iff $1 \in \mathfrak{I}$. Since by definition, an ideal $\mathfrak{I}$ is nonempty, there is some $a \in \mathfrak{I}$, and by (ID1) we get $0 = a - a \in \mathfrak{I}$. Then, for every $a \in \mathfrak{I}$, since $0 \in \mathfrak{I}$, by (ID1) we get $-a \in \mathfrak{I}$. Thus, an ideal is an additive subgroup of A. Because of (ID2), an ideal is also a subring.

Observe that if A is a field, then A only has two ideals, namely, the trivial ideal (0) and A itself. Indeed, if $\mathfrak{I} \neq (0)$, because every nonnull element has an inverse, then $1 \in \mathfrak{I}$, and thus, $\mathfrak{I} = A$.

Definition 30.5. Given a ring A, for any two elements $a, b \in A$ we say that b *is a multiple of* a *and that* a *divides* b if $b = ac$ for some $c \in A$; this is usually denoted by $a \mid b$.

Note that the principal ideal (a) is the set of all multiples of a, and that a divides b iff b is a multiple of a iff $b \in (a)$ iff $(b) \subseteq (a)$.

Note that every $a \in A$ divides 0. However, it is customary to say that a is a *zero divisor* iff $ac = 0$ for some $c \neq 0$. With this convention, 0 is a zero divisor unless $A = \{0\}$ (the trivial ring), and A is an integral domain iff 0 is the only zero divisor in A.

Given $a, b \in A$ with $a, b \neq 0$, if $(a) = (b)$ then there exist $c, d \in A$ such that $a = bc$ and $b = ad$. From this, we get $a = adc$ and $b = bcd$, that is, $a(1 - dc) = 0$ and $b(1 - cd) = 0$. If A is an integral domain, we get $dc = 1$ and $cd = 1$, that is, c is invertible with inverse d. Thus, when A is an integral domain, we have $b = ad$, with d invertible. The converse is obvious, if $b = ad$ with d invertible, then $(a) = (b)$.

It is worth recording this fact as the following proposition.

Proposition 30.5. *If A is an integral domain, for any $a, b \in A$ with $a, b \neq 0$, we have $(a) = (b)$ iff there exists some invertible $d \in A$ such that $b = ad$.*

An invertible element $u \in A$ is also called a *unit*.

Given two ideals $\mathfrak{I}$ and $\mathfrak{J}$, their sum

$$\mathfrak{I} + \mathfrak{J} = \{a + b \mid a \in \mathfrak{I},\ b \in \mathfrak{J}\}$$

is clearly an ideal. Given any nonempty subset J of A, the set

$$\{a_1 x_1 + \cdots + a_n x_n \mid x_1, \ldots, x_n \in A,\, a_1, \ldots, a_n \in J,\, n \geq 1\}$$

is easily seen to be an ideal, and in fact, it is the smallest ideal containing J. It is usually denoted by (J).

Ideals play a very important role in the study of rings. They tend to show up everywhere. For example, they arise naturally from homomorphisms.

Proposition 30.6. *Given any ring homomorphism $h\colon A \to B$, the kernel $\mathrm{Ker}\, h = \{a \in A \mid h(a) = 0\}$ of h is an ideal.*

Proof. Given $a, b \in A$, we have $a, b \in \mathrm{Ker}\, h$ iff $h(a) = h(b) = 0$, and since h is a homomorphism, we get

$$h(b - a) = h(b) - h(a) = 0,$$

and

$$h(ax) = h(a)h(x) = 0$$

for all $x \in A$, which shows that $\mathrm{Ker}\, h$ is an ideal. □

There is a sort of converse property. Given a ring A and an ideal $\mathfrak{I} \subseteq A$, we can define the quotient ring $A/\mathfrak{I}$, and there is a surjective homomorphism $\pi\colon A \to A/\mathfrak{I}$ whose kernel is precisely $\mathfrak{I}$.

Proposition 30.7. *Given any ring A and any ideal $\mathfrak{I} \subseteq A$, the equivalence relation $\equiv_{\mathfrak{I}}$ defined by $a \equiv_{\mathfrak{I}} b$ iff $b - a \in \mathfrak{I}$ is a congruence, which means that if $a_1 \equiv_{\mathfrak{I}} b_1$ and $a_2 \equiv_{\mathfrak{I}} b_2$, then*

1. *$a_1 + a_2 \equiv_{\mathfrak{I}} b_1 + b_2$, and*

2. *$a_1 a_2 \equiv_{\mathfrak{I}} b_1 b_2$.*

Then, the set $A/\mathfrak{I}$ of equivalence classes modulo $\mathfrak{I}$ is a ring under the operations

$$\begin{aligned} [a] + [b] &= [a + b] \\ [a][b] &= [ab]. \end{aligned}$$

The map $\pi\colon A \to A/\mathfrak{I}$ such that $\pi(a) = [a]$ is a surjective homomorphism whose kernel is precisely $\mathfrak{I}$.

Proof. Everything is straightforward. For example, if $a_1 \equiv_{\mathfrak{I}} b_1$ and $a_2 \equiv_{\mathfrak{I}} b_2$, then $b_1 - a_1 \in \mathfrak{I}$ and $b_2 - a_2 \in \mathfrak{I}$. Since $\mathfrak{I}$ is an ideal, we get

$$(b_1 - a_1)b_2 = b_1b_2 - a_1b_2 \in \mathfrak{I}$$

and

$$(b_2 - a_2)a_1 = a_1b_2 - a_1a_2 \in \mathfrak{I}.$$

Since $\mathfrak{I}$ is an ideal, and thus, an additive group, we get

$$b_1b_2 - a_1a_2 \in \mathfrak{I},$$

i.e., $a_1a_2 \equiv_{\mathfrak{I}} b_1b_2$. The equality $\operatorname{Ker} \pi = \mathfrak{I}$ holds because $\mathfrak{I}$ is an ideal. $\square$

Example 30.1.

1. In the ring $\mathbb{Z}$, for every $p \in \mathbb{Z}$, the subroup $p\mathbb{Z}$ is an ideal, and $\mathbb{Z}/p\mathbb{Z}$ is a ring, the ring of residues modulo p. This ring is a field iff p is a prime number.

2. The quotient of the polynomial ring $\mathbb{R}[X]$ by a prime ideal $\mathfrak{I}$ is an integral domain.

3. The quotient of the polynomial ring $\mathbb{R}[X]$ by a maximal ideal $\mathfrak{I}$ is a field. For example, if $\mathfrak{I} = (X^2 + 1)$, the principal ideal generated by $X^2 + 1$ (which is indeed a maximal ideal since $X^2 + 1$ has no real roots), then $\mathbb{R}[X]/(X^2 + 1) \cong \mathbb{C}$.

The following proposition yields a characterization of prime ideals and maximal ideals in terms of quotients.

Proposition 30.8. *Given a ring A, for any ideal $\mathfrak{I} \subseteq A$, the following properties hold.*

(1) The ideal $\mathfrak{I}$ is a prime ideal iff $A/\mathfrak{I}$ is an integral domain.

(2) The ideal $\mathfrak{I}$ is a maximal ideal iff $A/\mathfrak{I}$ is a field.

Proof. (1) Assume that $\mathfrak{I}$ is a prime ideal. Since $\mathfrak{I}$ is prime, $\mathfrak{I} \neq A$, and thus, $A/\mathfrak{I}$ is not the trivial ring (0). If $[a][b] = 0$, since $[a][b] = [ab]$, we have $ab \in \mathfrak{I}$, and since $\mathfrak{I}$ is prime, then either $a \in \mathfrak{I}$ or $b \in \mathfrak{I}$, so that either $[a] = 0$ or $[b] = 0$. Thus, $A/\mathfrak{I}$ is an integral domain.

Conversely, assume that $A/\mathfrak{I}$ is an integral domain. Since $A/\mathfrak{I}$ is not the trivial ring, $\mathfrak{I} \neq A$. Assume that $ab \in \mathfrak{I}$. Then, we have

$$\pi(ab) = \pi(a)\pi(b) = 0,$$

which implies that either $\pi(a) = 0$ or $\pi(b) = 0$, since $A/\mathfrak{I}$ is an integral domain (where $\pi\colon A \to A/\mathfrak{I}$ is the quotient map). Thus, either $a \in \mathfrak{I}$ or $b \in \mathfrak{I}$, and $\mathfrak{I}$ is a prime ideal.

(2) Assume that $\mathfrak{I}$ is a maximal ideal. As in (1), $A/\mathfrak{I}$ is not the trivial ring (0). Let $[a] \neq 0$ in $A/\mathfrak{I}$. We need to prove that $[a]$ has a multiplicative inverse. Since $[a] \neq 0$, we have $a \notin \mathfrak{I}$. Let $\mathfrak{I}_a$ be the ideal generated by $\mathfrak{I}$ and a. We have

$$\mathfrak{I} \subseteq \mathfrak{I}_a \quad \text{and} \quad \mathfrak{I} \neq \mathfrak{I}_a,$$

since $a \notin \mathfrak{I}$, and since $\mathfrak{I}$ is maximal, this implies that

$$\mathfrak{I}_a = A.$$

However, we know that

$$\mathfrak{I}_a = \{ax + h \mid x \in A,\, h \in \mathfrak{I}\},$$

and thus, there is some $x \in A$ so that

$$ax + h = 1,$$

which proves that $[a][x] = [1]$, as desired.

Conversely, assume that $A/\mathfrak{I}$ is a field. Again, since $A/\mathfrak{I}$ is not the trivial ring, $\mathfrak{I} \neq A$. Let $\mathfrak{J}$ be any proper ideal such that $\mathfrak{I} \subseteq \mathfrak{J}$, and assume that $\mathfrak{I} \neq \mathfrak{J}$. Thus, there is some $j \in \mathfrak{J} - \mathfrak{I}$, and since $\operatorname{Ker} \pi = \mathfrak{I}$, we have $\pi(j) \neq 0$. Since $A/\mathfrak{I}$ is a field and π is surjective, there is some $k \in A$ so that $\pi(j)\pi(k) = 1$, which implies that

$$jk - 1 = i$$

for some $i \in \mathfrak{I}$, and since $\mathfrak{I} \subset \mathfrak{J}$ and $\mathfrak{J}$ is an ideal, it follows that $1 = jk - i \in \mathfrak{J}$, showing that $\mathfrak{J} = A$, a contradiction. Therefore, $\mathfrak{I} = \mathfrak{J}$, and $\mathfrak{I}$ is a maximal ideal. □

As a corollary, we obtain the following useful result. It emphasizes the importance of maximal ideals.

Corollary 30.9. *Given any ring A, every maximal ideal $\mathfrak{I}$ in A is a prime ideal.*

Proof. If $\mathfrak{I}$ is a maximal ideal, then, by Proposition 30.8, the quotient ring $A/\mathfrak{I}$ is a field. However, a field is an integral domain, and by Proposition 30.8 (again), $\mathfrak{I}$ is a prime ideal. □

Observe that a ring A is an integral domain iff (0) is a prime ideal. This is an example of a prime ideal which is not a maximal ideal, as immediately seen in $A = \mathbb{Z}$, where (p) is a maximal ideal for every prime number p.

A less obvious example of a prime ideal which is not a maximal ideal is the ideal (X) in the ring of polynomials $\mathbb{Z}[X]$. Indeed, $(X, 2)$ is also a prime ideal, but (X) is properly contained in $(X, 2)$. The ideal (X) is the set of all polynomials of the form $XQ(X)$ for any $Q(X) \in \mathbb{Z}[X]$, in other words the set of all polynomials in $\mathbb{Z}[X]$ with constant term equal to 0, and the ideal $(X, 2)$ is the set of all polynomials of the form

$$XQ_1(X) + 2Q_2(X), \quad Q_1(X), Q_2(X) \in \mathbb{Z}[X],$$

which is just the set of all polynomials in $\mathbb{Z}[X]$ whose constant term is of the form $2c$ for some $c \in \mathbb{Z}$. The ideal (X) is indeed properly contained in the ideal $(X, 2)$. If $P(X)Q(X) \in (X, 2)$, let a be the constant term in $P(X)$ and let b be the constant term in $Q(X)$. Since $P(X)Q(X) \in (X, 2)$, we must have $ab = 2c$ for some $c \in \mathbb{Z}$, and since 2 is prime, either a is divisible by 2 or b is divisible by 2. It follows that either $P(X) \in (X, 2)$ or $Q(X) \in (X, 2)$, which shows that $(X, 2)$ is a prime ideal.

Definition 30.6. An integral domain in which every ideal is a principal ideal is called a *principal ring or principal ideal domain*, for short, a *PID*.

The ring $\mathbb{Z}$ is a PID. This is a consequence of the existence of a (Euclidean) division algorithm. As we shall see next, when K is a field, the ring $K[X]$ is also a principal ring.

However, when $n \geq 2$, the ring $K[X_1, \ldots, X_n]$ is not principal. For example, in the ring $K[X, Y]$, the ideal (X, Y) generated by X and Y is not principal. First, since (X, Y) is the set of all polynomials of the form $Xq_1 + Yq_2$, where $q_1, q_2 \in K[X, Y]$, except when $Xq_1 + Yq_2 = 0$, we have $\deg(Xq_1 + Yq_2) \geq 1$. Thus, $1 \notin (X, Y)$. Now if there was some $p \in K[X, Y]$ such that $(X, Y) = (p)$, since $1 \notin (X, Y)$, we must have $\deg(p) \geq 1$. But we would also have $X = pq_1$ and $Y = pq_2$, for some $q_1, q_2 \in K[X, Y]$. Since $\deg(X) = \deg(Y) = 1$, this is impossible.

Even though $K[X, Y]$ is not a principal ring, a suitable version of unique factorization in terms of irreducible factors holds. The ring $K[X, Y]$ (and more generally $K[X_1, \ldots, X_n]$) is what is called a *unique factorization domain*, for short, UFD, or a *factorial ring*.

From this point until Definition 30.11, we consider polynomials in one variable over a field K.

Remark: Although we already proved part (1) of Proposition 30.10 in a more general situation above, we reprove it in the special case of polynomials. This may offend the purists, but most readers will probably not mind.

Proposition 30.10. *Let K be a field. The following properties hold:*

(1) For any two nonzero polynomials $f, g \in K[X]$, $(f) = (g)$ iff there is some $\lambda \neq 0$ in K such that $g = \lambda f$.

(2) For every nonnull ideal $\mathfrak{I}$ in $K[X]$, there is a unique monic polynomial $f \in K[X]$ such that $\mathfrak{I} = (f)$.

Proof. (1) If $(f) = (g)$, there are some nonzero polynomials $q_1, q_2 \in K[X]$ such that $g = fq_1$ and $f = gq_2$. Thus, we have $f = fq_1q_2$, which implies $f(1 - q_1q_2) = 0$. Since K is a field, by Proposition 30.1, $K[X]$ has no zero divisor, and since we assumed $f \neq 0$, we must have $q_1q_2 = 1$. However, if either q_1 or q_2 is not a constant, by Proposition 30.1 again, $\deg(q_1q_2) = \deg(q_1) + \deg(q_2) \geq 1$, contradicting $q_1q_2 = 1$, since $\deg(1) = 0$. Thus, both $q_1, q_2 \in K - \{0\}$, and (1) holds with $\lambda = q_1$. In the other direction, it is obvious that $g = \lambda f$ implies that $(f) = (g)$.

(2) Since we are assuming that $\mathfrak{I}$ is not the null ideal, there is some polynomial of smallest degree in $\mathfrak{I}$, and since K is a field, by suitable multiplication by a scalar, we can make sure that this polynomial is monic. Thus, let f be a monic polynomial of smallest degree in $\mathfrak{I}$. By (ID2), it is clear that $(f) \subseteq \mathfrak{I}$. Now, let $g \in \mathfrak{I}$. Using the Euclidean algorithm, there exist unique $q, r \in K[X]$ such that

$$g = qf + r \quad \text{and} \quad \deg(r) < \deg(f).$$

If $r \neq 0$, there is some $\lambda \neq 0$ in K such that λr is a monic polynomial, and since $\lambda r = \lambda g - \lambda q f$, with $f, g \in \mathfrak{I}$, by (ID1) and (ID2), we have $\lambda r \in \mathfrak{I}$, where $\deg(\lambda r) < \deg(f)$ and λr is a monic polynomial, contradicting the minimality of the degree of f. Thus, $r = 0$, and $g \in (f)$. The uniqueness of the monic polynomial f follows from (1). $\square$

Proposition 30.10 shows that $K[X]$ is a principal ring when K is a field.

We now investigate the existence of a greatest common divisor (gcd) for two nonzero polynomials. Given any two nonzero polynomials $f, g \in K[X]$, recall that f divides g if $g = fq$ for some $q \in K[X]$.

Definition 30.7. Given any two nonzero polynomials $f, g \in K[X]$, a polynomial $d \in K[X]$ is a *greatest common divisor of f and g* (for short, a *gcd of f and g*) if d divides f and g and whenever $h \in K[X]$ divides f and g, then h divides d. We say that f and g are *relatively prime* if 1 is a gcd of f and g.

Note that f and g are relatively prime iff all of their gcd's are constants (scalars in K), or equivalently, if f, g have no divisor q of degree $\deg(q) \geq 1$.

In particular, note that f and g are relatively prime when f is a nonzero constant polynomial (a scalar $\lambda \neq 0$ in K) and g is any nonzero polynomial.

We can characterize gcd's of polynomials as follows.

Proposition 30.11. *Let K be a field and let $f, g \in K[X]$ be any two nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of f and g.

(2) The polynomial d divides f and g and there exist $u, v \in K[X]$ such that

$$d = uf + vg.$$

(3) The ideals $(f), (g)$, and (d) satisfy the equation

$$(d) = (f) + (g).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

Proof. Given any two nonzero polynomials $u, v \in K[X]$, observe that u divides v iff $(v) \subseteq (u)$. Now, (2) can be restated as $(f) \subseteq (d)$, $(g) \subseteq (d)$, and $d \in (f) + (g)$, which is equivalent to $(d) = (f) + (g)$, namely (3).

If (2) holds, since $d = uf + vg$, whenever $h \in K[X]$ divides f and g, then h divides d, and d is a gcd of f and g.

Assume that d is a gcd of f and g. Then, since d divides f and d divides g, we have $(f) \subseteq (d)$ and $(g) \subseteq (d)$, and thus $(f) + (g) \subseteq (d)$, and $(f) + (g)$ is nonempty since f and g are nonzero. By Proposition 30.10, there exists a monic polynomial $d_1 \in K[X]$ such that $(d_1) = (f) + (g)$. Then, d_1 divides both f and g, and since d is a gcd of f and g, then d_1 divides d, which shows that $(d) \subseteq (d_1) = (f) + (g)$. Consequently, $(f) + (g) = (d)$, and (3) holds.

Since $(d) = (f) + (g)$ and f and g are nonzero, the last part of the proposition is obvious. ∎

As a consequence of Proposition 30.11, two nonzero polynomials $f, g \in K[X]$ are relatively prime iff there exist $u, v \in K[X]$ such that

$$uf + vg = 1.$$

The identity

$$d = uf + vg$$

of part (2) of Proposition 30.11 is often called the *Bezout identity*.

We derive more useful consequences of Proposition 30.11.

Proposition 30.12. *Let K be a field and let $f, g \in K[X]$ be any two nonzero polynomials. For every gcd $d \in K[X]$ of f and g, the following properties hold:*

(1) For every nonzero polynomial $q \in K[X]$, the polynomial dq is a gcd of fq and gq.

(2) For every nonzero polynomial $q \in K[X]$, if q divides f and g, then d/q is a gcd of f/q and g/q.

Proof. (1) By Proposition 30.11 (2), d divides f and g, and there exist $u, v \in K[X]$, such that

$$d = uf + vg.$$

Then, dq divides fq and gq, and

$$dq = ufq + vgq.$$

By Proposition 30.11 (2), dq is a gcd of fq and gq. The proof of (2) is similar. ∎

The following proposition is used often.

Proposition 30.13. *(Euclid's proposition) Let K be a field and let $f, g, h \in K[X]$ be any nonzero polynomials. If f divides gh and f is relatively prime to g, then f divides h.*

Proof. From Proposition 30.11, f and g are relatively prime iff there exist some polynomials $u, v \in K[X]$ such that

$$uf + vg = 1.$$

Then, we have

$$ufh + vgh = h,$$

and since f divides gh, it divides both ufh and vgh, and so, f divides h. $\square$

Proposition 30.14. *Let K be a field and let $f, g_1, \ldots, g_m \in K[X]$ be some nonzero polynomials. If f and g_i are relatively prime for all i, $1 \leq i \leq m$, then f and $g_1 \cdots g_m$ are relatively prime.*

Proof. We proceed by induction on m. The case $m = 1$ is trivial. Let $h = g_2 \cdots g_m$. By the induction hypothesis, f and h are relatively prime. Let d be a gcd of f and $g_1 h$. We claim that d is relatively prime to g_1. Otherwise, d and g_1 would have some nonconstant gcd d_1 which would divide both f and g_1, contradicting the fact that f and g_1 are relatively prime. Now, by Proposition 30.13, since d divides $g_1 h$ and d and g_1 are relatively prime, d divides $h = g_2 \cdots g_m$. But then, d is a divisor of f and h, and since f and h are relatively prime, d must be a constant, and f and $g_1 \cdots g_m$ are relatively prime. $\square$

Definition 30.7 is generalized to any finite number of polynomials as follows.

Definition 30.8. Given any nonzero polynomials $f_1, \ldots, f_n \in K[X]$, where $n \geq 2$, a polynomial $d \in K[X]$ is a *greatest common divisor of* $f_1, \ldots, f_n$ (for short, a *gcd of* $f_1, \ldots, f_n$) if d divides each f_i and whenever $h \in K[X]$ divides each f_i, then h divides d. We say that $f_1, \ldots, f_n$ are *relatively prime* if 1 is a gcd of $f_1, \ldots, f_n$.

It is easily shown that Proposition 30.11 can be generalized to any finite number of polynomials, and similarly for its relevant corollaries. The details are left as an exercise.

Proposition 30.15. *Let K be a field and let $f_1, \ldots, f_n \in K[X]$ be any $n \geq 2$ nonzero polynomials. For every polynomial $d \in K[X]$, the following properties are equivalent:*

(1) The polynomial d is a gcd of $f_1, \ldots, f_n$.

(2) The polynomial d divides each f_i and there exist $u_1, \ldots, u_n \in K[X]$ such that

$$d = u_1 f_1 + \cdots + u_n f_n.$$

(3) The ideals (f_i), and (d) satisfy the equation

$$(d) = (f_1) + \cdots + (f_n).$$

In addition, $d \neq 0$, and d is unique up to multiplication by a nonzero scalar in K.

As a consequence of Proposition 30.15, some polynomials $f_1, \ldots, f_n \in K[X]$ are relatively prime iff there exist $u_1, \ldots, u_n \in K[X]$ such that

$$u_1 f_1 + \cdots + u_n f_n = 1.$$

The identity

$$u_1 f_1 + \cdots + u_n f_n = 1$$

of part (2) of Proposition 30.15 is also called the *Bezout identity*.

We now consider the factorization of polynomials of a single variable into irreducible factors.

30.5 Factorization and Irreducible Factors in $K[X]$

Definition 30.9. Given a field K, a polynomial $p \in K[X]$ is *irreducible or indecomposable or prime* if $\deg(p) \geq 1$ and if p is not divisible by any polynomial $q \in K[X]$ such that $1 \leq \deg(q) < \deg(p)$. Equivalently, p is irreducible if $\deg(p) \geq 1$ and if $p = q_1 q_2$, then either $q_1 \in K$ or $q_2 \in K$ (and of course, $q_1 \neq 0$, $q_2 \neq 0$).

Example 30.2. Every polynomial $aX + b$ of degree 1 is irreducible. Over the field $\mathbb{R}$, the polynomial $X^2 + 1$ is irreducible (why?), but $X^3 + 1$ is not irreducible, since

$$X^3 + 1 = (X + 1)(X^2 - X + 1).$$

The polynomial $X^2 - X + 1$ is irreducible over $\mathbb{R}$ (why?). It would seem that $X^4 + 1$ is irreducible over $\mathbb{R}$, but in fact,

$$X^4 + 1 = (X^2 - \sqrt{2}X + 1)(X^2 + \sqrt{2}X + 1).$$

However, in view of the above factorization, $X^4 + 1$ is irreducible over $\mathbb{Q}$.

It can be shown that the irreducible polynomials over $\mathbb{R}$ are the polynomials of degree 1, or the polynomials of degree 2 of the form $aX^2 + bX + c$, for which $b^2 - 4ac < 0$ (i.e., those having no real roots). This is not easy to prove! Over the complex numbers $\mathbb{C}$, the only irreducible polynomials are those of degree 1. This is a version of a fact often referred to as the "Fundamental theorem of Algebra", or, as the French sometimes say, as "d'Alembert's theorem"!

We already observed that for any two nonzero polynomials $f, g \in K[X]$, f divides g iff $(g) \subseteq (f)$. In view of the definition of a maximal ideal given in Definition 30.4, we now prove that a polynomial $p \in K[X]$ is irreducible iff (p) is a maximal ideal in $K[X]$.

Proposition 30.16. *A polynomial $p \in K[X]$ is irreducible iff (p) is a maximal ideal in $K[X]$.*

Proof. Since $K[X]$ is an integral domain, for all nonzero polynomials $p, q \in K[X]$, $\deg(pq) = \deg(p) + \deg(q)$, and thus, $(p) \neq K[X]$ iff $\deg(p) \geq 1$. Assume that $p \in K[X]$ is irreducible. Since every ideal in $K[X]$ is a principal ideal, every ideal in $K[X]$ is of the form (q), for some $q \in K[X]$. If $(p) \subseteq (q)$, with $\deg(q) \geq 1$, then q divides p, and since $p \in K[X]$ is irreducible, this implies that $p = \lambda q$ for some $\lambda \neq 0$ in K, and so, $(p) = (q)$. Thus, (p) is a maximal ideal. Conversely, assume that (p) is a maximal ideal. Then, as we showed above, $\deg(p) \geq 1$, and if q divides p, with $\deg(q) \geq 1$, then $(p) \subseteq (q)$, and since (p) is a maximal ideal, this implies that $(p) = (q)$, which means that $p = \lambda q$ for some $\lambda \neq 0$ in K, and so, p is irreducible. ☐

Let $p \in K[X]$ be irreducible. Then, for every nonzero polynomial $g \in K[X]$, either p and g are relatively prime, or p divides g. Indeed, if d is any gcd of p and g, if d is a constant, then p and g are relatively prime, and if not, because p is irreducible, we have $d = \lambda p$ for some $\lambda \neq 0$ in K, and thus, p divides g. As a consequence, if $p, q \in K[X]$ are both irreducible, then either p and q are relatively prime, or $p = \lambda q$ for some $\lambda \neq 0$ in K. In particular, if $p, q \in K[X]$ are both irreducible monic polynomials and $p \neq q$, then p and q are relatively prime.

We now prove the (unique) factorization of polynomials into irreducible factors.

Theorem 30.17. *Given any field K, for every nonzero polynomial*

$$f = a_d X^d + a_{d-1} X^{d-1} + \cdots + a_0$$

of degree $d = \deg(f) \geq 1$ in $K[X]$, there exists a unique set $\{\langle p_1, k_1 \rangle, \ldots, \langle p_m, k_m \rangle\}$ such that

$$f = a_d p_1^{k_1} \cdots p_m^{k_m},$$

where the $p_i \in K[X]$ are distinct irreducible monic polynomials, the k_i are (not necessarily distinct) integers, and $m \geq 1$, $k_i \geq 1$.

Proof. First, we prove the existence of such a factorization by induction on $d = \deg(f)$. Clearly, it is enough to prove the result for monic polynomials f of degree $d = \deg(f) \geq 1$. If $d = 1$, then $f = X + a_0$, which is an irreducible monic polynomial.

Assume $d \geq 2$, and assume the induction hypothesis for all monic polynomials of degree $< d$. Consider the set S of all monic polynomials g such that $\deg(g) \geq 1$ and g divides f. Since $f \in S$, the set S is nonempty, and thus, S contains some monic polynomial p_1 of minimal degree. Since $\deg(p_1) \geq 1$, the monic polynomial p_1 must be irreducible. Otherwise we would have $p_1 = g_1 g_2$, for some monic polynomials g_1, g_2 such that $\deg(p_1) > \deg(g_1) \geq 1$ and $\deg(p_1) > \deg(g_2) \geq 1$, and since p_1 divide f, then g_1 would divide f, contradicting the minimality of the degree of p_1. Thus, we have $f = p_1 q$, for some irreducible monic polynomial p_1, with q also monic. Since $\deg(p_1) \geq 1$, we have $\deg(q) < \deg(f)$, and we can apply the induction hypothesis to q. Thus, we obtain a factorization of the desired form.

We now prove uniqueness. Assume that

$$f = a_d p_1^{k_1} \cdots p_m^{k_m},$$

and

$$f = a_d q_1^{h_1} \cdots q_n^{h_n}.$$

Thus, we have

$$a_d p_1^{k_1} \cdots p_m^{k_m} = a_d q_1^{h_1} \cdots q_n^{h_n}.$$

We prove that $m = n$, $p_i = q_i$ and $h_i = k_i$, for all i, with $1 \leq i \leq n$.

The proof proceeds by induction on $h_1 + \cdots + h_n$.

If $h_1 + \cdots + h_n = 1$, then $n = 1$ and $h_1 = 1$. Then, since $K[X]$ is an integral domain, we have

$$p_1^{k_1} \cdots p_m^{k_m} = q_1,$$

and since q_1 and the p_i are irreducible monic, we must have $m = 1$ and $p_1 = q_1$.

If $h_1 + \cdots + h_n \geq 2$, since $K[X]$ is an integral domain and since $h_1 \geq 1$, we have

$$p_1^{k_1} \cdots p_m^{k_m} = q_1 q,$$

with

$$q = q_1^{h_1 - 1} \cdots q_n^{h_n},$$

where $(h_1 - 1) + \cdots + h_n \geq 1$ (and $q_1^{h_1-1} = 1$ if $h_1 = 1$). Now, if q_1 is not equal to any of the p_i, by a previous remark, q_1 and p_i are relatively prime, and by Proposition 30.14, q_1 and $p_1^{k_1} \cdots p_m^{k_m}$ are relatively prime. But this contradicts the fact that q_1 divides $p_1^{k_1} \cdots p_m^{k_m}$. Thus, q_1 is equal to one of the p_i. Without loss of generality, we can assume that $q_1 = p_1$. Then, since $K[X]$ is an integral domain, we have

$$p_1^{k_1 - 1} \cdots p_m^{k_m} = q_1^{h_1 - 1} \cdots q_n^{h_n},$$

where $p_1^{k_1-1} = 1$ if $k_1 = 1$, and $q_1^{h_1-1} = 1$ if $h_1 = 1$. Now, $(h_1 - 1) + \cdots + h_n < h_1 + \cdots + h_n$, and we can apply the induction hypothesis to conclude that $m = n$, $p_i = q_i$ and $h_i = k_i$, with $1 \leq i \leq n$. $\square$

The above considerations about unique factorization into irreducible factors can be extended almost without changes to more general rings known as *Euclidean domains*. In such rings, some abstract version of the division theorem is assumed to hold.

Definition 30.10. A *Euclidean domain (or Euclidean ring)* is an integral domain A such that there exists a function $\varphi\colon A \to \mathbb{N}$ with the following property: For all $a, b \in A$ with $b \neq 0$, there are some $q, r \in A$ such that

$$a = bq + r \quad \text{and} \quad \varphi(r) < \varphi(b).$$

Note that the pair (q, r) is not necessarily unique.

Actually, unique factorization holds in principal ideal domains (PID's), see Theorem 32.12. As shown below, every Euclidean domain is a PID, and thus, unique factorization holds for Euclidean domains.

Proposition 30.18. *Every Euclidean domain A is a PID.*

Proof. Let $\mathfrak{I}$ be a nonnull ideal in A. Then, the set

$$\{\varphi(a) \mid a \in \mathfrak{I}\}$$

is nonempty, and thus, has a smallest element m. Let b be any (nonnull) element of $\mathfrak{I}$ such that $m = \varphi(b)$. We claim that $\mathfrak{I} = (b)$. Given any $a \in \mathfrak{I}$, we can write

$$a = bq + r$$

for some $q, r \in A$, with $\varphi(r) < \varphi(b)$. Since $b \in \mathfrak{I}$ and $\mathfrak{I}$ is an ideal, we also have $bq \in \mathfrak{I}$, and since $a, bq \in \mathfrak{I}$ and $\mathfrak{I}$ is an ideal, then $r \in \mathfrak{I}$ with $\varphi(r) < \varphi(b) = m$, contradicting the minimality of m. Thus, $r = 0$ and $a \in (b)$. But then,

$$\mathfrak{I} \subseteq (b),$$

and since $b \in \mathfrak{I}$, we get

$$\mathfrak{I} = (b),$$

and A is a PID. $\square$

As a corollary of Proposition 30.18, the ring $\mathbb{Z}$ is a Euclidean domain (using the function $\varphi(a) = |a|$) and thus, a PID. If K is a field, the function φ on $K[X]$ defined such that

$$\varphi(f) = \begin{cases} 0 & \text{if } f = 0, \\ \deg(f) + 1 & \text{if } f \neq 0, \end{cases}$$

shows that $K[X]$ is a Euclidean domain.

Example 30.3. A more interesting example of a Euclidean domain is the ring $\mathbb{Z}[i]$ of *Gaussian integers*, i.e., the subring of $\mathbb{C}$ consisting of all complex numbers of the form $a + ib$, where $a, b \in \mathbb{Z}$. Using the function φ defined such that

$$\varphi(a + ib) = a^2 + b^2,$$

we leave it as an interesting exercise to prove that $\mathbb{Z}[i]$ is a Euclidean domain.

Not every PID is a Euclidean ring.

Remark: Given any integer $d \in \mathbb{Z}$ such that $d \neq 0, 1$ and d does not have any square factor greater than one, the *quadratic field* $\mathbb{Q}(\sqrt{d})$ is the field consisting of all complex numbers of the form $a + ib\sqrt{-d}$ if $d < 0$, and of all the real numbers of the form $a + b\sqrt{d}$ if $d > 0$, with $a, b \in \mathbb{Q}$. The subring of $\mathbb{Q}(\sqrt{d})$ consisting of all elements as above for which $a, b \in \mathbb{Z}$ is denoted by $\mathbb{Z}[\sqrt{d}]$. We define the *ring of integers* of the field $\mathbb{Q}(\sqrt{d})$ as the subring of $\mathbb{Q}(\sqrt{d})$ consisting of the following elements:

(1) If $d \equiv 2 \pmod 4$ or $d \equiv 3 \pmod 4$, then all elements of the form $a + ib\sqrt{-d}$ (if $d < 0$) or all elements of the form $a + b\sqrt{d}$ (if $d > 0$), with $a, b \in \mathbb{Z}$;

(2) If $d \equiv 1 \pmod 4$, then all elements of the form $(a + ib\sqrt{-d})/2$ (if $d < 0$) or all elements of the form $(a + b\sqrt{d})/2$ (if $d > 0$), with $a, b \in \mathbb{Z}$ and with a, b either both even or both odd.

Observe that when $d \equiv 2 \pmod 4$ or $d \equiv 3 \pmod 4$, the ring of integers of $\mathbb{Q}(\sqrt{d})$ is equal to $\mathbb{Z}[\sqrt{d}]$.

It can be shown that the rings of integers of the fields $\mathbb{Q}(\sqrt{-d})$ where $d = 19$, 43, 67, 163 are PID's, but not Euclidean rings. The proof is hard and long. First, it can be shown that these rings are UFD's (refer to Definition 32.2), see Stark [164] (Chapter 8, Theorems 8.21 and 8.22). Then, we use the fact that the ring of integers of the field $\mathbb{Q}(\sqrt{d})$ (with $d \neq 0, 1$ any square-free integers) is a certain kind of integral domain called a Dedekind ring; see Atiyah-MacDonald [8] (Chapter 9, Theorem 9.5) or Samuel [143] (Chapter III, Section 3.4). Finally, we use the fact that if a Dedekind ring is a UFD then it is a PID, which follows from Proposition 32.13.

Actually, the rings of integers of $\mathbb{Q}(\sqrt{d})$ that are Euclidean domains are completely determined but the proof is quite difficult. It turns out that there are twenty one such rings corresponding to the integers: $-11, -7, -3, -2, -1,\ 2, 3, 5, 6, 7, 11,\ 13, 17, 19, 21,\ 29, 33, 37, 41, 57$ and 73, see Stark [164] (Chapter 8). For more on quadratic fields and their rings of integers, see Stark [164] (Chapter 8) or Niven, Zuckerman and Montgomery [132] (Chapter 9).

It is possible to characterize a larger class of rings (in terms of ideals), *factorial rings (or unique factorization domains)*, for which unique factorization holds (see Section 32.1). We now consider zeros (or roots) of polynomials.

30.6 Roots of Polynomials

We go back to the general case of an arbitrary ring for a little while.

Definition 30.11. Given a ring A and any polynomial $f \in A[X]$, we say that some $\alpha \in A$ is *a zero of f, or a root of f*, if $f(\alpha) = 0$. Similarly, given a polynomial $f \in A[X_1, \ldots, X_n]$, we say that $(\alpha_1, \ldots, \alpha_n) \in A^n$ is a *a zero of f, or a root of f*, if $f(\alpha_1, \ldots, \alpha_n) = 0$.

When $f \in A[X]$ is the null polynomial, every $\alpha \in A$ is trivially a zero of f. This case being trivial, we usually assume that we are considering zeros of nonnull polynomials.

Example 30.4. Considering the polynomial $f(X) = X^2 - 1$, both $+1$ and -1 are zeros of $f(X)$. Over the field of reals, the polynomial $g(X) = X^2 + 1$ has no zeros. Over the field $\mathbb{C}$ of complex numbers, $g(X) = X^2 + 1$ has two roots i and $-i$, the square roots of -1, which are "imaginary numbers."

We have the following basic proposition showing the relationship between polynomial division and roots.

Proposition 30.19. *Let $f \in A[X]$ be any polynomial and $\alpha \in A$ any element of A. If the result of dividing f by $X - \alpha$ is $f = (X - \alpha)q + r$, then $r = 0$ iff $f(\alpha) = 0$, i.e., α is a root of f iff $r = 0$.*

Proof. We have $f = (X - \alpha)q + r$, with $\deg(r) < 1 = \deg(X - \alpha)$. Thus, r is a constant in K, and since $f(\alpha) = (\alpha - \alpha)q(\alpha) + r$, we get $f(\alpha) = r$, and the proposition is trivial. $\square$

We now consider the issue of multiplicity of a root.

Proposition 30.20. *Let $f \in A[X]$ be any nonnull polynomial and $h \geq 0$ any integer. The following conditions are equivalent.*

(1) f is divisible by $(X - \alpha)^h$ but not by $(X - \alpha)^{h+1}$.

(2) There is some $g \in A[X]$ such that $f = (X - \alpha)^h g$ and $g(\alpha) \neq 0$.

Proof. Assume (1). Then, we have $f = (X - \alpha)^h g$ for some $g \in A[X]$. If we had $g(\alpha) = 0$, by Proposition 30.19, g would be divisible by $(X - \alpha)$, and then f would be divisible by $(X - \alpha)^{h+1}$, contradicting (1).

Assume (2), that is, $f = (X - \alpha)^h g$ and $g(\alpha) \neq 0$. If f is divisible by $(X - \alpha)^{h+1}$, then we have $f = (X - \alpha)^{h+1} g_1$, for some $g_1 \in A[X]$. Then, we have

$$(X - \alpha)^h g = (X - \alpha)^{h+1} g_1,$$

and thus

$$(X - \alpha)^h (g - (X - \alpha)g_1) = 0,$$

and since the leading coefficient of $(X - \alpha)^h$ is 1 (show this by induction), by Proposition 30.1, $(X - \alpha)^h$ is not a zero divisor, and we get $g - (X - \alpha)g_1 = 0$, i.e., $g = (X - \alpha)g_1$, and so $g(\alpha) = 0$, contrary to the hypothesis. $\square$

As a consequence of Proposition 30.20, for every nonnull polynomial $f \in A[X]$ and every $\alpha \in A$, there is a unique integer $h \geq 0$ such that f is divisible by $(X - \alpha)^h$ but not by $(X - \alpha)^{h+1}$. Indeed, since f is divisible by $(X - \alpha)^h$, we have $h \leq \deg(f)$. When $h = 0$, α is not a root of f, i.e., $f(\alpha) \neq 0$. The interesting case is when α is a root of f.

Definition 30.12. Given a ring A and any nonnull polynomial $f \in A[X]$, given any $\alpha \in A$, the unique $h \geq 0$ such that f is divisible by $(X-\alpha)^h$ but not by $(X-\alpha)^{h+1}$ is called the *order, or multiplicity, of* α. We have $h = 0$ iff α is not a root of f, and when α is a root of f, if $h = 1$, we call α a *simple root*, if $h = 2$, a *double root*, and generally, a root of multiplicity $h \geq 2$ is called a *multiple root*.

Observe that Proposition 30.20 (2) implies that if $A \subseteq B$, where A and B are rings, for every nonnull polynomial $f \in A[X]$, if $\alpha \in A$ is a root of f, then the multiplicity of α with respect to $f \in A[X]$ and the multiplicity of α with respect to f considered as a polynomial in $B[X]$, is the same.

We now show that if the ring A is an integral domain, the number of roots of a nonzero polynomial is at most its degree.

Proposition 30.21. *Let $f, g \in A[X]$ be nonnull polynomials, let $\alpha \in A$, and let $h \geq 0$ and $k \geq 0$ be the multiplicities of α with respect to f and g. The following properties hold.*

(1) If l is the multiplicity of α with respect to $(f+g)$, then $l \geq \min(h,k)$. If $h \neq k$, then $l = \min(h,k)$.

(2) If m is the multiplicity of α with respect to fg, then $m \geq h+k$. If A is an integral domain, then $m = h+k$.

Proof. (1) We have $f(X) = (X-\alpha)^h f_1(X)$, $g(X) = (X-\alpha)^k g_1(X)$, with $f_1(\alpha) \neq 0$ and $g_1(\alpha) \neq 0$. Clearly, $l \geq \min(h,k)$. If $h \neq k$, assume $h < k$. Then, we have

$$f(X) + g(X) = (X-\alpha)^h f_1(X) + (X-\alpha)^k g_1(X) = (X-\alpha)^h (f_1(X) + (X-\alpha)^{k-h} g_1(X)),$$

and since $(f_1(X) + (X-\alpha)^{k-h} g_1(X))(\alpha) = f_1(\alpha) \neq 0$, we have $l = h = \min(h,k)$.

(2) We have

$$f(X)g(X) = (X-\alpha)^{h+k} f_1(X) g_1(X),$$

with $f_1(\alpha) \neq 0$ and $g_1(\alpha) \neq 0$. Clearly, $m \geq h+k$. If A is an integral domain, then $f_1(\alpha)g_1(\alpha) \neq 0$, and so $m = h+k$. ☐

Proposition 30.22. *Let A be an integral domain. Let f be any nonnull polynomial $f \in A[X]$ and let $\alpha_1, \ldots, \alpha_m \in A$ be $m \geq 1$ distinct roots of f of respective multiplicities $k_1, \ldots, k_m$. Then, we have*

$$f(X) = (X-\alpha_1)^{k_1} \cdots (X-\alpha_m)^{k_m} g(X),$$

where $g \in A[X]$ and $g(\alpha_i) \neq 0$ for all i, $1 \leq i \leq m$.

Proof. We proceed by induction on m. The case $m = 1$ is obvious in view of Definition 30.12 (which itself, is justified by Proposition 30.20). If $m \geq 2$, by the induction hypothesis, we have

$$f(X) = (X-\alpha_1)^{k_1} \cdots (X-\alpha_{m-1})^{k_{m-1}} g_1(X),$$

where $g_1 \in A[X]$ and $g_1(\alpha_i) \neq 0$, for $1 \leq i \leq m-1$. Since A is an integral domain and $\alpha_i \neq \alpha_j$ for $i \neq j$, since α_m is a root of f, we have

$$0 = (\alpha_m - \alpha_1)^{k_1} \cdots (\alpha_m - \alpha_{m-1})^{k_{m-1}} g_1(\alpha_m),$$

which implies that $g_1(\alpha_m) = 0$. Now, by Proposition 30.21 (2), since α_m is not a root of the polynomial $(X - \alpha_1)^{k_1} \cdots (X - \alpha_{m-1})^{k_{m-1}}$ and since A is an integral domain, α_m must be a root of multiplicity k_m of g_1, which means that

$$g_1(X) = (X - \alpha_m)^{k_m} g(X),$$

with $g(\alpha_m) \neq 0$. Since $g_1(\alpha_i) \neq 0$ for $1 \leq i \leq m-1$ and A is an integral domain, we must also have $g(\alpha_i) \neq 0$, for $1 \leq i \leq m-1$. Thus, we have

$$f(X) = (X - \alpha_1)^{k_1} \cdots (X - \alpha_m)^{k_m} g(X),$$

where $g \in A[X]$, and $g(\alpha_i) \neq 0$ for $1 \leq i \leq m$. □

As a consequence of Proposition 30.22, we get the following important result.

Theorem 30.23. *Let A be an integral domain. For every nonnull polynomial $f \in A[X]$, if the degree of f is $n = \deg(f)$ and $k_1, \ldots, k_m$ are the multiplicities of all the distinct roots of f (where $m \geq 0$), then $k_1 + \cdots + k_m \leq n$.*

Proof. Immediate from Proposition 30.22. □

Since fields are integral domains, Theorem 30.23 holds for nonnull polynomials over fields and in particular, for $\mathbb{R}$ and $\mathbb{C}$. An important consequence of Theorem 30.23 is the following.

Proposition 30.24. *Let A be an integral domain. For any two polynomials $f, g \in A[X]$, if $\deg(f) \leq n$, $\deg(g) \leq n$, and if there are $n+1$ distinct elements $\alpha_1, \alpha_2, \ldots, \alpha_{n+1} \in A$ (with $\alpha_i \neq \alpha_j$ for $i \neq j$) such that $f(\alpha_i) = g(\alpha_i)$ for all i, $1 \leq i \leq n+1$, then $f = g$.*

Proof. Assume $f \neq g$, then, $(f-g)$ is nonnull, and since $f(\alpha_i) = g(\alpha_i)$ for all i, $1 \leq i \leq n+1$, the polynomial $(f-g)$ has $n+1$ distinct roots. Thus, $(f-g)$ has $n+1$ distinct roots and is of degree at most n, which contradicts Theorem 30.23. □

Proposition 30.24 is often used to show that polynomials coincide. We will use it to show some interpolation formulae due to Lagrange and Hermite. But first, we characterize the multiplicity of a root of a polynomial. For this, we need the notion of derivative familiar in analysis. Actually, we can simply define this notion algebraically.

First, we need to rule out some undesirable behaviors. Given a field K, as we saw in Example 2.8, we can define a homomorphism $\chi \colon \mathbb{Z} \to K$ given by

$$\chi(n) = n \cdot 1,$$

where 1 is the multiplicative identity of K. Recall that we define $n \cdot a$ by

$$n \cdot a = \underbrace{a + \cdots + a}_{n}$$

if $n \geq 0$ (with $0 \cdot a = 0$) and

$$n \cdot a = -(-n) \cdot a$$

if $n < 0$. We say that the field K is of *characteristic zero* if the homomorphism χ is injective. Then, for any $a \in K$ with $a \neq 0$, we have $n \cdot a \neq 0$ for all $n \neq 0$

The fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are of characteristic zero. In fact, it is easy to see that every field of characteristic zero contains a subfield isomorphic to $\mathbb{Q}$. Thus, finite fields can't be of characteristic zero.

Remark: If a field is not of characteristic zero, it is not hard to show that its characteristic, that is, the smallest $n \geq 2$ such that $n \cdot 1 = 0$, is a prime number p. The characteristic p of K is the generator of the principal ideal $p\mathbb{Z}$, the kernel of the homomorphism $\chi\colon \mathbb{Z} \to K$. Thus, every finite field is of characteristic some prime p. Infinite fields of nonzero characteristic also exist.

Definition 30.13. Let A be a ring. The *derivative f', or $\mathrm{D}f$, or $\mathrm{D}^1 f$, of a polynomial* $f \in A[X]$ is defined inductively as follows:

$$\begin{aligned}
f' &= 0, \quad \text{if } f = 0, \text{ the null polynomial,}\\
f' &= 0, \quad \text{if } f = a,\ a \neq 0,\ a \in A,\\
f' &= na_n X^{n-1} + (n-1)a_{n-1}X^{n-2} + \cdots + 2a_2 X + a_1,\\
&\qquad \text{if } f = a_n X^n + a_{n-1}X^{n-1} + \cdots + a_0, \text{ with } n = \deg(f) \geq 1.
\end{aligned}$$

If $A = K$ is a field of characteristic zero, if $\deg(f) \geq 1$, the leading coefficient na_n of f' is nonzero, and thus, f' is not the null polynomial. Thus, if $A = K$ is a field of characteristic zero, when $n = \deg(f) \geq 1$, we have $\deg(f') = n - 1$.

For rings or for fields of characteristic $p \geq 2$, we could have $f' = 0$, for a polynomial f of degree ≥ 1.

The following standard properties of derivatives are recalled without proof (prove them as an exercise).

Given any two polynomials, $f, g \in A[X]$, we have

$$\begin{aligned}
(f+g)' &= f' + g',\\
(fg)' &= f'g + fg'.
\end{aligned}$$

For example, if $f = (X - \alpha)^k g$ and $k \geq 1$, we have

$$f' = k(X-\alpha)^{k-1}g + (X-\alpha)^k g'.$$

We can now give a criterion for the existence of simple roots. The first proposition holds for any ring.

Proposition 30.25. *Let A be any ring. For every nonnull polynomial $f \in A[X]$, $\alpha \in A$ is a simple root of f iff α is a root of f and α is not a root of f'.*

Proof. Since $\alpha \in A$ is a root of f, we have $f = (X - \alpha)g$ for some $g \in A[X]$. Now, α is a simple root of f iff $g(\alpha) \neq 0$. However, we have $f' = g + (X - \alpha)g'$, and so $f'(\alpha) = g(\alpha)$. Thus, α is a simple root of f iff $f'(\alpha) \neq 0$. $\square$

We can improve the previous proposition as follows.

Proposition 30.26. *Let A be any ring. For every nonnull polynomial $f \in A[X]$, let $\alpha \in A$ be a root of multiplicity $k \geq 1$ of f. Then, α is a root of multiplicity at least $k - 1$ of f'. If A is a field of characteristic zero, then α is a root of multiplicity $k - 1$ of f'.*

Proof. Since $\alpha \in A$ is a root of multiplicity k of f, we have $f = (X - \alpha)^k g$ for some $g \in A[X]$ and $g(\alpha) \neq 0$. Since

$$f' = k(X - \alpha)^{k-1} g + (X - \alpha)^k g' = (X - \alpha)^{k-1}(kg + (X - \alpha)g'),$$

it is clear that the multiplicity of α w.r.t. f' is at least $k-1$. Now, $(kg+(X-\alpha)g')(\alpha) = kg(\alpha)$, and if A is of characteristic zero, since $g(\alpha) \neq 0$, then $kg(\alpha) \neq 0$. Thus, α is a root of multiplicity $k - 1$ of f'. $\square$

As a consequence, we obtain the following test for the existence of a root of multiplicity k for a polynomial f:

Given a field K of characteristic zero, for any nonnull polynomial $f \in K[X]$, any $\alpha \in K$ is a root of multiplicity $k \geq 1$ of f iff α is a root of $f, \mathrm{D}^1 f, \mathrm{D}^2 f, \ldots, \mathrm{D}^{k-1} f$, but not a root of $\mathrm{D}^k f$.

We can now return to polynomial functions and tie up some loose ends. Given a ring A, recall that every polynomial $f \in A[X_1, \ldots, X_n]$ induces a function $f_A \colon A^n \to A$ defined such that $f_A(\alpha_1, \ldots, \alpha_n) = f(\alpha_1, \ldots, \alpha_n)$, for every $(\alpha_1, \ldots, \alpha_n) \in A^n$. We now give a sufficient condition for the mapping $f \mapsto f_A$ to be injective.

Proposition 30.27. *Let A be an integral domain. For every polynomial $f \in A[X_1, \ldots, X_n]$, if $A_1, \ldots, A_n$ are n infinite subsets of A such that $f(\alpha_1, \ldots, \alpha_n) = 0$ for all $(\alpha_1, \ldots, \alpha_n) \in A_1 \times \cdots \times A_n$, then $f = 0$, i.e., f is the null polynomial. As a consequence, if A is an infinite integral domain, then the map $f \mapsto f_A$ is injective.*

Proof. We proceed by induction on n. Assume $n = 1$. If $f \in A[X_1]$ is nonnull, let $m = \deg(f)$ be its degree. Since A_1 is infinite and $f(\alpha_1) = 0$ for all $\alpha_1 \in A_1$, then f has an infinite number of roots. But since f is of degree m, this contradicts Theorem 30.23. Thus, $f = 0$.

If $n \geq 2$, we can view $f \in A[X_1, \ldots, X_n]$ as a polynomial

$$f = g_m X_n^m + g_{m-1} X_n^{m-1} + \cdots + g_0,$$

where the coefficients g_i are polynomials in $A[X_1, \ldots, X_{n-1}]$. Now, for every $(\alpha_1, \ldots, \alpha_{n-1}) \in A_1 \times \cdots \times A_{n-1}$, $f(\alpha_1, \ldots, \alpha_{n-1}, X_n)$ determines a polynomial $h(X_n) \in A[X_n]$, and since A_n is infinite and $h(\alpha_n) = f(\alpha_1, \ldots, \alpha_{n-1}, \alpha_n) = 0$ for all $\alpha_n \in A_n$, by the induction hypothesis, we have $g_i(\alpha_1, \ldots, \alpha_{n-1}) = 0$. Now, since $A_1, \ldots, A_{n-1}$ are infinite, using the induction hypothesis again, we get $g_i = 0$, which shows that f is the null polynomial. The second part of the proposition follows immediately from the first, by letting $A_i = A$. $\square$

When A is an infinite integral domain, in particular an infinite field, since the map $f \mapsto f_A$ is injective, we identify the polynomial f with the polynomial function f_A, and we write f_A simply as f.

The following proposition can be very useful to show polynomial identities.

Proposition 30.28. *Let A be an infinite integral domain and $f, g_1, \ldots, g_m \in A[X_1, \ldots, X_n]$ be polynomials. If the g_i are nonnull polynomials and if*

$$f(\alpha_1, \ldots, \alpha_n) = 0 \textit{ whenever } g_i(\alpha_1, \ldots, \alpha_n) \neq 0 \textit{ for all } i,\ 1 \leq i \leq m,$$

for every $(\alpha_1, \ldots, \alpha_n) \in A^n$, then

$$f = 0,$$

i.e., f is the null polynomial.

Proof. If f is not the null polynomial, since the g_i are nonnull and A is an integral domain, then the product $f g_1 \cdots g_m$ is nonnull. By Proposition 30.27, only the null polynomial maps to the zero function, and thus there must be some $(\alpha_1, \ldots, \alpha_n) \in A^n$, such that

$$f(\alpha_1, \ldots, \alpha_n) g_1(\alpha_1, \ldots, \alpha_n) \cdots g_m(\alpha_1, \ldots, \alpha_n) \neq 0,$$

but this contradicts the hypothesis. $\square$

Proposition 30.28 is often called the *principle of extension of algebraic identities*. Another perhaps more illuminating way of stating this proposition is as follows: For any polynomial $g \in A[X_1, \ldots, X_n]$, let

$$V(g) = \{(\alpha_1, \ldots, \alpha_n) \in A^n \mid g(\alpha_1, \ldots, \alpha_n) = 0\},$$

the set of zeros of g. Note that $V(g_1) \cup \cdots \cup V(g_m) = V(g_1 \cdots g_m)$. Then, Proposition 30.28 can be stated as:

If $f(\alpha_1, \ldots, \alpha_n) = 0$ for every $(\alpha_1, \ldots, \alpha_n) \in A^n - V(g_1 \cdots g_m)$, then $f = 0$.

In other words, if the algebraic identity $f(\alpha_1, \ldots, \alpha_n) = 0$ holds on the complement of $V(g_1) \cup \cdots \cup V(g_m) = V(g_1 \cdots g_m)$, then $f(\alpha_1, \ldots, \alpha_n) = 0$ holds everywhere in A^n. With this second formulation, we understand better the terminology "principle of extension of algebraic identities."

Remark: Letting $U(g) = A - V(g)$, the identity $V(g_1) \cup \cdots \cup V(g_m) = V(g_1 \cdots g_m)$ translates to $U(g_1) \cap \cdots \cap U(g_m) = U(g_1 \cdots g_m)$. This suggests to define a topology on A whose basis of open sets consists of the sets $U(g)$. In this topology (called the Zariski topology), the sets of the form $V(g)$ are closed sets. Also, when $g_1, \ldots, g_m \in A[X_1, \ldots, X_n]$ and $n \geq 2$, understanding the structure of the closed sets of the form $V(g_1) \cap \cdots \cap V(g_m)$ is quite difficult, and it is the object of algebraic geometry (at least, its classical part).

When $f \in A[X_1, \ldots, X_n]$ and $n \geq 2$, one should not apply Proposition 30.27 abusively. For example, let

$$f(X, Y) = X^2 + Y^2 - 1,$$

considered as a polynomial in $\mathbb{R}[X, Y]$. Since $\mathbb{R}$ is an infinite field and since

$$f\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right) = \frac{(1-t^2)^2}{(1+t^2)^2} + \frac{(2t)^2}{(1+t^2)^2} - 1 = 0,$$

for every $t \in \mathbb{R}$, it would be tempting to say that $f = 0$. But what's wrong with the above reasoning is that there are no two infinite subsets R_1, R_2 of $\mathbb{R}$ such that $f(\alpha_1, \alpha_2) = 0$ for all $(\alpha_1, \alpha_2) \in \mathbb{R}^2$. For every $\alpha_1 \in \mathbb{R}$, there are at most two $\alpha_2 \in \mathbb{R}$ such that $f(\alpha_1, \alpha_2) = 0$. What the example shows though, is that a nonnull polynomial $f \in A[X_1, \ldots, X_n]$ where $n \geq 2$ can have an infinite number of zeros. This is in contrast with nonnull polynomials in one variables over an infinite field (which have a number of roots bounded by their degree).

We now look at polynomial interpolation.

30.7 Polynomial Interpolation (Lagrange, Newton, Hermite)

Let K be a field. First, we consider the following interpolation problem: Given a sequence $(\alpha_1, \ldots, \alpha_{m+1})$ of pairwise distinct scalars in K and any sequence $(\beta_1, \ldots, \beta_{m+1})$ of scalars in K, where the β_j are not necessarily distinct, find a polynomial $P(X)$ of degree $\leq m$ such that

$$P(\alpha_1) = \beta_1, \ldots, P(\alpha_{m+1}) = \beta_{m+1}.$$

First, observe that if such a polynomial exists, then it is unique. Indeed, this is a consequence of Proposition 30.24. Thus, we just have to find any polynomial of degree $\leq m$. Consider the following so-called *Lagrange polynomials*:

$$L_i(X) = \frac{(X - \alpha_1) \cdots (X - \alpha_{i-1})(X - \alpha_{i+1}) \cdots (X - \alpha_{m+1})}{(\alpha_i - \alpha_1) \cdots (\alpha_i - \alpha_{i-1})(\alpha_i - \alpha_{i+1}) \cdots (\alpha_i - \alpha_{m+1})}.$$

Note that $L(\alpha_i) = 1$ and that $L(\alpha_j) = 0$, for all $j \neq i$. But then,

$$P(X) = \beta_1 L_1 + \cdots + \beta_{m+1} L_{m+1}$$

is the unique desired polynomial, since clearly, $P(\alpha_i) = \beta_i$. Such a polynomial is called a *Lagrange interpolant*. Also note that the polynomials $(L_1, \ldots, L_{m+1})$ form a basis of the vector space of all polynomials of degree $\leq m$. Indeed, if we had

$$\lambda_1 L_1(X) + \cdots + \lambda_{m+1} L_{m+1}(X) = 0,$$

setting X to α_i, we would get $\lambda_i = 0$. Thus, the L_i are linearly independent, and by the previous argument, they are a set of generators. We we call $(L_1, \ldots, L_{m+1})$ the *Lagrange basis* (of order $m+1$).

It is known from numerical analysis that from a computational point of view, the Lagrange basis is not very good. Newton proposed another solution, the method of divided differences.

Consider the polynomial $P(X)$ of degree $\leq m$, called the *Newton interpolant*,

$$P(X) = \lambda_0 + \lambda_1(X - \alpha_1) + \lambda_2(X - \alpha_1)(X - \alpha_2) + \cdots + \lambda_m(X - \alpha_1)(X - \alpha_2) \cdots (X - \alpha_m).$$

Then, the λ_i can be determined by successively setting X to, $\alpha_1, \alpha_2, \ldots, \alpha_{m+1}$. More precisely, we define inductively the polynomials $Q(X)$ and $Q(\alpha_1, \ldots, \alpha_i, X)$, for $1 \leq i \leq m$, as follows:

$$\begin{aligned}
Q(X) &= P(X) \\
Q_1(\alpha_1, X) &= \frac{Q(X) - Q(\alpha_1)}{X - \alpha_1} \\
Q(\alpha_1, \alpha_2, X) &= \frac{Q(\alpha_1, X) - Q(\alpha_1, \alpha_2)}{X - \alpha_2} \\
&\cdots \\
Q(\alpha_1, \ldots, \alpha_i, X) &= \frac{Q(\alpha_1, \ldots, \alpha_{i-1}, X) - Q(\alpha_1, \ldots, \alpha_{i-1}, \alpha_i)}{X - \alpha_i}, \\
&\cdots \\
Q(\alpha_1, \ldots, \alpha_m, X) &= \frac{Q(\alpha_1, \ldots, \alpha_{m-1}, X) - Q(\alpha_1, \ldots, \alpha_{m-1}, \alpha_m)}{X - \alpha_m}.
\end{aligned}$$

By induction on i, $1 \leq i \leq m-1$, it is easily verified that

$$\begin{aligned}
Q(X) &= P(X), \\
Q(\alpha_1, \ldots, \alpha_i, X) &= \lambda_i + \lambda_{i+1}(X - \alpha_{i+1}) + \cdots + \lambda_m(X - \alpha_{i+1}) \cdots (X - \alpha_m), \\
Q(\alpha_1, \ldots, \alpha_m, X) &= \lambda_m.
\end{aligned}$$

From the above expressions, it is clear that

$$\begin{aligned}
\lambda_0 &= Q(\alpha_1), \\
\lambda_i &= Q(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}), \\
\lambda_m &= Q(\alpha_1, \ldots, \alpha_m, \alpha_{m+1}).
\end{aligned}$$

The expression $Q(\alpha_1, \alpha_2, \ldots, \alpha_{i+1})$ is called the *i-th difference quotient*. Then, we can compute the λ_i in terms of $\beta_1 = P(\alpha_1), \ldots, \beta_{m+1} = P(\alpha_{m+1})$, using the inductive formulae for the $Q(\alpha_1, \ldots, \alpha_i, X)$ given above, initializing the $Q(\alpha_i)$ such that $Q(\alpha_i) = \beta_i$.

The above method is called the method of *divided differences* and it is due to Newton.

An astute observation may be used to optimize the computation. Observe that if $P_i(X)$ is the polynomial of degree $\leq i$ taking the values $\beta_1, \ldots, \beta_{i+1}$ at the points $\alpha_1, \ldots, \alpha_{i+1}$, then the coefficient of X^i in $P_i(X)$ is $Q(\alpha_1, \alpha_2, \ldots, \alpha_{i+1})$, which is the value of λ_i in the Newton interpolant

$$P_i(X) = \lambda_0 + \lambda_1(X - \alpha_1) + \lambda_2(X - \alpha_1)(X - \alpha_2) + \cdots + \lambda_i(X - \alpha_1)(X - \alpha_2) \cdots (X - \alpha_i).$$

As a consequence, $Q(\alpha_1, \alpha_2, \ldots, \alpha_{i+1})$ does not depend on the specific ordering of the α_j and there are better ways of computing it. For example, $Q(\alpha_1, \alpha_2, \ldots, \alpha_{i+1})$ can be computed using

$$Q(\alpha_1, \ldots, \alpha_{i+1}) = \frac{Q(\alpha_2, \ldots, \alpha_{i+1}) - Q(\alpha_1, \ldots, \alpha_i)}{\alpha_{i+1} - \alpha_1}.$$

Then, the computation can be arranged into a triangular array reminiscent of Pascal's triangle, as follows:

Initially, $Q(\alpha_j) = \beta_j$, $1 \leq j \leq m + 1$, and

$$\begin{array}{cccc}
Q(\alpha_1) & & & \\
 & Q(\alpha_1, \alpha_2) & & \\
Q(\alpha_2) & & Q(\alpha_1, \alpha_2, \alpha_3) & \\
 & Q(\alpha_2, \alpha_3) & & \ldots \\
Q(\alpha_3) & & Q(\alpha_2, \alpha_3, \alpha_4) & \\
 & Q(\alpha_3, \alpha_4) & \ldots & \\
Q(\alpha_4) & \ldots & & \\
\ldots & & &
\end{array}$$

In this computation, each successive column is obtained by forming the difference quotients of the preceeding column according to the formula

$$Q(\alpha_k, \ldots, \alpha_{i+k}) = \frac{Q(\alpha_{k+1}, \ldots, \alpha_{i+k}) - Q(\alpha_k, \ldots, \alpha_{i+k-1})}{\alpha_{i+k} - \alpha_k}.$$

The λ_i are the elements of the descending diagonal.

Observe that if we performed the above computation starting with a polynomial $Q(X)$ of degree m, we could extend it by considering new given points α_{m+2}, α_{m+3}, etc. Then, from what we saw above, the $(m + 1)$th column consists of λ_m in the expression of $Q(X)$ as a Newton interpolant and the $(m + 2)$th column consists of zeros. Such divided differences are used in numerical analysis.

Newton's method can be used to compute the value $P(\alpha)$ at some α of the interpolant $P(X)$ taking the values $\beta_1, \ldots, \beta_{m+1}$ for the (distinct) arguments $\alpha_1, \ldots, \alpha_{m+1}$. We also mention that inductive methods for computing $P(\alpha)$ without first computing the coefficients of the Newton interpolant exist, for example, Aitken's method. For this method, the reader is referred to Farin [58].

It has been observed that Lagrange interpolants oscillate quite badly as their degree increases, and thus, this makes them undesirable as a stable method for interpolation. A standard example due to Runge, is the function

$$f(x) = \frac{1}{1 + x^2},$$

in the interval $[-5, +5]$. Assuming a uniform distribution of points on the curve in the interval $[-5, +5]$, as the degree of the Lagrange interpolant increases, the interpolant shows wilder and wilder oscillations around the points $x = -5$ and $x = +5$. This phenomenon becomes quite noticeable beginning for degree 14, and gets worse and worse. For degree 22, things are quite bad! Equivalently, one may consider the function

$$f(x) = \frac{1}{1 + 25x^2},$$

in the interval $[-1, +1]$.

We now consider a more general interpolation problem which will lead to the Hermite polynomials.

We consider the following interpolation problem:

Given a sequence $(\alpha_1, \ldots, \alpha_{m+1})$ of pairwise distinct scalars in K, integers $n_1, \ldots, n_{m+1}$ where $n_j \geq 0$, and $m + 1$ sequences $(\beta_j^0, \ldots, \beta_j^{n_j})$ of scalars in K, letting

$$n = n_1 + \cdots + n_{m+1} + m,$$

find a polynomial P of degree $\leq n$, such that

$$\begin{array}{rcr}
P(\alpha_1) = \beta_1^0, & \ldots & P(\alpha_{m+1}) = \beta_{m+1}^0, \\
\mathrm{D}^1 P(\alpha_1) = \beta_1^1, & \ldots & \mathrm{D}^1 P(\alpha_{m+1}) = \beta_{m+1}^1, \\
 & \ldots & \\
\mathrm{D}^i P(\alpha_1) = \beta_1^i, & \ldots & \mathrm{D}^i P(\alpha_{m+1}) = \beta_{m+1}^i, \\
 & \ldots & \\
\mathrm{D}^{n_1} P(\alpha_1) = \beta_1^{n_1}, & \ldots & \mathrm{D}^{n_{m+1}} P(\alpha_{m+1}) = \beta_{m+1}^{n_{m+1}}.
\end{array}$$

Note that the above equations constitute $n+1$ constraints, and thus, we can expect that there is a unique polynomial of degree $\leq n$ satisfying the above problem. This is indeed the case and such a polynomial is called a *Hermite polynomial*. We call the above problem the *Hermite interpolation problem*.

Proposition 30.29. *The Hermite interpolation problem has a unique solution of degree* $\leq n$, *where* $n = n_1 + \cdots + n_{m+1} + m$.

Proof. First, we prove that the Hermite interpolation problem has at most one solution. Assume that P and Q are two distinct solutions of degree $\leq n$. Then, by Proposition 30.26 and the criterion following it, $P - Q$ has among its roots α_1 of multiplicity at least $n_1+1, \ldots$, α_{m+1} of multiplicity at least $n_{m+1} + 1$. However, by Theorem 30.23, we should have

$$n_1 + 1 + \cdots + n_{m+1} + 1 = n_1 + \cdots + n_{m+1} + m + 1 \leq n,$$

which is a contradiction, since $n = n_1 + \cdots + n_{m+1} + m$. Thus, $P = Q$. We are left with proving the existence of a Hermite interpolant. A quick way to do so is to use Proposition 7.12, which tells us that given a square matrix A over a field K, the following properties hold:

For every column vector B, there is a unique column vector X such that $AX = B$ iff the only solution to $AX = 0$ is the trivial vector $X = 0$ iff $D(A) \neq 0$.

If we let $P = y_0 + y_1 X + \cdots + y_n X^n$, the Hermite interpolation problem yields a linear system of equations in the unknowns $(y_0, \ldots, y_n)$ with some associated $(n+1)\times(n+1)$ matrix A. Now, the system $AY = 0$ has a solution iff P has among its roots α_1 of multiplicity at least $n_1 + 1, \ldots$, α_{m+1} of multiplicity at least $n_{m+1} + 1$. By the previous argument, since P has degree $\leq n$, we must have $P = 0$, that is, $Y = 0$. This concludes the proof. $\square$

Proposition 30.29 shows the existence of unique polynomials $H^i_j(X)$ of degree $\leq n$ such that $\mathrm{D}^i H^i_j(\alpha_j) = 1$ and $\mathrm{D}^k H^i_j(\alpha_l) = 0$, for $k \neq i$ or $l \neq j$, $1 \leq j, l \leq m+1$, $0 \leq i, k \leq n_j$. The polynomials H^i_j are called *Hermite basis polynomials*.

One problem with Proposition 30.29 is that it does not give an explicit way of computing the Hermite basis polynomials. We first show that this can be done explicitly in the special cases $n_1 = \ldots = n_{m+1} = 1$, and $n_1 = \ldots = n_{m+1} = 2$, and then suggest a method using a generalized Newton interpolant.

Assume that $n_1 = \ldots = n_{m+1} = 1$. We try $H^0_j = (a(X - \alpha_j) + b)L_j^2$, and $H^1_j = (c(X - \alpha_j) + d)L_j^2$, where L_j is the Lagrange interpolant determined earlier. Since

$$\mathrm{D}H^0_j = aL_j^2 + 2(a(X - \alpha_j) + b)L_j \mathrm{D}L_j,$$

requiring that $H^0_j(\alpha_j) = 1$, $H^0_j(\alpha_k) = 0$, $\mathrm{D}H^0_j(\alpha_j) = 0$, and $\mathrm{D}H^0_j(\alpha_k) = 0$, for $k \neq j$, implies $b = 1$ and $a = -2\mathrm{D}L_j(\alpha_j)$. Similarly, from the requirements $H^1_j(\alpha_j) = 0$, $H^1_j(\alpha_k) = 0$, $\mathrm{D}H^1_j(\alpha_j) = 1$, and $\mathrm{D}H^1_j(\alpha_k) = 0$, $k \neq j$, we get $c = 1$ and $d = 0$.

Thus, we have the Hermite polynomials

$$H^0_j = (1 - 2\mathrm{D}L_j(\alpha_j)(X - \alpha_j))L_j^2, \qquad H^1_j = (X - \alpha_j)L_j^2.$$

In the special case where $m = 1$, $\alpha_1 = 0$, and $\alpha_2 = 1$, we leave as an exercise to show that the Hermite polynomials are

$$\begin{aligned} H_0^0 &= 2X^3 - 3X^2 + 1, \\ H_1^0 &= -2X^3 + 3X^2, \\ H_0^1 &= X^3 - 2X^2 + X, \\ H_1^1 &= X^3 - X^2. \end{aligned}$$

As a consequence, the polynomial P of degree 3 such that $P(0) = x_0$, $P(1) = x_1$, $P'(0) = m_0$, and $P'(1) = m_1$, can be written as

$$P(X) = x_0(2X^3 - 3X^2 + 1) + m_0(X^3 - 2X^2 + X) + m_1(X^3 - X^2) + x_1(-2X^3 + 3X^2).$$

If we want the polynomial P of degree 3 such that $P(a) = x_0$, $P(b) = x_1$, $P'(a) = m_0$, and $P'(b) = m_1$, where $b \neq a$, then we have

$$P(X) = x_0(2t^3 - 3t^2 + 1) + (b-a)m_0(t^3 - 2t^2 + t) + (b-a)m_1(t^3 - t^2) + x_1(-2t^3 + 3t^2),$$

where

$$t = \frac{X - a}{b - a}.$$

Observe the presence of the extra factor $(b - a)$ in front of m_0 and m_1, the formula would be false otherwise!

We now consider the case where $n_1 = \ldots = n_{m+1} = 2$. Let us try

$$H_j^i(X) = (a^i(X - \alpha_j)^2 + b^i(X - \alpha_j) + c^i)L_j^3,$$

where $0 \leq i \leq 2$. Sparing the readers some (tedious) computations, we find:

$$\begin{aligned} H_j^0(X) &= \Big(\big(6(\mathrm{D}L_j(\alpha_j))^2 - \frac{3}{2}\mathrm{D}^2L_j(\alpha_j)\big)(X - \alpha_j)^2 - 3\mathrm{D}L_j(\alpha_j)(X - \alpha_j) + 1\Big)L_j^3(X), \\ H_j^1(X) &= \Big(9(\mathrm{D}L_j(\alpha_j))^2(X - \alpha_j)^2 - 3\mathrm{D}L_j(\alpha_j)(X - \alpha_j)\Big)L_j^3(X), \\ H_j^2(X) &= \frac{1}{2}(X - \alpha_j)^2 L_j^3(X). \end{aligned}$$

Going back to the general problem, it seems to us that a kind of Newton interpolant will be more manageable. Let

$$\begin{aligned} &P_0^0(X) = 1, \\ &P_j^0(X) = (X - \alpha_1)^{n_1+1} \cdots (X - \alpha_j)^{n_j+1}, \ 1 \leq j \leq m \\ &P_0^i(X) = (X - \alpha_1)^i(X - \alpha_2)^{n_2+1} \cdots (X - \alpha_{m+1})^{n_{m+1}+1}, \ 1 \leq i \leq n_1, \\ &P_j^i(X) = (X - \alpha_1)^{n_1+1} \cdots (X - \alpha_j)^{n_j+1}(X - \alpha_{j+1})^i(X - \alpha_{j+2})^{n_{j+2}+1} \cdots (X - \alpha_{m+1})^{n_{m+1}+1}, \\ &\qquad 1 \leq j \leq m-1, \ 1 \leq i \leq n_{j+1}, \\ &P_m^i(X) = (X - \alpha_1)^{n_1+1} \cdots (X - \alpha_m)^{n_m+1}(X - \alpha_{m+1})^i, \ 1 \leq i \leq n_{m+1}, \end{aligned}$$

and let

$$P(X) = \sum_{j=0,i=0}^{j=m,i=n_{j+1}} \lambda_j^i P_j^i(X).$$

We can think of $P(X)$ as a generalized Newton interpolant. We can compute the derivatives $\mathrm{D}^k P_j^i$, for $1 \leq k \leq n_{j+1}$, and if we look for the Hermite basis polynomials $H_j^i(X)$ such that $\mathrm{D}^i H_j^i(\alpha_j) = 1$ and $\mathrm{D}^k H_j^i(\alpha_l) = 0$, for $k \neq i$ or $l \neq j$, $1 \leq j, l \leq m+1$, $0 \leq i, k \leq n_j$, we find that we have to solve triangular systems of linear equations. Thus, as in the simple case $n_1 = \ldots = n_{m+1} = 0$, we can solve successively for the λ_j^i. Obviously, the computations are quite formidable and we leave such considerations for further study.

Chapter 31

Annihilating Polynomials and the Primary Decomposition

In this chapter all vector spaces are defined over an arbitrary field K.

In Section 7.7 we explained that if $f\colon E \to E$ is a linear map on a K-vector space E, then for any polynomial $p(X) = a_0X^d + a_1X^{d-1} + \cdots + a_d$ with coefficients in the field K, we can define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0f^d + a_1f^{d-1} + \cdots + a_d\mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0f^d(u) + a_1f^{d-1}(u) + \cdots + a_du,$$

for every vector $u \in E$. Then we showed that if E is finite-dimensional and if $\chi_f(X) = \det(XI - f)$ is the characteristic polynomial of f, by the Cayley–Hamilton theorem, we have

$$\chi_f(f) = 0.$$

This fact suggests looking at the set of all polynomials $p(X)$ such that

$$p(f) = 0.$$

Such polynomials are called *annihilating polynomials* of f, the set of all these polynomials, denoted $\mathrm{Ann}(f)$, is called the *annihilator* of f, and the Cayley-Hamilton theorem shows that it is nontrivial since it contains a polynomial of positive degree. It turns out that $\mathrm{Ann}(f)$ contains a polynomial m_f of smallest degree that generates $\mathrm{Ann}(f)$, and this polynomial divides the characteristic polynomial. Furthermore, the polynomial m_f encapsulates a lot of information about f, in particular whether f can be diagonalized. One of the main reasons for this is that a scalar $\lambda \in K$ is a zero of the minimal polynomial m_f if and only if λ is an eigenvalue of f.

The first main result is Theorem 31.6 which states that if $f\colon E \to E$ is a linear map on a finite-dimensional space E, then f is diagonalizable iff its minimal polynomial m is of the form

$$m = (X - \lambda_1) \cdots (X - \lambda_k),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of K.

One of the technical tools used to prove this result is the notion of *f-conductor*; see Definition 31.2. As a corollary of Theorem 31.6 we obtain results about finite commuting families of diagonalizable or triangulable linear maps.

If $f\colon E \to E$ is a linear map and $\lambda \in K$ is an eigenvalue of f, recall that the eigenspace E_λ associated with λ is the kernel of the linear map $\lambda \mathrm{id} - f$. If all the eigenvalues $\lambda_1 \ldots, \lambda_k$ of f are in K and if f is diagonalizable, then

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_k},$$

but in general there are not enough eigenvectors to span E. A remedy is to generalize the notion of eigenvector and look for (nonzero) vectors u (called generalized eigenvectors) such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1.$$

Then, it turns out that if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

then $r = r_i$ does the job for λ_i; that is, if we let

$$W_i = \mathrm{Ker}\,(\lambda_i \mathrm{id} - f)^{r_i},$$

then

$$E = W_1 \oplus \cdots \oplus W_k.$$

The above facts are parts of the *primary decomposition theorem* (Theorem 31.11). It is a special case of a more general result involving the factorization of the minimal polynomial m into its irreducible monic factors; see Theorem 31.10.

Theorem 31.11 implies that every linear map f that has all its eigenvalues in K can be written as $f = D + N$, where D is diagonalizable and N is nilpotent (which means that $N^r = 0$ for some positive integer r). Furthermore D and N commute and are unique. This is the *Jordan decomposition*, Theorem 31.12.

The Jordan decomposition suggests taking a closer look at nilpotent maps. We prove that for any nilpotent linear map $f\colon E \to E$ on a finite-dimensional vector space E of dimension n over a field K, there is a basis of E such that the matrix N of f is of the form

$$N = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$; see Theorem 31.16. As a corollary we obtain the *Jordan form*; which involves matrices of the form

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix},$$

called *Jordan blocks*; see Theorem 31.17.

31.1 Annihilating Polynomials and the Minimal Polynomial

Given a linear map $f\colon E \to E$, it is easy to check that the set $\mathrm{Ann}(f)$ of polynomials that annihilate f is an ideal. Furthermore, when E is finite-dimensional, the Cayley–Hamilton Theorem implies that $\mathrm{Ann}(f)$ is not the zero ideal. Therefore, by Proposition 30.10, there is a unique monic polynomial m_f that generates $\mathrm{Ann}(f)$. Results from Chapter 30, especially about gcd's of polynomials, will come handy.

Definition 31.1. If $f\colon E \to E$ is a linear map on a finite-dimensional vector space E, the unique monic polynomial $m_f(X)$ that generates the ideal $\mathrm{Ann}(f)$ of polynomials which annihilate f (the *annihilator* of f) is called the *minimal polynomial* of f.

The minimal polynomial m_f of f is the monic polynomial of smallest degree that annihilates f. Thus, the minimal polynomial divides the characteristic polynomial χ_f, and $\deg(m_f) \geq 1$. For simplicity of notation, we often write m instead of m_f.

If A is any $n \times n$ matrix, the set $\mathrm{Ann}(A)$ of polynomials that annihilate A is the set of polynomials

$$p(X) = a_0X^d + a_1X^{d-1} + \cdots + a_{d-1}X + a_d$$

such that

$$a_0A^d + a_1A^{d-1} + \cdots + a_{d-1}A + a_dI = 0.$$

It is clear that $\mathrm{Ann}(A)$ is a nonzero ideal and its unique monic generator is called the *minimal polynomial* of A. We check immediately that if Q is an invertible matrix, then A and $Q^{-1}AQ$ have the same minimal polynomial. Also, if A is the matrix of f with respect to some basis, then f and A have the same minimal polynomial.

The zeros (in K) of the minimal polynomial of f and the eigenvalues of f (in K) are intimately related.

Proposition 31.1. *Let $f\colon E \to E$ be a linear map on some finite-dimensional vector space E. Then $\lambda \in K$ is a zero of the minimal polynomial $m_f(X)$ of f iff λ is an eigenvalue of f*

iff λ is a zero of $\chi_f(X)$. Therefore, the minimal and the characteristic polynomials have the same zeros (in K), except for multiplicities.

Proof. First assume that $m(\lambda) = 0$ (with $\lambda \in K$, and writing m instead of m_f). If so, using polynomial division, m can be factored as

$$m = (X - \lambda)q,$$

with $\deg(q) < \deg(m)$. Since m is the minimal polynomial, $q(f) \neq 0$, so there is some nonzero vector $v \in E$ such that $u = q(f)(v) \neq 0$. But then, because m is the minimal polynomial,

$$\begin{aligned} 0 &= m(f)(v) \\ &= (f - \lambda \mathrm{id})(q(f)(v)) \\ &= (f - \lambda \mathrm{id})(u), \end{aligned}$$

which shows that λ is an eigenvalue of f.

Conversely, assume that $\lambda \in K$ is an eigenvalue of f. This means that for some $u \neq 0$, we have $f(u) = \lambda u$. Now it is easy to show that

$$m(f)(u) = m(\lambda)u,$$

and since m is the minimal polynomial of f, we have $m(f)(u) = 0$, so $m(\lambda)u = 0$, and since $u \neq 0$, we must have $m(\lambda) = 0$. $\square$

Proposition 31.2. *Let $f\colon E \to E$ be a linear map on some finite-dimensional vector space E. If f diagonalizable, then its minimal polynomial is a product of distinct factors of degree 1.*

Proof. If we assume that f is diagonalizable, then its eigenvalues are all in K, and if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f, and then by Proposition 31.1, the minimal polynomial m of f must be a product of powers of the polynomials $(X - \lambda_i)$. Actually, we claim that

$$m = (X - \lambda_1) \cdots (X - \lambda_k).$$

For this we just have to show that m annihilates f. However, for any eigenvector u of f, one of the linear maps $f - \lambda_i \mathrm{id}$ sends u to 0, so

$$m(f)(u) = (f - \lambda_1 \mathrm{id}) \circ \cdots \circ (f - \lambda_k \mathrm{id})(u) = 0.$$

Since E is spanned by the eigenvectors of f, we conclude that

$$m(f) = 0. \qquad \square$$

It turns out that the converse of Proposition 31.2 is true, but this will take a little work to establish it.

31.2 Minimal Polynomials of Diagonalizable Linear Maps

In this section we prove that if the minimal polynomial m_f of a linear map f is of the form

$$m_f = (X - \lambda_1) \cdots (X - \lambda_k)$$

for distinct scalars $\lambda_1, \ldots, \lambda_k \in K$, then f is diagonalizable. This is a powerful result that has a number of implications. But first we need of few properties of invariant subspaces.

Given a linear map $f\colon E \to E$, recall that a subspace W of E is *invariant under* f if $f(u) \in W$ for all $u \in W$. For example, if $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ is $f(x, y) = (-x, y)$, the y-axis is invariant under f.

Proposition 31.3. *Let W be a subspace of E invariant under the linear map $f\colon E \to E$ (where E is finite-dimensional). Then the minimal polynomial of the restriction $f \mid W$ of f to W divides the minimal polynomial of f, and the characteristic polynomial of $f \mid W$ divides the characteristic polynomial of f.*

Sketch of proof. The key ingredient is that we can pick a basis $(e_1, \ldots, e_n)$ of E in which $(e_1, \ldots, e_k)$ is a basis of W. The matrix of f over this basis is a block matrix of the form

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix},$$

where B is a $k \times k$ matrix, D is an $(n-k) \times (n-k)$ matrix, and C is a $k \times (n-k)$ matrix. Then

$$\det(XI - A) = \det(XI - B)\det(XI - D),$$

which implies the statement about the characteristic polynomials. Furthermore,

$$A^i = \begin{pmatrix} B^i & C_i \\ 0 & D^i \end{pmatrix},$$

for some $k \times (n-k)$ matrix C_i. It follows that any polynomial which annihilates A also annihilates B and D. So the minimal polynomial of B divides the minimal polynomial of A. $\square$

For the next step, there are at least two ways to proceed. We can use an old-fashion argument using Lagrange interpolants, or we can use a slight generalization of the notion of annihilator. We pick the second method because it illustrates nicely the power of principal ideals.

What we need is the notion of conductor (also called transporter).

Definition 31.2. Let $f\colon E \to E$ be a linear map on a finite-dimensional vector space E, let W be an invariant subspace of f, and let u be any vector in E. The set $S_f(u, W)$ consisting of all polynomials $q \in K[X]$ such that $q(f)(u) \in W$ is called the *f-conductor of u into W*.

Observe that the minimal polynomial m_f of f always belongs to $S_f(u, W)$, so this is a nontrivial set. Also, if $W = (0)$, then $S_f(u, (0))$ is just the annihilator of f. The crucial property of $S_f(u, W)$ is that it is an ideal.

Proposition 31.4. *If W is an invariant subspace for f, then for each $u \in E$, the f-conductor $S_f(u, W)$ is an ideal in $K[X]$.*

We leave the proof as a simple exercise, using the fact that if W invariant under f, then W is invariant under every polynomial $q(f)$ in $S_f(u, W)$.

Since $S_f(u, W)$ is an ideal, it is generated by a unique monic polynomial q of smallest degree, and because the minimal polynomial m_f of f is in $S_f(u, W)$, the polynomial q divides m.

Definition 31.3. The unique monic polynomial which generates $S_f(u, W)$ is called the *conductor of u into W.*

Example 31.1. For example, suppose $f \colon \mathbb{R}^2 \to \mathbb{R}^2$ where $f(x, y) = (x, 0)$. Observe that $W = \{(x, 0) \in \mathbb{R}^2\}$ is invariant under f. By representing f as $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, we see that $m_f(X) = \chi_f(X) = X^2 - X$. Let $u = (0, y)$. Then $S_f(u, W) = (X)$ and we say X is the conductor of u into W.

Proposition 31.5. *Let $f \colon E \to E$ be a linear map on a finite-dimensional space E and assume that the minimal polynomial m of f is of the form*

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K. If W is a proper subspace of E which is invariant under f, then there is a vector $u \in E$ with the following properties:

(a) $u \notin W$;

(b) $(f - \lambda \mathrm{id})(u) \in W$, for some eigenvalue λ of f.

Proof. Observe that (a) and (b) together assert that the conductor of u into W is a polynomial of the form $X - \lambda_i$. Pick any vector $v \in E$ not in W, and let g be the conductor of v into W, i.e. $g(f)(v) \in W$. Since g divides m and $v \notin W$, the polynomial g is not a constant, and thus it is of the form

$$g = (X - \lambda_1)^{s_1} \cdots (X - \lambda_k)^{s_k},$$

with at least some $s_i > 0$. Choose some index j such that $s_j > 0$. Then $X - \lambda_j$ is a factor of g, so we can write

$$g = (X - \lambda_j)q. \tag{*}$$

By definition of g, the vector $u = q(f)(v)$ cannot be in W, since otherwise g would not be of minimal degree. However, $(*)$ implies that

$$\begin{aligned}(f - \lambda_j \mathrm{id})(u) &= (f - \lambda_j \mathrm{id})(q(f)(v)) \\ &= g(f)(v)\end{aligned}$$

is in W, which concludes the proof. $\square$

We can now prove the main result of this section.

Theorem 31.6. *Let $f\colon E \to E$ be a linear map on a finite-dimensional space E. Then f is diagonalizable iff its minimal polynomial m is of the form*

$$m = (X - \lambda_1) \cdots (X - \lambda_k),$$

where $\lambda_1, \ldots, \lambda_k$ are distinct elements of K.

Proof. We already showed in Proposition 31.2 that if f is diagonalizable, then its minimal polynomial is of the above form (where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f).

For the converse, let W be the subspace spanned by all the eigenvectors of f. If $W \neq E$, since W is invariant under f, by Proposition 31.5, there is some vector $u \notin W$ such that for some λ_j, we have

$$(f - \lambda_j \mathrm{id})(u) \in W.$$

Let $v = (f - \lambda_j \mathrm{id})(u) \in W$. Since $v \in W$, we can write

$$v = w_1 + \cdots + w_k$$

where $f(w_i) = \lambda_i w_i$ (either $w_i = 0$ or w_i is an eigenvector for λ_i), and so for every polynomial h, we have

$$h(f)(v) = h(\lambda_1) w_1 + \cdots + h(\lambda_k) w_k,$$

which shows that $h(f)(v) \in W$ for every polynomial h. We can write

$$m = (X - \lambda_j) q$$

for some polynomial q, and also

$$q - q(\lambda_j) = p(X - \lambda_j)$$

for some polynomial p. We know that $p(f)(v) \in W$, and since m is the minimal polynomial of f, we have

$$0 = m(f)(u) = (f - \lambda_j \mathrm{id})(q(f)(u)),$$

which implies that $q(f)(u) \in W$ (either $q(f)(u) = 0$, or it is an eigenvector associated with λ_j). However,

$$q(f)(u) - q(\lambda_j) u = p(f)((f - \lambda_j \mathrm{id})(u)) = p(f)(v),$$

and since $p(f)(v) \in W$ and $q(f)(u) \in W$, we conclude that $q(\lambda_j) u \in W$. But, $u \notin W$, which implies that $q(\lambda_j) = 0$, so λ_j is a double root of m, a contradiction. Therefore, we must have $W = E$. $\square$

Remark: Proposition 31.5 can be used to give a quick proof of Theorem 15.5.

31.3 Commuting Families of Diagonalizable and Triangulable Maps

Using Theorem 31.6, we can give a short proof about commuting diagonalizable linear maps.

Definition 31.4. If $\mathcal{F}$ is a family of linear maps on a vector space E, we say that $\mathcal{F}$ is a *commuting family* iff $f \circ g = g \circ f$ for all $f, g \in \mathcal{F}$.

Proposition 31.7. *Let $\mathcal{F}$ be a finite commuting family of diagonalizable linear maps on a vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by a diagonal matrix.*

Proof. We proceed by induction on $n = \dim(E)$. If $n = 1$, there is nothing to prove. If $n > 1$, there are two cases. If all linear maps in $\mathcal{F}$ are of the form $\lambda\mathrm{id}$ for some $\lambda \in K$, then the proposition holds trivially. In the second case, let $f \in \mathcal{F}$ be some linear map in $\mathcal{F}$ which is not a scalar multiple of the identity. In this case, f has at least two distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and because f is diagonalizable, E is the direct sum of the corresponding eigenspaces $E_{\lambda_1}, \ldots, E_{\lambda_k}$. For every index i, the eigenspace E_{λ_i} is invariant under f and under every other linear map g in $\mathcal{F}$, since for any $g \in \mathcal{F}$ and any $u \in E_{\lambda_i}$, because f and g commute, we have

$$f(g(u)) = g(f(u)) = g(\lambda_i u) = \lambda_i g(u)$$

so $g(u) \in E_{\lambda_i}$. Let $\mathcal{F}_i$ be the family obtained by restricting each $f \in \mathcal{F}$ to E_{λ_i}. By Proposition 31.3, the minimal polynomial of every linear map $f \mid E_{\lambda_i}$ in $\mathcal{F}_i$ divides the minimal polynomial m_f of f, and since f is diagonalizable, m_f is a product of distinct linear factors, so the minimal polynomial of $f \mid E_{\lambda_i}$ is also a product of distinct linear factors. By Theorem 31.6, the linear map $f \mid E_{\lambda_i}$ is diagonalizable. Since $k > 1$, we have $\dim(E_{\lambda_i}) < \dim(E)$ for $i = 1, \ldots, k$, and by the induction hypothesis, for each i there is a basis of E_{λ_i} over which $f \mid E_{\lambda_i}$ is represented by a diagonal matrix. Since the above argument holds for all i, by combining the bases of the E_{λ_i}, we obtain a basis of E such that the matrix of every linear map $f \in \mathcal{F}$ is represented by a diagonal matrix. $\square$

Remark: Proposition 31.7 also holds for infinite commuting families $\mathcal{F}$ of diagonalizable linear maps, because E being finite dimensional, there is a finite subfamily of linearly independent linear maps in $\mathcal{F}$ spanning $\mathcal{F}$.

There is also an analogous result for commuting families of linear maps represented by upper triangular matrices. To prove this we need the following proposition.

Proposition 31.8. *Let $\mathcal{F}$ be a nonempty finite commuting family of triangulable linear maps on a finite-dimensional vector space E. Let W be a proper subspace of E which is invariant under $\mathcal{F}$. Then there exists a vector $u \in E$ such that:*

1. $u \notin W$.

2. For every $f \in \mathcal{F}$, the vector $f(u)$ belongs to the subspace $W \oplus Ku$ spanned by W and u.

Proof. By renaming the elements of $\mathcal{F}$ if necessary, we may assume that $(f_1, \ldots, f_r)$ is a basis of the subspace of $\mathrm{End}(E)$ spanned by $\mathcal{F}$. We prove by induction on r that there exists some vector $u \in E$ such that

1. $u \notin W$.

2. $(f_i - \alpha_i \mathrm{id})(u) \in W$ for $i = 1, \ldots, r$, for some scalars $\alpha_i \in K$.

Consider the base case $r = 1$. Since f_1 is triangulable, its eigenvalues all belong to K since they are the diagonal entries of the triangular matrix associated with f_1 (this is the easy direction of Theorem 15.5), so the minimal polynomial of f_1 is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_1 belong to K. We conclude by applying Proposition 31.5.

Next assume that $r \geq 2$ and that the induction hypothesis holds for $f_1, \ldots, f_{r-1}$. Thus, there is a vector $u_{r-1} \in E$ such that

1. $u_{r-1} \notin W$.

2. $(f_i - \alpha_i \mathrm{id})(u_{r-1}) \in W$ for $i = 1, \ldots, r-1$, for some scalars $\alpha_i \in K$.

Let
$$V_{r-1} = \{w \in E \mid (f_i - \alpha_i \mathrm{id})(w) \in W,\ i = 1, \ldots, r-1\}.$$
Clearly, $W \subseteq V_{r-1}$ and $u_{r-1} \in V_{r-1}$. We claim that V_{r-1} is invariant under $\mathcal{F}$. This is because, for any $v \in V_{r-1}$ and any $f \in \mathcal{F}$, since f and f_i commute, we have

$$(f_i - \alpha_i \mathrm{id})(f(v)) = f((f_i - \alpha_i \mathrm{id})(v)), \quad 1 \leq i \leq r-1.$$

Now $(f_i - \alpha_i \mathrm{id})(v) \in W$ because $v \in V_{r-1}$, and W is invariant under $\mathcal{F}$, so $f(f_i - \alpha_i \mathrm{id})(v)) \in W$, that is, $(f_i - \alpha_i \mathrm{id})(f(v)) \in W$.

Consider the restriction g_r of f_r to V_{r-1}. The minimal polynomial of g_r divides the minimal polynomial of f_r, and since f_r is triangulable, just as we saw for f_1, the minimal polynomial of f_r is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

where the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f_r belong to K, so the minimal polynomial of g_r is of the same form. By Proposition 31.5, there is some vector $u_r \in V_{r-1}$ such that

1. $u_r \notin W$.

2. $(g_r - \alpha_r \mathrm{id})(u_r) \in W$ for some scalars $\alpha_r \in K$.

Now since $u_r \in V_{r-1}$, we have $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r-1$, so $(f_i - \alpha_i \mathrm{id})(u_r) \in W$ for $i = 1, \ldots, r$ (since g_r is the restriction of f_r), which concludes the proof of the induction step. Finally, since every $f \in \mathcal{F}$ is the linear combination of $(f_1, \ldots, f_r)$, Condition (2) of the inductive claim implies Condition (2) of the proposition. $\square$

We can now prove the following result.

Proposition 31.9. *Let $\mathcal{F}$ be a nonempty finite commuting family of triangulable linear maps on a finite-dimensional vector space E. There exists a basis of E such that every linear map in $\mathcal{F}$ is represented in that basis by an upper triangular matrix.*

Proof. Let $n = \dim(E)$. We construct inductively a basis $(u_1, \ldots, u_n)$ of E such that if W_i is the subspace spanned by $(u_1 \ldots, u_i)$, then for every $f \in \mathcal{F}$,

$$f(u_i) = a^f_{1i} u_1 + \cdots + a^f_{ii} u_i,$$

for some $a^f_{ij} \in K$; that is, $f(u_i)$ belongs to the subspace W_i.

We begin by applying Proposition 31.8 to the subspace $W_0 = (0)$ to get u_1 so that for all $f \in \mathcal{F}$,

$$f(u_1) = \alpha^f_1 u_1.$$

For the induction step, since W_i invariant under $\mathcal{F}$, we apply Proposition 31.8 to the subspace W_i, to get $u_{i+1} \in E$ such that

1. $u_{i+1} \notin W_i$.

2. For every $f \in \mathcal{F}$, the vector $f(u_{i+1})$ belong to the subspace spanned by W_i and u_{i+1}.

Condition (1) implies that $(u_1, \ldots, u_i, u_{i+1})$ is linearly independent, and Condition (2) means that for every $f \in \mathcal{F}$,

$$f(u_{i+1}) = a^f_{1i+1} u_1 + \cdots + a^f_{i+1i+1} u_{i+1},$$

for some $a^f_{i+1j} \in K$, establishing the induction step. After n steps, each $f \in \mathcal{F}$ is represented by an upper triangular matrix. $\square$

Observe that if $\mathcal{F}$ consists of a single linear map f and if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

with all $\lambda_i \in K$, using Proposition 31.5 instead of Proposition 31.8, the proof of Proposition 31.9 yields another proof of Theorem 15.5.

31.4 The Primary Decomposition Theorem

If $f\colon E \to E$ is a linear map and $\lambda \in K$ is an eigenvalue of f, recall that the eigenspace E_λ associated with λ is the kernel of the linear map $\lambda \mathrm{id} - f$. If all the eigenvalues $\lambda_1 \ldots, \lambda_k$ of f are in K, it may happen that

$$E = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_k},$$

but in general there are not enough eigenvectors to span E. What if we generalize the notion of eigenvector and look for (nonzero) vectors u such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1?$$

It turns out that if the minimal polynomial of f is of the form

$$m = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k},$$

then $r = r_i$ does the job for λ_i; that is, if we let

$$W_i = \mathrm{Ker}\,(\lambda_i \mathrm{id} - f)^{r_i},$$

then

$$E = W_1 \oplus \cdots \oplus W_k.$$

This result is very nice but seems to require that the eigenvalues of f all belong to K. Actually, it is a special case of a more general result involving the factorization of the minimal polynomial m into its irreducible monic factors (see Theorem 30.17),

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K.

Theorem 31.10. *(Primary Decomposition Theorem) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. Write the minimal polynomial m of f as*

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K, and the r_i are positive integers. Let

$$W_i = \mathrm{Ker}\,(p_i^{r_i}(f)), \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.

(b) Each W_i is invariant under f.

(c) The minimal polynomial of the restriction $f \mid W_i$ of f to W_i is $p_i^{r_i}$.

Proof. The trick is to construct projections π_i using the polynomials $p_j^{r_j}$ so that the range of π_i is equal to W_i. Let

$$g_i = m/p_i^{r_i} = \prod_{j \neq i} p_j^{r_j}.$$

Note that

$$p_i^{r_i} g_i = m.$$

Since $p_1, \ldots, p_k$ are irreducible and distinct, they are relatively prime. Then using Proposition 30.14, it is easy to show that $g_1, \ldots, g_k$ are relatively prime. Otherwise, some irreducible polynomial p would divide all of $g_1, \ldots, g_k$, so by Proposition 30.14 it would be equal to one of the irreducible factors p_i. But that p_i is missing from g_i, a contradiction. Therefore, by Proposition 30.15, there exist some polynomials $h_1, \ldots, h_k$ such that

$$g_1 h_1 + \cdots + g_k h_k = 1.$$

Let $q_i = g_i h_i$ and let $\pi_i = q_i(f) = g_i(f) h_i(f)$. We have

$$q_1 + \cdots + q_k = 1,$$

and since m divides $q_i q_j$ for $i \neq j$, we get

$$\begin{aligned} \pi_1 + \cdots + \pi_k &= \mathrm{id} \\ \pi_i \pi_j &= 0, \quad i \neq j. \end{aligned}$$

(We implicitly used the fact that if p, q are two polynomials, the linear maps $p(f) \circ q(f)$ and $q(f) \circ p(f)$ are the same since $p(f)$ and $q(f)$ are polynomials in the powers of f, which commute.) Composing the first equation with π_i and using the second equation, we get

$$\pi_i^2 = \pi_i.$$

Therefore, the π_i are projections, and E is the direct sum of the images of the π_i. Indeed, every $u \in E$ can be expressed as

$$u = \pi_1(u) + \cdots + \pi_k(u).$$

Also, if

$$\pi_1(u) + \cdots + \pi_k(u) = 0,$$

then by applying π_i we get

$$0 = \pi_i^2(u) = \pi_i(u), \quad i = 1, \ldots k.$$

To finish proving (a), we need to show that

$$W_i = \mathrm{Ker}\,(p_i^{r_i}(f)) = \pi_i(E).$$

If $v \in \pi_i(E)$, then $v = \pi_i(u)$ for some $u \in E$, so

$$\begin{aligned} p_i^{r_i}(f)(v) &= p_i^{r_i}(f)(\pi_i(u)) \\ &= p_i^{r_i}(f)g_i(f)h_i(f)(u) \\ &= h_i(f)p_i^{r_i}(f)g_i(f)(u) \\ &= h_i(f)m(f)(u) = 0, \end{aligned}$$

because m is the minimal polynomial of f. Therefore, $v \in W_i$.

Conversely, assume that $v \in W_i = \mathrm{Ker}\,(p_i^{r_i}(f))$. If $j \neq i$, then $g_j h_j$ is divisible by $p_i^{r_i}$, so

$$g_j(f)h_j(f)(v) = \pi_j(v) = 0, \quad j \neq i.$$

Then since $\pi_1 + \cdots + \pi_k = \mathrm{id}$, we have $v = \pi_i v$, which shows that v is in the range of π_i. Therefore, $W_i = \mathrm{Im}(\pi_i)$, and this finishes the proof of (a).

If $p_i^{r_i}(f)(u) = 0$, then $p_i^{r_i}(f)(f(u)) = f(p_i^{r_i}(f)(u)) = 0$, so (b) holds.

If we write $f_i = f \mid W_i$, then $p_i^{r_i}(f_i) = 0$, because $p_i^{r_i}(f) = 0$ on W_i (its kernel). Therefore, the minimal polynomial of f_i divides $p_i^{r_i}$. Conversely, let q be any polynomial such that $q(f_i) = 0$ (on W_i). Since $m = p_i^{r_i} g_i$, the fact that $m(f)(u) = 0$ for all $u \in E$ shows that

$$p_i^{r_i}(f)(g_i(f)(u)) = 0, \quad u \in E,$$

and thus $\mathrm{Im}(g_i(f)) \subseteq \mathrm{Ker}\,(p_i^{r_i}(f)) = W_i$. Consequently, since $q(f)$ is zero on W_i,

$$q(f)g_i(f) = 0 \quad \text{for all } u \in E.$$

But then qg_i is divisible by the minimal polynomial $m = p_i^{r_i} g_i$ of f, and since $p_i^{r_i}$ and g_i are relatively prime, by Euclid's proposition, $p_i^{r_i}$ must divide q. This finishes the proof that the minimal polynomial of f_i is $p_i^{r_i}$, which is (c). $\square$

To best understand the projection constructions of Theorem 31.10, we provide the following two explicit examples of the primary decomposition theorem.

Example 31.2. First let $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ be defined as $f(x, y, z) = (y, -x, z)$. In terms of the standard basis f is represented by the 3×3 matrix $X_f := \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Then a simple calculation shows that $m_f(x) = \chi_f(x) = (x^2+1)(x-1)$. Using the notation of the preceding proof set

$$m = p_1 p_2, \qquad p_1 = x^2 + 1, \qquad p_2 = x - 1.$$

Then

$$g_1 = \frac{m}{p_1} = x - 1, \qquad g_2 = \frac{m}{p_2} = x^2 + 1.$$

We must find $h_1, h_2 \in \mathbb{R}[x]$ such that $g_1h_1 + g_2h_2 = 1$. In general this is the hard part of the projection construction. But since we are only working with two relatively prime polynomials g_1, g_2, we may apply the Euclidean algorithm to discover that

$$-\frac{x+1}{2}(x-1) + \frac{1}{2}(x^2+1) = 1,$$

where $h_1 = -\frac{x+1}{2}$ while $h_2 = \frac{1}{2}$. By definition

$$\pi_1 = g_1(f)h_1(f) = -\frac{1}{2}(X_f - \mathrm{id})(X_f + \mathrm{id}) = -\frac{1}{2}(X_f^2 - \mathrm{id}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$\pi_2 = g_2(f)h_2(f) = \frac{1}{2}(X_f^2 + \mathrm{id}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $\mathbb{R}^3 = W_1 \oplus W_2$, where

$$W_1 = \pi_1(\mathbb{R}^3) = \mathrm{Ker}\,(p_1(X_f)) = \mathrm{Ker}\,(X_f^2 + \mathrm{id}) = \mathrm{Ker}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \{(x, y, 0) \in \mathbb{R}^3\},$$

$$W_2 = \pi_2(\mathbb{R}^3) = \mathrm{Ker}\,(p_2(X_f)) = \mathrm{Ker}\,(X_f - \mathrm{id}) = \mathrm{Ker}\begin{pmatrix} -1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \{(0, 0, z) \in \mathbb{R}^3\}.$$

Example 31.3. For our second example of the primary decomposition theorem let $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ be defined as $f(x, y, z) = (y, -x + z, -y)$, with standard matrix representation $X_f = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$. A simple calculation shows that $m_f(x) = \chi_f(x) = x(x^2 + 2)$. Set

$$p_1 = x^2 + 2, \qquad p_2 = x, \qquad g_1 = \frac{m_f}{p_1} = x, \qquad g_2 = \frac{m_f}{p_2} = x^2 + 2.$$

Since $\gcd(g_1, g_2) = 1$, we use the Euclidean algorithm to find

$$h_1 = -\frac{1}{2}x, \qquad h_2 = \frac{1}{2},$$

such that $g_1h_1 + g_2h_2 = 1$. Then

$$\pi_1 = g_1(f)h_1(f) = -\frac{1}{2}X_f^2 = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix},$$

while

$$\pi_2 = g_2(f)h_2(f) = \frac{1}{2}(X_f^2 + 2\mathrm{id}) = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Although it is not entirely obvious, π_1 and π_2 are indeed projections since

$$\pi_1^2 = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \pi_1,$$

and

$$\pi_2^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \pi_2.$$

Furthermore observe that $\pi_1 + \pi_2 = \mathrm{id}$. The primary decomposition theorem implies that $\mathbb{R}^3 = W_1 \oplus W_2$ where

$$\begin{aligned} W_1 &= \pi_1(\mathbb{R}^3) = \mathrm{Ker}\,(p_1(f)) = \mathrm{Ker}\,(X^2+2) = \mathrm{Ker}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \mathrm{span}\{(0,1,0),(1,0,-1)\}, \\ W_2 &= \pi_2(\mathbb{R}^3) = \mathrm{Ker}\,(p_2(f)) = \mathrm{Ker}\,(X) = \mathrm{span}\{(1,0,1)\}. \end{aligned}$$

See Figure 31.1.

If all the eigenvalues of f belong to the field K, we obtain the following result.

Theorem 31.11. *(Primary Decomposition Theorem, Version 2) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, write*

$$m = (X-\lambda_1)^{r_1}\cdots(X-\lambda_k)^{r_k}$$

for the minimal polynomial of f,

$$\chi_f = (X-\lambda_1)^{n_1}\cdots(X-\lambda_k)^{n_k}$$

for the characteristic polynomial of f, with $1 \leq r_i \leq n_i$, and let

$$W_i = \mathrm{Ker}\,(\lambda_i\mathrm{id} - f)^{r_i}, \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.

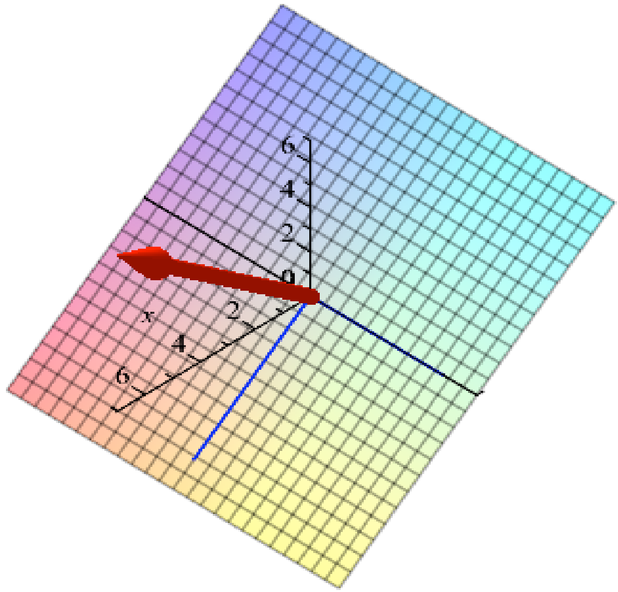


Figure 31.1: The direct sum decomposition of $\mathbb{R}^3 = W_1 \oplus W_2$ where W_1 is the plane $x + z = 0$ and W_2 is line $t(1, 0, 1)$. The spanning vectors of W_1 are in blue.

(b) Each W_i is invariant under f.

(c) $\dim(W_i) = n_i$.

(d) The minimal polynomial of the restriction $f \mid W_i$ of f to W_i is $(X - \lambda_i)^{r_i}$.

Proof. Parts (a), (b) and (d) have already been proven in Theorem 31.10, so it remains to prove (c). Since W_i is invariant under f, let f_i be the restriction of f to W_i. The characteristic polynomial χ_{f_i} of f_i divides $\chi(f)$, and since $\chi(f)$ has all its roots in K, so does $\chi_i(f)$. By Theorem 15.5, there is a basis of W_i in which f_i is represented by an upper triangular matrix, and since $(\lambda_i \mathrm{id} - f)^{r_i} = 0$, the diagonal entries of this matrix are equal to λ_i. Consequently,

$$\chi_{f_i} = (X - \lambda_i)^{\dim(W_i)},$$

and since χ_{f_i} divides $\chi(f)$, we conclude hat

$$\dim(W_i) \leq n_i, \quad i = 1, \ldots, k.$$

Because E is the direct sum of the W_i, we have $\dim(W_1) + \cdots + \dim(W_k) = n$, and since $n_1 + \cdots + n_k = n$, we must have

$$\dim(W_i) = n_i, \quad i = 1, \ldots, k,$$

proving (c). $\square$

Definition 31.5. If $\lambda \in K$ is an eigenvalue of f, we define a *generalized eigenvector* of f as a nonzero vector $u \in E$ such that

$$(\lambda \mathrm{id} - f)^r(u) = 0, \quad \text{for some } r \geq 1.$$

The *index* of λ is defined as the smallest $r \geq 1$ such that

$$\mathrm{Ker}\,(\lambda \mathrm{id} - f)^r = \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{r+1}.$$

It is clear that $\mathrm{Ker}\,(\lambda \mathrm{id} - f)^i \subseteq \mathrm{Ker}\,(\lambda \mathrm{id} - f)^{i+1}$ for all $i \geq 1$. By Theorem 31.11(d), if $\lambda = \lambda_i$, the index of λ_i is equal to r_i.

31.5 Jordan Decomposition

Recall that a linear map $g\colon E \to E$ is said to be *nilpotent* if there is some positive integer r such that $g^r = 0$. Another important consequence of Theorem 31.11 is that f can be written as the sum of a diagonalizable and a nilpotent linear map (which commute). For example $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the $\mathbb{R}$-linear map $f(x, y) = (x, x + y)$ with standard matrix representation $X_f = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. A basic calculation shows that $m_f(x) = \chi_f(x) = (x-1)^2$. By Theorem 31.6 we know that f is not diagonalizable over $\mathbb{R}$. But since the eigenvalue $\lambda_1 = 1$ of f does belong to $\mathbb{R}$, we may use the projection construction inherent within Theorem 31.11 to write $f = D + N$, where D is a diagonalizable linear map and N is a nilpotent linear map. The proof of Theorem 31.10 implies that

$$p_1^{r_1} = (x-1)^2, \qquad g_1 = 1 = h_1, \qquad \pi_1 = g_1(f)h_1(f) = \mathrm{id}.$$

Then

$$D = \lambda_1 \pi_1 = \mathrm{id}, \qquad N = f - D = f(x, y) - \mathrm{id}(x, y) = (x, x + y) - (x, y) = (0, y),$$

which is equivalent to the matrix decomposition

$$X_f = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

This example suggests that the diagonal summand of f is related to the projection constructions associated with the proof of the primary decomposition theorem. If we write

$$D = \lambda_1 \pi_1 + \cdots + \lambda_k \pi_k,$$

where π_i is the projection from E onto the subspace W_i defined in the proof of Theorem 31.10, since

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we have

$$f = f\pi_1 + \cdots + f\pi_k,$$

and so we get

$$N = f - D = (f - \lambda_1 \mathrm{id})\pi_1 + \cdots + (f - \lambda_k \mathrm{id})\pi_k.$$

We claim that $N = f - D$ is a nilpotent operator. Since by construction the π_i are polynomials in f, they commute with f, using the properties of the π_i, we get

$$N^r = (f - \lambda_1 \mathrm{id})^r \pi_1 + \cdots + (f - \lambda_k \mathrm{id})^r \pi_k.$$

Therefore, if $r = \max\{r_i\}$, we have $(f - \lambda_k \mathrm{id})^r = 0$ for $i = 1, \ldots, k$, which implies that

$$N^r = 0.$$

It remains to show that D is diagonalizable. Since N is a polynomial in f, it commutes with f, and thus with D. From

$$D = \lambda_1 \pi_1 + \cdots + \lambda_k \pi_k,$$

and

$$\pi_1 + \cdots + \pi_k = \mathrm{id},$$

we see that

$$\begin{aligned} D - \lambda_i \mathrm{id} &= \lambda_1 \pi_1 + \cdots + \lambda_k \pi_k - \lambda_i(\pi_1 + \cdots + \pi_k) \\ &= (\lambda_1 - \lambda_i)\pi_1 + \cdots + (\lambda_{i-1} - \lambda_i)\pi_{i-1} + (\lambda_{i+1} - \lambda_i)\pi_{i+1} + \cdots + (\lambda_k - \lambda_i)\pi_k. \end{aligned}$$

Since the projections π_j with $j \neq i$ vanish on W_i, the above equation implies that $D - \lambda_i \mathrm{id}$ vanishes on W_i and that $(D - \lambda_j \mathrm{id})(W_i) \subseteq W_i$, and thus that the minimal polynomial of D is

$$(X - \lambda_1) \cdots (X - \lambda_k).$$

Since the λ_i are distinct, by Theorem 31.6, the linear map D is diagonalizable.

In summary we have shown that when all the eigenvalues of f belong to K, there exist a diagonalizable linear map D and a nilpotent linear map N such that

$$\begin{aligned} f &= D + N \\ DN &= ND, \end{aligned}$$

and N and D are polynomials in f.

Definition 31.6. A decomposition of f as $f = D + N$ as above is called a *Jordan decomposition*.

In fact, we can prove more: the maps D and N are uniquely determined by f.

Theorem 31.12. *(Jordan Decomposition) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. If all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of f belong to K, then there exist a diagonalizable linear map D and a nilpotent linear map N such that*

$$\begin{aligned} f &= D + N \\ DN &= ND. \end{aligned}$$

Furthermore, D and N are uniquely determined by the above equations and they are polynomials in f.

Proof. We already proved the existence part. Suppose we also have $f = D' + N'$, with $D'N' = N'D'$, where D' is diagonalizable, N' is nilpotent, and both are polynomials in f. We need to prove that $D = D'$ and $N = N'$.

Since D' and N' commute with one another and $f = D' + N'$, we see that D' and N' commute with f. Then D' and N' commute with any polynomial in f; hence they commute with D and N. From

$$D + N = D' + N',$$

we get

$$D - D' = N' - N,$$

and D, D', N, N' commute with one another. Since D and D' are both diagonalizable and commute, by Proposition 31.7, they are simultaneousy diagonalizable, so $D - D'$ is diagonalizable. Since N and N' commute, by the binomial formula, for any $r \geq 1$,

$$(N' - N)^r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} (N')^{r-j} N^j.$$

Since both N and N' are nilpotent, we have $N^{r_1} = 0$ and $(N')^{r_2} = 0$, for some $r_1, r_2 > 0$, so for $r \geq r_1 + r_2$, the right-hand side of the above expression is zero, which shows that $N' - N$ is nilpotent. (In fact, it is easy that $r_1 = r_2 = n$ works). It follows that $D - D' = N' - N$ is both diagonalizable and nilpotent. Clearly, the minimal polynomial of a nilpotent linear map is of the form X^r for some $r > 0$ (and $r \leq \dim(E)$). But $D - D'$ is diagonalizable, so its minimal polynomial has simple roots, which means that $r = 1$. Therefore, the minimal polynomial of $D - D'$ is X, which says that $D - D' = 0$, and then $N = N'$. □

If K is an algebraically closed field, then Theorem 31.12 holds. This is the case when $K = \mathbb{C}$. This theorem reduces the study of linear maps (from E to itself) to the study of nilpotent operators. There is a special normal form for such operators which is discussed in the next section.

31.6 Nilpotent Linear Maps and Jordan Form

This section is devoted to a normal form for nilpotent maps. We follow Godement's exposition [76]. Let $f\colon E \to E$ be a nilpotent linear map on a finite-dimensional vector space over a field K, and assume that f is not the zero map. There is a smallest positive integer $r \geq 1$ such $f^r \neq 0$ and $f^{r+1} = 0$. Clearly, the polynomial X^{r+1} annihilates f, and it is the minimal polynomial of f since $f^r \neq 0$. It follows that $r + 1 \leq n = \dim(E)$. Let us define the subspaces N_i by

$$N_i = \mathrm{Ker}\,(f^i), \quad i \geq 0.$$

Note that $N_0 = (0)$, $N_1 = \mathrm{Ker}\,(f)$, and $N_{r+1} = E$. Also, it is obvious that

$$N_i \subseteq N_{i+1}, \quad i \geq 0.$$

Proposition 31.13. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$ as above, the inclusions in the following sequence are strict:*

$$(0) = N_0 \subset N_1 \subset \cdots \subset N_r \subset N_{r+1} = E.$$

Proof. We proceed by contradiction. Assume that $N_i = N_{i+1}$ for some i with $0 \leq i \leq r$. Since $f^{r+1} = 0$, for every $u \in E$, we have

$$0 = f^{r+1}(u) = f^{i+1}(f^{r-i}(u)),$$

which shows that $f^{r-i}(u) \in N_{i+1}$. Since $N_i = N_{i+1}$, we get $f^{r-i}(u) \in N_i$, and thus $f^r(u) = 0$. Since this holds for all $u \in E$, we see that $f^r = 0$, a contradiction. $\square$

Proposition 31.14. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, for any integer i with $1 \leq i \leq r$, for any subspace U of E, if $U \cap N_i = (0)$, then $f(U) \cap N_{i-1} = (0)$, and the restriction of f to U is an isomorphism onto $f(U)$.*

Proof. Pick $v \in f(U) \cap N_{i-1}$. We have $v = f(u)$ for some $u \in U$ and $f^{i-1}(v) = 0$, which means that $f^i(u) = 0$. Then $u \in U \cap N_i$, so $u = 0$ since $U \cap N_i = (0)$, and $v = f(u) = 0$. Therefore, $f(U) \cap N_{i-1} = (0)$. The restriction of f to U is obviously surjective on $f(U)$. Suppose that $f(u) = 0$ for some $u \in U$. Then $u \in U \cap N_1 \subseteq U \cap N_i = (0)$ (since $i \geq 1$), so $u = 0$, which proves that f is also injective on U. $\square$

Proposition 31.15. *Given a nilpotent linear map f with $f^r \neq 0$ and $f^{r+1} = 0$, there exists a sequence of subspace $U_1, \ldots, U_{r+1}$ of E with the following properties:*

(1) $N_i = N_{i-1} \oplus U_i$, for $i = 1, \ldots, r + 1$.

(2) We have $f(U_i) \subseteq U_{i-1}$, and the restriction of f to U_i is an injection, for $i = 2, \ldots, r+1$.

See Figure 31.2.

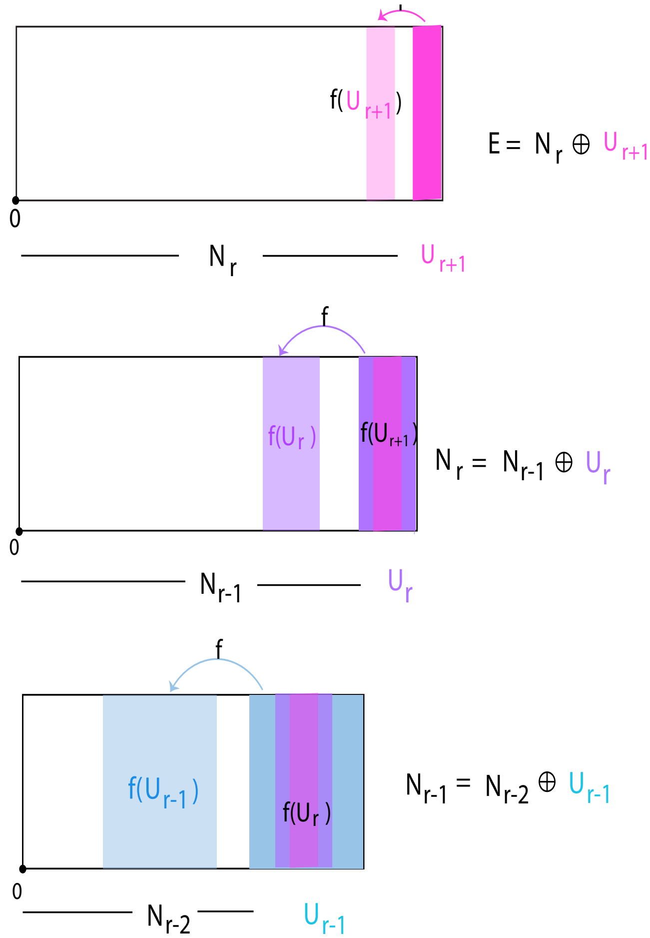


Figure 31.2: A schematic illustration of $N_i = N_{i-1} \oplus U_i$ with $f(U_i) \subseteq U_{i-1}$ for $i = r+1, r, r-1$.

Proof. We proceed inductively, by defining the sequence $U_{r+1}, U_r, \ldots, U_1$. We pick U_{r+1} to be any supplement of N_r in $N_{r+1} = E$, so that

$$E = N_{r+1} = N_r \oplus U_{r+1}.$$

Since $f^{r+1} = 0$ and $N_r = \mathrm{Ker}\,(f^r)$, we have $f(U_{r+1}) \subseteq N_r$, and by Proposition 31.14, as $U_{r+1} \cap N_r = (0)$, we have $f(U_{r+1}) \cap N_{r-1} = (0)$. As a consequence, we can pick a supplement U_r of N_{r-1} in N_r so that $f(U_{r+1}) \subseteq U_r$. We have

$$N_r = N_{r-1} \oplus U_r \quad \text{and} \quad f(U_{r+1}) \subseteq U_r.$$

By Proposition 31.14, f is an injection from U_{r+1} to U_r. Assume inductively that $U_{r+1}, \ldots, U_i$ have been defined for $i \geq 2$ and that they satisfy (1) and (2). Since

$$N_i = N_{i-1} \oplus U_i,$$

we have $U_i \subseteq N_i$, so $f^{i-1}(f(U_i)) = f^i(U_i) = (0)$, which implies that $f(U_i) \subseteq N_{i-1}$. Also, since $U_i \cap N_{i-1} = (0)$, by Proposition 31.14, we have $f(U_i) \cap N_{i-2} = (0)$. It follows that there is a supplement U_{i-1} of N_{i-2} in N_{i-1} that contains $f(U_i)$. We have

$$N_{i-1} = N_{i-2} \oplus U_{i-1} \quad \text{and} \quad f(U_i) \subseteq U_{i-1}.$$

The fact that f is an injection from U_i into U_{i-1} follows from Proposition 31.14. Therefore, the induction step is proven. The construction stops when $i = 1$. $\square$

Because $N_0 = (0)$ and $N_{r+1} = E$, we see that E is the direct sum of the U_i:

$$E = U_1 \oplus \cdots \oplus U_{r+1},$$

with $f(U_i) \subseteq U_{i-1}$, and f an injection from U_i to U_{i-1}, for $i = r+1, \ldots, 2$. By a clever choice of bases in the U_i, we obtain the following nice theorem.

Theorem 31.16. *For any nilpotent linear map $f \colon E \to E$ on a finite-dimensional vector space E of dimension n over a field K, there is a basis of E such that the matrix N of f is of the form*

$$N = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$.

Proof. First apply Proposition 31.15 to obtain a direct sum $E = \bigoplus_{i=1}^{r+1} U_i$. Then we define a basis of E inductively as follows. First we choose a basis

$$e_1^{r+1}, \ldots, e_{n_{r+1}}^{r+1}$$

of U_{r+1}. Next, for $i = r+1, \ldots, 2$, given the basis

$$e_1^i, \ldots, e_{n_i}^i$$

of U_i, since f is injective on U_i and $f(U_i) \subseteq U_{i-1}$, the vectors $f(e_1^i), \ldots, f(e_{n_i}^i)$ are linearly independent, so we define a basis of U_{i-1} by completing $f(e_1^i), \ldots, f(e_{n_i}^i)$ to a basis in U_{i-1}:

$$e_1^{i-1}, \ldots, e_{n_i}^{i-1}, e_{n_i+1}^{i-1}, \ldots, e_{n_{i-1}}^{i-1}$$

with

$$e_j^{i-1} = f(e_j^i), \quad j = 1 \ldots, n_i.$$

Since $U_1 = N_1 = \mathrm{Ker}\,(f)$, we have

$$f(e_j^1) = 0, \quad j = 1, \ldots, n_1.$$

These basis vectors can be arranged as the rows of the following matrix:

$$\begin{pmatrix}
e_1^{r+1} & \cdots & e_{n_{r+1}}^{r+1} & & & & & & & & & \\
\vdots & & \vdots & & & & & & & & & \\
e_1^{r} & \cdots & e_{n_{r+1}}^{r} & e_{n_{r+1}+1}^{r} & \cdots & e_{n_r}^{r} & & & & & & \\
\vdots & & \vdots & \vdots & & \vdots & & & & & & \\
e_1^{r-1} & \cdots & e_{n_{r+1}}^{r-1} & e_{n_{r+1}+1}^{r-1} & \cdots & e_{n_r}^{r-1} & e_{n_r+1}^{r-1} & \cdots & e_{n_{r-1}}^{r-1} & & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & & \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & & \\
e_1^{1} & \cdots & e_{n_{r+1}}^{1} & e_{n_{r+1}+1}^{1} & \cdots & e_{n_r}^{1} & e_{n_r+1}^{1} & \cdots & e_{n_{r-1}}^{1} & \cdots & \cdots & e_{n_1}^{1}
\end{pmatrix}$$

Finally, we define the basis $(e_1, \ldots, e_n)$ by listing each column of the above matrix from the bottom-up, starting with column one, then column two, *etc*. This means that we list the vectors e_j^i in the following order:

For $j = 1, \ldots, n_{r+1}$, list $e_j^1, \ldots, e_j^{r+1}$;

In general, for $i = r, \ldots, 1$,

for $j = n_{i+1} + 1, \ldots, n_i$, list $e_j^1, \ldots, e_j^i$.

Then because $f(e_j^1) = 0$ and $e_j^{i-1} = f(e_j^i)$ for $i \geq 2$, either

$$f(e_i) = 0 \quad \text{or} \quad f(e_i) = e_{i-1},$$

which proves the theorem. □

As an application of Theorem 31.16, we obtain the *Jordan form* of a linear map.

Definition 31.7. A *Jordan block* is an $r \times r$ matrix $J_r(\lambda)$, of the form

$$J_r(\lambda) = \begin{pmatrix}
\lambda & 1 & 0 & \cdots & 0 \\
0 & \lambda & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \ddots & 1 \\
0 & 0 & 0 & \cdots & \lambda
\end{pmatrix},$$

where $\lambda \in K$, with $J_1(\lambda) = (\lambda)$ if $r = 1$. A *Jordan matrix*, J, is an $n \times n$ block diagonal matrix of the form

$$J = \begin{pmatrix}
J_{r_1}(\lambda_1) & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & J_{r_m}(\lambda_m)
\end{pmatrix},$$

where each $J_{r_k}(\lambda_k)$ is a Jordan block associated with some $\lambda_k \in K$, and with $r_1 + \cdots + r_m = n$.

To simplify notation, we often write $J(\lambda)$ for $J_r(\lambda)$. Here is an example of a Jordan matrix with four blocks:

$$J = \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}.$$

Theorem 31.17. *(Jordan form) Let E be a vector space of dimension n over a field K and let $f\colon E \to E$ be a linear map. The following properties are equivalent:*

(1) The eigenvalues of f all belong to K (i.e. the roots of the characteristic polynomial χ_f all belong to K).

(2) There is a basis of E in which the matrix of f is a Jordan matrix.

Proof. Assume (1). First we apply Theorem 31.11, and we get a direct sum $E = \bigoplus_{j=1}^k W_k$, such that the restriction of $g_i = f - \lambda_j \mathrm{id}$ to W_i is nilpotent. By Theorem 31.16, there is a basis of W_i such that the matrix of the restriction of g_i is of the form

$$G_i = \begin{pmatrix} 0 & \nu_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \nu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \nu_{n_i} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

where $\nu_i = 1$ or $\nu_i = 0$. Furthermore, over any basis, $\lambda_i \mathrm{id}$ is represented by the diagonal matrix D_i with λ_i on the diagonal. Then it is clear that we can split $D_i + G_i$ into Jordan blocks by forming a Jordan block for every uninterrupted chain of 1s. By putting the bases of the W_i together, we obtain a matrix in Jordan form for f.

Now assume (2). If f can be represented by a Jordan matrix, it is obvious that the diagonal entries are the eigenvalues of f, so they all belong to K. $\square$

Observe that Theorem 31.17 applies if $K = \mathbb{C}$. It turns out that there are uniqueness properties of the Jordan blocks. There are also other fundamental normal forms for linear maps, such as the rational canonical form, but to prove these results, it is better to develop more powerful machinery about finitely generated modules over a PID. To accomplish this most effectively, we need some basic knowledge about tensor products.

If a complex $n \times n$ matrix A is expressed in terms of its Jordan decomposition as $A = D + N$, since D and N commute, by Proposition 9.21, the exponential of A is given by

$$e^A = e^D e^N,$$

and since N is an $n \times n$ nilpotent matrix, $N^{n-1} = 0$, so we obtain

$$e^A = e^D \left(I + \frac{N}{1!} + \frac{N^2}{2!} + \cdots + \frac{N^{n-1}}{(n-1)!} \right).$$

In particular, the above applies if A is a Jordan matrix. This fact can be used to solve (at least in theory) systems of first-order linear differential equations. Such systems are of the form

$$\frac{dX}{dt} = AX, \tag{$*$}$$

where A is an $n \times n$ matrix and X is an n-dimensional vector of functions of the parameter t.

It can be shown that the columns of the matrix e^{tA} form a basis of the vector space of solutions of the system of linear differential equations $(*)$; see Artin [7] (Chapter 4). Furthermore, for any matrix B and any invertible matrix P, if $A = PBP^{-1}$, then the system $(*)$ is equivalent to

$$P^{-1}\frac{dX}{dt} = BP^{-1}X,$$

so if we make the change of variable $Y = P^{-1}X$, we obtain the system

$$\frac{dY}{dt} = BY. \tag{$**$}$$

Consequently, if B is such that the exponential e^{tB} can be easily computed, we obtain an explicit solution Y of $(**)$, and $X = PY$ is an explicit solution of $(*)$. This is the case when B is a Jordan form of A. In this case, it suffices to consider the Jordan blocks of B. Then we have

$$J_r(\lambda) = \lambda I_r + \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} = \lambda I_r + N,$$

and the powers N^k are easily computed.

For example, if

$$B = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} = 3I_3 + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

we obtain

$$tB = t \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} = 3tI_3 + \begin{pmatrix} 0 & t & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{pmatrix}$$

and so

$$e^{tB} = \begin{pmatrix} e^{3t} & 0 & 0 \\ 0 & e^{3t} & 0 \\ 0 & 0 & e^{3t} \end{pmatrix} \begin{pmatrix} 1 & t & (1/2)t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{3t} & te^{3t} & (1/2)t^2e^{3t} \\ 0 & e^{3t} & te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix}.$$

The columns of e^{tB} form a basis of the space of solutions of the system of linear differential equations

$$\begin{aligned} \frac{dY_1}{dt} &= 3Y_1 + Y_2 \\ \frac{dY_2}{dt} &= 3Y_2 + Y_3 \\ \frac{dY_3}{dt} &= 3Y_3, \end{aligned}$$

in matrix form,

$$\begin{pmatrix} \frac{dY_1}{dt} \\ \frac{dY_2}{dt} \\ \frac{dY_3}{dt} \end{pmatrix} = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}.$$

Explicitly, the general solution of the above system is

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = c_1 \begin{pmatrix} e^{3t} \\ 0 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} te^{3t} \\ e^{3t} \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} (1/2)t^2e^{3t} \\ te^{3t} \\ e^{3t} \end{pmatrix},$$

with $c_1, c_2, c_3 \in \mathbb{R}$. Solving systems of first-order linear differential equations is discussed in Artin [7] and more extensively in Hirsh and Smale [93].

31.7 Summary

The main concepts and results of this chapter are listed below:

- Ideals, principal ideals, greatest common divisors.
- Monic polynomial, irreducible polynomial, relatively prime polynomials.
- Annihilator of a linear map.
- Minimal polynomial of a linear map.

- Invariant subspace.
- f-conductor of u into W; conductor of u into W.
- Diagonalizable linear maps.
- Commuting families of linear maps.
- Primary decomposition.
- Generalized eigenvectors.
- Nilpotent linear map.
- Normal form of a nilpotent linear map.
- Jordan decomposition.
- Jordan block.
- Jordan matrix.
- Jordan normal form.
- Systems of first-order linear differential equations.

31.8 Problems

Problem 31.1. Given a linear map $f\colon E \to E$, prove that the set $\mathrm{Ann}(f)$ of polynomials that annihilate f is an ideal.

Problem 31.2. Provide the details of Proposition 31.3.

Problem 31.3. Prove that the f-conductor $S_f(u, W)$ is an ideal in $K[X]$ (Proposition 31.4).

Problem 31.4. Prove that the polynomials $g_1, \ldots, g_k$ used in the proof of Theorem 31.10 are relatively prime.

Problem 31.5. Find the minimal polynomial of the matrix

$$A = \begin{pmatrix} 6 & -3 & -2 \\ 4 & -1 & -2 \\ 10 & -5 & -3 \end{pmatrix}.$$

Problem 31.6. Find the Jordan decomposition of the matrix

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 2 & 2 & -1 \\ 2 & 2 & 0 \end{pmatrix}.$$

Problem 31.7. Let $f\colon E \to E$ be a linear map on a finite-dimensional vector space. Prove that if f has rank 1, then either f is diagonalizable or f is nilpotent but not both.

Problem 31.8. Find the Jordan form of the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Problem 31.9. Let N be a 3×3 nilpotent matrix over $\mathbb{C}$. Prove that the matrix $A = I + (1/2)N - (1/8)N^2$ satisfies the equation

$$A^2 = I + N.$$

In other words, A is a square root of $I + N$.

Generalize the above fact to any $n \times n$ nilpotent matrix N over $\mathbb{C}$ using the binomial series for $(1+t)^{1/2}$.

Problem 31.10. Let K be an algebraically closed field (for example, $K = \mathbb{C}$). Prove that every 4×4 matrix is similar to a Jordan matrix of the following form:

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}, \quad \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}, \quad \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix},$$
$$\begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \mu & 1 \\ 0 & 0 & 0 & \mu \end{pmatrix}, \quad \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}.$$

Problem 31.11. In this problem the field K is of characteristic 0. Consider an $(r \times r)$ Jordan block

$$J_r(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

Prove that for any polynomial $f(X)$, we have

$$f(J_r(\lambda)) = \begin{pmatrix} f(\lambda) & f_1(\lambda) & f_2(\lambda) & \cdots & f_{r-1}(\lambda) \\ 0 & f(\lambda) & f_1(\lambda) & \cdots & f_{r-2}(\lambda) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & f_1(\lambda) \\ 0 & 0 & 0 & \cdots & f(\lambda) \end{pmatrix},$$

where

$$f_k(X) = \frac{f^{(k)}(X)}{k!},$$

and $f^{(k)}(X)$ is the kth derivative of $f(X)$.

Chapter 32

UFD's, Noetherian Rings, Hilbert's Basis Theorem

32.1 Unique Factorization Domains (Factorial Rings)

We saw in Section 30.5 that if K is a field, then every nonnull polynomial in $K[X]$ can be factored as a product of irreducible factors, and that such a factorization is essentially unique. The same property holds for the ring $K[X_1, \ldots, X_n]$ where $n \geq 2$, but a different proof is needed.

The reason why unique factorization holds for $K[X_1, \ldots, X_n]$ is that if A is an integral domain for which unique factorization holds in some suitable sense, then the property of unique factorization lifts to the polynomial ring $A[X]$. Such rings are called factorial rings, or unique factorization domains. The first step if to define the notion of irreducible element in an integral domain, and then to define a factorial ring. If will turn out that in a factorial ring, any nonnull element a is irreducible (or prime) iff the principal ideal (a) is a prime ideal.

Recall that given a ring A, a *unit* is any invertible element (w.r.t. multiplication). The set of units of A is denoted by A^*. It is a multiplicative subgroup of A, with identity 1. Also, given $a, b \in A$, recall that a *divides* b if $b = ac$ for some $c \in A$; equivalently, a divides b iff $(b) \subseteq (a)$. Any nonzero $a \in A$ is divisible by any unit u, since $a = u(u^{-1}a)$. The relation "a divides b," often denoted by $a \mid b$, is reflexive and transitive, and thus, a preorder on $A - \{0\}$.

Definition 32.1. Let A be an integral domain. Some element $a \in A$ is *irreducible* if $a \neq 0$, $a \notin A^*$ (a is not a unit), and whenever $a = bc$, then either b or c is a unit (where $b, c \in A$). Equivalently, $a \in A$ is *reducible* if $a = 0$, or $a \in A^*$ (a is a unit), or $a = bc$ where $b, c \notin A^*$ (a, b are both noninvertible) and $b, c \neq 0$.

Observe that if $a \in A$ is irreducible and $u \in A$ is a unit, then ua is also irreducible. Generally, if $a \in A$, $a \neq 0$, and u is a unit, then a and ua are said to be *associated*. This is the equivalence relation on nonnull elements of A induced by the divisibility preorder.

The following simple proposition gives a sufficient condition for an element $a \in A$ to be irreducible.

Proposition 32.1. *Let A be an integral domain. For any $a \in A$ with $a \neq 0$, if the principal ideal (a) is a prime ideal, then a is irreducible.*

Proof. If (a) is prime, then $(a) \neq A$ and a is not a unit. Assume that $a = bc$. Then, $bc \in (a)$, and since (a) is prime, either $b \in (a)$ or $c \in (a)$. Consider the case where $b \in (a)$, the other case being similar. Then, $b = ax$ for some $x \in A$. As a consequence,

$$a = bc = axc,$$

and since A is an integral domain and $a \neq 0$, we get

$$1 = xc,$$

which proves that $c = x^{-1}$ is a unit. $\square$

It should be noted that the converse of Proposition 32.1 is generally false. However, it holds for factorial rings, defined next.

Definition 32.2. A *factorial ring* or *unique factorization domain (UFD)* (or *unique factorization ring*) is an integral domain A such that the following two properties hold:

(1) For every nonnull $a \in A$, if $a \notin A^*$ (a is not a unit), then a can be factored as a product

$$a = a_1 \cdots a_m$$

where each $a_i \in A$ is irreducible ($m \geq 1$).

(2) For every nonnull $a \in A$, if $a \notin A^*$ (a is not a unit) and if

$$a = a_1 \cdots a_m = b_1 \cdots b_n$$

where $a_i \in A$ and $b_j \in A$ are irreducible, then $m = n$ and there is a permutation σ of $\{1, \ldots, m\}$ and some units $u_1, \ldots, u_m \in A^*$ such that $a_i = u_i b_{\sigma(i)}$ for all i, $1 \leq i \leq m$.

Example 32.1. The ring $\mathbb{Z}$ of integers if a typical example of a UFD. Given a field K, the polynomial ring $K[X]$ is a UFD. More generally, we will show later that every PID is a UFD (see Theorem 32.12). Thus, in particular, $\mathbb{Z}[X]$ is a UFD. However, we leave as an exercise to prove that the ideal $(2X, X^2)$ generated by $2X$ and X^2 is not principal, and thus, $\mathbb{Z}[X]$ is not a PID.

First, we prove that condition (2) in Definition 32.2 is equivalent to the usual "Euclidean" condition.

There are integral domains that are not UFD's. For example, the subring $\mathbb{Z}[\sqrt{-5}]$ of $\mathbb{C}$ consisting of the complex numbers of the form $a + bi\sqrt{5}$ where $a, b \in \mathbb{Z}$ is not a UFD. Indeed, we have

$$9 = 3 \cdot 3 = (2 + i\sqrt{5})(2 - i\sqrt{5}),$$

and it can be shown that 3, $2 + i\sqrt{5}$, and $2 - i\sqrt{5}$ are irreducible, and that the units are ± 1. The uniqueness condition (2) fails and $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

Remark: For $d \in \mathbb{Z}$ with $d < 0$, it is known that the ring of integers of $\mathbb{Q}(\sqrt{d})$ is a UFD iff d is one of the nine primes, $d = -1, -2, -3, -7, -11, -19, -43, -67$ and -163. This is a hard theorem that was conjectured by Gauss but not proved until 1966, independently by Stark and Baker. Heegner had published a proof of this result in 1952 but there was some doubt about its validity. After finding his proof, Stark reexamined Heegner's proof and concluded that it was essentially correct after all. In sharp contrast, when d is a positive integer, the problem of determining which of the rings of integers of $\mathbb{Q}(\sqrt{d})$ are UFD's, is still open. It can also be shown that if $d < 0$, then the ring $\mathbb{Z}[\sqrt{d}]$ is a UFD iff $d = -1$ or $d = -2$. If $d \equiv 1 \pmod 4$, then $\mathbb{Z}[\sqrt{d}]$ is never a UFD. For more details about these remarkable results, see Stark [164] (Chapter 8).

Proposition 32.2. *Let A be an integral domain satisfying condition* (1) *in Definition 32.2. Then, condition* (2) *in Definition 32.2 is equivalent to the following condition:*

(2′) If $a \in A$ is irreducible and a divides the product bc, where $b, c \in A$ and $b, c \neq 0$, then either a divides b or a divides c.

Proof. First, assume that (2) holds. Let $bc = ad$, where $d \in A$, $d \neq 0$. If b is a unit, then

$$c = adb^{-1},$$

and c is divisible by a. A similar argument applies to c. Thus, we may assume that b and c are not units. In view of (1), we can write

$$b = p_1 \cdots p_m \quad \text{and} \quad c = p_{m+1} \cdots q_{m+n},$$

where $p_i \in A$ is irreducible. Since $bc = ad$, a is irreducible, and b, c are not units, d cannot be a unit. In view of (1), we can write

$$d = q_1 \cdots q_r,$$

where $q_i \in A$ is irreducible. Thus,

$$p_1 \cdots p_m p_{m+1} \cdots p_{m+n} = aq_1 \cdots q_r,$$

where all the factors involved are irreducible. By (2), we must have

$$a = u_{i_0} p_{i_0}$$

for some unit $u_{i_0} \in A$ and some index i_0, $1 \leq i_0 \leq m+n$. As a consequence, if $1 \leq i_0 \leq m$, then a divides b, and if $m+1 \leq i_0 \leq m+n$, then a divides c. This proves that (2′) holds.

Let us now assume that (2′) holds. Assume that

$$a = a_1 \cdots a_m = b_1 \cdots b_n,$$

where $a_i \in A$ and $b_j \in A$ are irreducible. Without loss of generality, we may assume that $m \leq n$. We proceed by induction on m. If $m = 1$,

$$a_1 = b_1 \cdots b_n,$$

and since a_1 is irreducible, $u = b_1 \cdots b_{i-1} b_{i+1} b_n$ must be a unit for some i, $1 \leq i \leq n$. Thus, (2) holds with $n = 1$ and $a_1 = b_i u$. Assume that $m > 1$ and that the induction hypothesis holds for $m - 1$. Since

$$a_1 a_2 \cdots a_m = b_1 \cdots b_n,$$

a_1 divides $b_1 \cdots b_n$, and in view of (2′), a_1 divides some b_j. Since a_1 and b_j are irreducible, we must have $b_j = u_j a_1$, where $u_j \in A$ is a unit. Since A is an integral domain,

$$a_1 a_2 \cdots a_m = b_1 \cdots b_{j-1} u_j a_1 b_{j+1} \cdots b_n$$

implies that

$$a_2 \cdots a_m = (u_j b_1) \cdots b_{j-1} b_{j+1} \cdots b_n,$$

and by the induction hypothesis, $m - 1 = n - 1$ and $a_i = v_i b_{\tau(i)}$ for some units $v_i \in A$ and some bijection τ between $\{2, \ldots, m\}$ and $\{1, \ldots, j-1, j+1, \ldots, m\}$. However, the bijection τ extends to a permutation σ of $\{1, \ldots, m\}$ by letting $\sigma(1) = j$, and the result holds by letting $v_1 = u_j^{-1}$. $\square$

As a corollary of Proposition 32.2. we get the converse of Proposition 32.1.

Proposition 32.3. *Let A be a factorial ring. For any $a \in A$ with $a \neq 0$, the principal ideal (a) is a prime ideal iff a is irreducible.*

Proof. In view of Proposition 32.1, we just have to prove that if $a \in A$ is irreducible, then the principal ideal (a) is a prime ideal. Indeed, if $bc \in (a)$, then a divides bc, and by Proposition 32.2, property (2′) implies that either a divides b or a divides c, that is, either $b \in (a)$ or $c \in (a)$, which means that (a) is prime. $\square$

Because Proposition 32.3 holds, in a UFD, an irreducible element is often called a *prime*.

In a UFD A, every nonzero element $a \in A$ that is not a unit can be expressed as a product $a = a_1 \cdots a_n$ of irreducible elements a_i, and by property (2), the number n of factors only depends on a, that is, it is the same for all factorizations into irreducible factors. We agree that this number is 0 for a unit.

Remark: If A is a UFD, we can state the factorization properties so that they also applies to units:

(1) For every nonnull $a \in A$, a can be factored as a product

$$a = ua_1 \cdots a_m$$

where $u \in A^*$ (u is a unit) and each $a_i \in A$ is irreducible ($m \geq 0$).

(2) For every nonnull $a \in A$, if

$$a = ua_1 \cdots a_m = vb_1 \cdots b_n$$

where $u, v \in A^*$ (u, v are units) and $a_i \in A$ and $b_j \in A$ are irreducible, then $m = n$, and if $m = n = 0$ then $u = v$, else if $m \geq 1$, then there is a permutation σ of $\{1, \ldots, m\}$ and some units $u_1, \ldots, u_m \in A^*$ such that $a_i = u_i b_{\sigma(i)}$ for all i, $1 \leq i \leq m$.

We are now ready to prove that if A is a UFD, then the polynomial ring $A[X]$ is also a UFD.

First, observe that the units of $A[X]$ are just the units of A. The fact that nonnull and nonunit polynomials in $A[X]$ factor as products of irreducible polynomials is easier to prove than uniqueness. We will show in the proof of Theorem 32.10 that we can proceed by induction on the pairs (m, n) where m is the degree of $f(X)$ and n is either 0 if the coefficient f_m of X^m in $f(X)$ is a unit of n is f_m is the product of n irreducible elements.

For the uniqueness of the factorization, by Proposition 32.2, it is enough to prove that condition (2′) holds. This is a little more tricky. There are several proofs, but they all involve a pretty Lemma due to Gauss.

First, note the following trivial fact. Given a ring A, for any $a \in A$, $a \neq 0$, if a divides every coefficient of some nonnull polynomial $f(X) \in A[X]$, then a divides $f(X)$. If A is an integral domain, we get the following converse.

Proposition 32.4. *Let A be an integral domain. For any $a \in A$, $a \neq 0$, if a divides a nonnull polynomial $f(X) \in A[X]$, then a divides every coefficient of $f(X)$.*

Proof. Assume that $f(X) = ag(X)$, for some $g(X) \in A[X]$. Since $a \neq 0$ and A is an integral ring, $f(X)$ and $g(X)$ have the same degree m, and since for every i ($0 \leq i \leq m$) the coefficient of X^i in $f(X)$ is equal to the coefficient of X^i in $ag(x)$, we have $f_i = ag_i$, and whenever $f_i \neq 0$, we see that a divides f_i. □

Lemma 32.5. *(Gauss's lemma) Let A be a UFD. For any $a \in A$, if a is irreducible and a divides the product $f(X)g(X)$ of two polynomials $f(X), g(X) \in A[X]$, then either a divides $f(X)$ or a divides $g(X)$.*

Proof. Let $f(X) = f_m X^m + \cdots + f_i X^i + \cdots + f_0$ and $g(X) = g_n X^n + \cdots + g_j X^j + \cdots + g_0$. Assume that a divides neither $f(X)$ nor $g(X)$. By the (easy) converse of Proposition 32.4, there is some i ($0 \leq i \leq m$) such that a does not divide f_i, and there is some j ($0 \leq j \leq n$)

such that a does not divide g_j. Pick i and j minimal such that a does not divide f_i and a does not divide g_j. The coefficient c_{i+j} of X^{i+j} in $f(X)g(X)$ is

$$c_{i+j} = f_0 g_{i+j} + f_1 g_{i+j-1} + \cdots + f_i g_j + \cdots + f_{i+j} g_0$$

(letting $f_h = 0$ if $h > m$ and $g_k = 0$ if $k > n$). From the choice of i and j, a cannot divide $f_i g_j$, since a being irreducible, by (2′) of Proposition 32.2, a would divide f_i or g_j. However, by the choice of i and j, a divides every other nonnull term in the sum for c_{i+j}, and since a is irreducible and divides $f(X)g(X)$, by Proposition 32.4, a divides c_{i+j}, which implies that a divides $f_i g_j$, a contradiction. Thus, either a divides $f(X)$ or a divides $g(X)$. ☐

As a corollary, we get the following proposition.

Proposition 32.6. *Let A be a UFD. For any $a \in A$, $a \neq 0$, if a divides the product $f(X)g(X)$ of two polynomials $f(X), g(X) \in A[X]$ and $f(X)$ is irreducible and of degree at least 1, then a divides $g(X)$.*

Proof. The Proposition is trivial is a is a unit. Otherwise, $a = a_1 \cdots a_m$ where $a_i \in A$ is irreducible. Using induction and applying Lemma 32.5, we conclude that a divides $g(X)$. ☐

We now show that Lemma 32.5 also applies to the case where a is an irreducible polynomial. This requires a little excursion involving the fraction field F of A.

Remark: If A is a UFD, it is possible to prove the uniqueness condition (2) for $A[X]$ directly without using the fraction field of A, see Malliavin [119], Chapter 3.

Given an integral domain A, we can construct a field F such that every element of F is of the form a/b, where $a, b \in A$, $b \neq 0$, using essentially the method for constructing the field $\mathbb{Q}$ of rational numbers from the ring $\mathbb{Z}$ of integers.

Proposition 32.7. *Let A be an integral domain.*

(1) There is a field F and an injective ring homomorphism $i\colon A \to F$ such that every element of F is of the form $i(a)i(b)^{-1}$, where $a, b \in A$, $b \neq 0$.

(2) For every field K and every injective ring homomorphism $h\colon A \to K$, there is a (unique) field homomorphism $\widehat{h}\colon F \to K$ such that

$$\widehat{h}(i(a)i(b)^{-1}) = h(a)h(b)^{-1}$$

for all $a, b \in A$, $b \neq 0$.

(3) The field F in (1) is unique up to isomorphism.

Proof. (1) Consider the binary relation $\simeq$ on $A \times (A - \{0\})$ defined as follows:

$$(a, b) \simeq (a', b') \quad \text{iff} \quad ab' = a'b.$$

It is easily seen that $\simeq$ is an equivalence relation. Note that the fact that A is an integral domain is used to prove transitivity. The equivalence class of (a, b) is denoted by a/b. Clearly, $(0, b) \simeq (0, 1)$ for all $b \in A$, and we denote the class of $(0, 1)$ also by 0. The equivalence class $a/1$ of $(a, 1)$ is also denoted by a. We define addition and multiplication on $A \times (A - \{0\})$ as follows:

$$\begin{aligned}(a, b) + (a', b') &= (ab' + a'b, bb'),\\ (a, b) \cdot (a', b') &= (aa', bb').\end{aligned}$$

It is easily verified that $\simeq$ is congruential w.r.t. $+$ and $\cdot$, which means that $+$ and $\cdot$ are well-defined on equivalence classes modulo $\simeq$. When $a, b \neq 0$, the inverse of a/b is b/a, and it is easily verified that F is a field. The map $i\colon A \to F$ defined such that $i(a) = a/1$ is an injection of A into F and clearly

$$\frac{a}{b} = i(a)i(b)^{-1}.$$

(2) Given an injective ring homomorphism $h\colon A \to K$ into a field K,

$$\frac{a}{b} = \frac{a'}{b'} \quad \text{iff} \quad ab' = a'b,$$

which implies that

$$h(a)h(b') = h(a')h(b),$$

and since h is injective and $b, b' \neq 0$, we get

$$h(a)h(b)^{-1} = h(a')h(b')^{-1}.$$

Thus, there is a map $\widehat{h}\colon F \to K$ such that

$$\widehat{h}(a/b) = \widehat{h}(i(a)i(b)^{-1}) = h(a)h(b)^{-1}$$

for all $a, b \in A$, $b \neq 0$, and it is easily checked that $\widehat{h}$ is a field homomorphism. The map $\widehat{h}$ is clearly unique.

(3) The uniqueness of F up to isomorphism follows from (2), and is left as an exercise. $\square$

The field F given by Proposition 32.7 is called the *fraction field of* A, and it is denoted by $\mathrm{Frac}(A)$.

In particular, given an integral domain A, since $A[X_1, \ldots, X_n]$ is also an integral domain, we can form the fraction field of the polynomial ring $A[X_1, \ldots, X_n]$, denoted by $F(X_1, \ldots, X_n)$, where $F = \mathrm{Frac}(A)$ is the fraction field of A. It is also called the field of *rational functions* over F, although the terminology is a bit misleading, since elements of $F(X_1, \ldots, X_n)$ only define functions when the dominator is nonnull.

We now have the following crucial lemma which shows that if a polynomial $f(X)$ is reducible over $F[X]$ where F is the fraction field of A, then $f(X)$ is already reducible over $A[X]$.

Lemma 32.8. *Let A be a UFD and let F be the fraction field of A. For any nonnull polynomial $f(X) \in A[X]$ of degree m, if $f(X)$ is not the product of two polynomials of degree strictly smaller than m, then $f(X)$ is irreducible in $F[X]$.*

Proof. Assume that $f(X)$ is reducible in $F[X]$ and that $f(X)$ is neither null nor a unit. Then,

$$f(X) = G(X)H(X),$$

where $G(X), H(X) \in F[X]$ are polynomials of degree $p, q \geq 1$. Let a be the product of the denominators of the coefficients of $G(X)$, and b the product of the denominators of the coefficients of $H(X)$. Then, $a, b \neq 0$, $g_1(X) = aG(X) \in A[X]$ has degree $p \geq 1$, $h_1(X) = bH(X) \in A[X]$ has degree $q \geq 1$, and

$$abf(X) = g_1(X)h_1(X).$$

Let $c = ab$. If c is a unit, then $f(X)$ is also reducible in $A[X]$. Otherwise, $c = c_1 \cdots c_n$, where $c_i \in A$ is irreducible. We now use induction on n to prove that

$$f(X) = g(X)h(X),$$

for some polynomials $g(X) \in A[X]$ of degree $p \geq 1$ and $h(X) \in A[X]$ of degree $q \geq 1$.

If $n = 1$, since $c = c_1$ is irreducible, by Lemma 32.5, either c divides $g_1(X)$ or c divides $h_1(X)$. Say that c divides $g_1(X)$, the other case being similar. Then, $g_1(X) = cg(X)$ for some $g(X) \in A[X]$ of degree $p \geq 1$, and since $A[X]$ is an integral ring, we get

$$f(X) = g(X)h_1(X),$$

showing that $f(X)$ is reducible in $A[X]$. If $n > 1$, since

$$c_1 \cdots c_n f(X) = g_1(X)h_1(X),$$

c_1 divides $g_1(X)h_1(X)$, and as above, either c_1 divides $g_1(X)$ or c divides $h_1(X)$. In either case, we get

$$c_2 \cdots c_n f(X) = g_2(X)h_2(X)$$

for some polynomials $g_2(X) \in A[X]$ of degree $p \geq 1$ and $h_2(X) \in A[X]$ of degree $q \geq 1$. By the induction hypothesis, we get

$$f(X) = g(X)h(X),$$

for some polynomials $g(X) \in A[X]$ of degree $p \geq 1$ and $h(X) \in A[X]$ of degree $q \geq 1$, showing that $f(X)$ is reducible in $A[X]$. ☐

Finally, we can prove that (2′) holds.

Lemma 32.9. *Let A be a UFD. Given any three nonnull polynomials $f(X), g(X), h(X) \in A[X]$, if $f(X)$ is irreducible and $f(X)$ divides the product $g(X)h(X)$, then either $f(X)$ divides $g(X)$ or $f(X)$ divides $h(X)$.*

Proof. If $f(X)$ has degree 0, then the result follows from Lemma 32.5. Thus, we may assume that the degree of $f(X)$ is $m \geq 1$. Let F be the fraction field of A. By Lemma 32.8, $f(X)$ is also irreducible in $F[X]$. Since $F[X]$ is a UFD (by Theorem 30.17), either $f(X)$ divides $g(X)$ or $f(X)$ divides $h(X)$, in $F[X]$. Assume that $f(X)$ divides $g(X)$, the other case being similar. Then,

$$g(X) = f(X)G(X),$$

for some $G(X) \in F[X]$. If a is the product the denominators of the coefficients of G, we have

$$ag(X) = q_1(X)f(X),$$

where $q_1(X) = aG(X) \in A[X]$. If a is a unit, we see that $f(X)$ divides $g(X)$. Otherwise, $a = a_1 \cdots a_n$, where $a_i \in A$ is irreducible. We prove by induction on n that

$$g(X) = q(X)f(X)$$

for some $q(X) \in A[X]$.

If $n = 1$, since $f(X)$ is irreducible and of degree $m \geq 1$ and

$$a_1g(X) = q_1(X)f(X),$$

by Lemma 32.5, a_1 divides $q_1(X)$. Thus, $q_1(X) = a_1q(X)$ where $q(X) \in A[X]$. Since $A[X]$ is an integral domain, we get

$$g(X) = q(X)f(X),$$

and $f(X)$ divides $g(X)$. If $n > 1$, from

$$a_1 \cdots a_n g(X) = q_1(X)f(X),$$

we note that a_1 divides $q_1(X)f(X)$, and as in the previous case, a_1 divides $q_1(X)$. Thus, $q_1(X) = a_1q_2(X)$ where $q_2(X) \in A[X]$, and we get

$$a_2 \cdots a_n g(X) = q_2(X)f(X).$$

By the induction hypothesis, we get

$$g(X) = q(X)f(X)$$

for some $q(X) \in A[X]$, and $f(X)$ divides $g(X)$. ∎

We finally obtain the fact that $A[X]$ is a UFD when A is.

Theorem 32.10. *If A is a UFD then the polynomial ring $A[X]$ is also a UFD.*

Proof. As we said earlier, the factorization property (1) is easier to prove than uniqueness. Assume that $f(X)$ has degree m and let f_m be the coefficient of X^m in $f(X)$. Either f_m is a unit or it is the product of $n \geq 1$ irreducible elements. If f_m is a unit we set $n = 0$. We proceed by induction on the pair (m, n), using the well-founded ordering on pairs, i.e.,

$$(m, n) \leq (m', n')$$

iff either $m < m'$, or $m = m'$ and $n < n'$. If $f(X)$ is a nonnull polynomial of degree 0 which is not a unit, then $f(X) \in A$, and $f(X) = f_m = a_1 \cdots a_n$ for some irreducible $a_i \in A$, since A is a UFD. This proves the base case.

If $f(X)$ has degree $m > 0$ and $f(X)$ is reducible, then

$$f(X) = g(X)h(X),$$

where $g(X)$ and $h(X)$ have degree $p, q \leq m$ and are not units. There are two cases.

(1) f_m is a unit (so $n = 0$).

If so, since $f_m = g_p h_q$ (where g_p is the coefficient of X^p in $g(X)$ and h_q is the coefficient of X^q in $h(X)$), then g_p and h_q are both units. We claim that $p, q \geq 1$. Otherwise, $p = 0$ or $q = 0$, but then either $g(X) = g_0$ is a unit or $h(X) = h_0$ is a unit, a contradiction.

Now, since $m = p + q$ and $p, q \geq 1$, we have $p, q < m$ so $(p, 0) < (m, 0)$ and $(q, 0) < (m, 0)$, and by the induction hypothesis, both $g(X)$ and $h(X)$ can be written as products of irreducible factors, thus so can $f(X)$.

(2) f_m is not a unit, say $f_m = a_1 \cdots a_n$ where $a_1, \ldots, a_n$ are irreducible and $n \geq 1$.

(a) If $p, q < m$, then $(p, n_1) < (m, n)$ and $(q, n_2) < (m, n)$ where n_1 is the number of irreducible factors of g_p or $n_1 = 0$ if g_p is irreducible, and similarly n_2 is the number of irreducible factors of h_p or $n_2 = 0$ if h_p is irreducible (note that $n_1, n_2 \leq n$ and it is possible that $n_1 = n$ if h_q is irreducible or $n_2 = n$ if g_p is irreducible). By the induction hypothesis, $g(X)$ and $h(X)$ can be written as products of irreducible polynomials, thus so can $f(X)$.

(b) If $p = 0$ and $q = m$, then $g(X) = g_p$ and by hypothesis g_p is not a unit. Since $f_m = a_1 \cdots a_n = g_p h_q$ and g_p is not a unit, either h_q is not a unit in which case, by the uniqueness of the number of irreducible elements in the decomposition of f_m (since A is a UFD), h_q is the product of $n_2 < n$ irreducible elements, or $n_2 = 0$ if h_q is irreducible. Since $n \geq 1$, this implies that $(m, n_2) < (m, n)$, and by the induction hypothesis $h(X)$ can be written as products of irreducible polynomials. Since $g_p \in A$ is not a unit, it can also be written as a product of irreducible elements, thus so can $f(X)$.

The case where $p = m$ and $q = 0$ is similar to the previous case.

Property $(2')$ follows by Lemma 32.9. By Proposition 32.2, $A[X]$ is a UFD. $\square$

As a corollary of Theorem 32.10 and using induction, we note that for any field K, the polynomial ring $K[X_1, \ldots, X_n]$ is a UFD.

For the sake of completeness, we shall prove that every PID is a UFD. First, we review the notion of gcd and the characterization of gcd's in a PID.

Given an integral domain A, for any two elements $a, b \in A$, $a, b \neq 0$, we say that $d \in A$ ($d \neq 0$) is a *greatest common divisor (gcd)* of a and b if

(1) d divides both a and b.

(2) For any $h \in A$ ($h \neq 0$), if h divides both a and b, then h divides d.

We also say that a and b are *relatively prime* if 1 is a gcd of a and b.

Note that a and b are relatively prime iff every gcd of a and b is a unit. If A is a PID, then gcd's are characterized as follows.

Proposition 32.11. *Let A be a PID.*

(1) For any $a, b, d \in A$ $(a, b, d \neq 0)$, d is a gcd of a and b iff

$$(d) = (a, b) = (a) + (b),$$

i.e., d generates the principal ideal generated by a and b.

(2) (Bezout identity) Two nonnull elements $a, b \in A$ are relatively prime iff there are some $x, y \in A$ such that

$$ax + by = 1.$$

Proof. (1) Recall that the ideal generated by a and b is the set

$$(a) + (b) = aA + bA = \{ax + by \mid x, y \in A\}.$$

First, assume that d is a gcd of a and b. If so, $a \in Ad$, $b \in Ad$, and thus, $(a) \subseteq (d)$ and $(b) \subseteq (d)$, so that

$$(a) + (b) \subseteq (d).$$

Since A is a PID, there is some $t \in A$, $t \neq 0$, such that

$$(a) + (b) = (t),$$

and thus, $(a) \subseteq (t)$ and $(b) \subseteq (t)$, which means that t divides both a and b. Since d is a gcd of a and b, t must divide d. But then,

$$(d) \subseteq (t) = (a) + (b),$$

and thus, $(d) = (a) + (b)$.

Assume now that

$$(d) = (a) + (b) = (a, b).$$

Since $(a) \subseteq (d)$ and $(b) \subseteq (d)$, d divides both a and b. Assume that t divides both a and b, so that $(a) \subseteq (t)$ and $(b) \subseteq (t)$. Then,

$$(d) = (a) + (b) \subseteq (t),$$

which means that t divides d, and d is indeed a gcd of a and b.

(2) By (1), if a and b are relatively prime, then

$$(1) = (a) + (b),$$

which yields the result. Conversely, if

$$ax + by = 1,$$

then

$$(1) = (a) + (b),$$

and 1 is a gcd of a and b. $\square$

Given two nonnull elements $a, b \in A$, if a is an irreducible element and a does not divide b, then a and b are relatively prime. Indeed, if d is not a unit and d divides both a and b, then $a = dp$ and $b = dq$ where p must be a unit, so that

$$b = ap^{-1}q,$$

and a divides b, a contradiction.

Theorem 32.12. *Let A be ring. If A is a PID, then A is a UFD.*

Proof. First, we prove that every nonnull element that is a not a unit can be factored as a product of irreducible elements. Let $\mathcal{S}$ be the set of nontrivial principal ideals (a) such that $a \neq 0$ is not a unit and cannot be factored as a product of irreducible elements (in particular, a is not irreducible). Assume that $\mathcal{S}$ is nonempty. We claim that every ascending chain in $\mathcal{S}$ is finite. Otherwise, consider an infinite ascending chain

$$(a_1) \subset (a_2) \subset \cdots \subset (a_n) \subset \cdots .$$

It is immediately verified that

$$\bigcup_{n \geq 1} (a_n)$$

is an ideal in A. Since A is a PID,

$$\bigcup_{n \geq 1} (a_n) = (a)$$

for some $a \in A$. However, there must be some n such that $a \in (a_n)$, and thus,

$$(a_n) \subseteq (a) \subseteq (a_n),$$

and the chain stabilizes at (a_n).

As a consequence, there are maximal ideals in $\mathcal{S}$. Let (a) be a maximal ideal in $\mathcal{S}$. Then, for any ideal (d) such that

$$(a) \subset (d) \quad \text{and} \quad (a) \neq (d),$$

we must have $d \notin \mathcal{S}$, since otherwise (a) would not be a maximal ideal in $\mathcal{S}$. Observe that a is not irreducible, since $(a) \in \mathcal{S}$, and thus,

$$a = bc$$

for some $b, c \in A$, where neither b nor c is a unit. Then,

$$(a) \subseteq (b) \quad \text{and} \quad (a) \subseteq (c).$$

If $(a) = (b)$, then $b = au$ for some $u \in A$, and then

$$a = auc,$$

so that

$$1 = uc,$$

since A is an integral domain, and thus, c is a unit, a contradiction. Thus, $(a) \neq (b)$, and similarly, $(a) \neq (c)$. But then, by a previous observation $b \notin \mathcal{S}$ and $c \notin \mathcal{S}$, and since a and b are not units, both b and c factor as products of irreducible elements and so does $a = bc$, a contradiction. This implies that $\mathcal{S} = \emptyset$, so every nonnull element that is a not a unit can be factored as a product of irreducible elements

To prove the uniqueness of factorizations, we use Proposition 32.2. Assume that a is irreducible and that a divides bc. If a does not divide b, by a previous remark, a and b are relatively prime, and by Proposition 32.11, there are some $x, y \in A$ such that

$$ax + by = 1.$$

Thus,

$$acx + bcy = c,$$

and since a divides bc, we see that a must divide c, as desired. $\square$

Thus, we get another justification of the fact that $\mathbb{Z}$ is a UFD and that if K is a field, then $K[X]$ is a UFD.

It should also be noted that in a UFD, gcd's of nonnull elements always exist. Indeed, this is trivial if a or b is a unit, and otherwise, we can write

$$a = p_1 \cdots p_m \quad \text{and} \quad b = q_1 \cdots q_n$$

where $p_i, q_j \in A$ are irreducible, and the product of the common factors of a and b is a gcd of a and b (it is 1 is there are no common factors).

We conclude this section on UFD's by proving a proposition characterizing when a UFD is a PID. The proof is nontrivial and makes use of Zorn's lemma (several times).

Proposition 32.13. *Let A be a ring that is a UFD, and not a field. Then, A is a PID iff every nonzero prime ideal is maximal.*

Proof. Assume that A is a PID that is not a field. Consider any nonzero prime ideal, (p), and pick any proper ideal $\mathfrak{A}$ in A such that

$$(p) \subseteq \mathfrak{A}.$$

Since A is a PID, the ideal $\mathfrak{A}$ is a principal ideal, so $\mathfrak{A} = (q)$, and since $\mathfrak{A}$ is a proper nonzero ideal, $q \neq 0$ and q is not a unit. Since

$$(p) \subseteq (q),$$

q divides p, and we have $p = qp_1$ for some $p_1 \in A$. Now, by Proposition 32.1, since $p \neq 0$ and (p) is a prime ideal, p is irreducible. But then, since $p = qp_1$ and p is irreducible, p_1 must be a unit (since q is not a unit), which implies that

$$(p) = (q);$$

that is, (p) is a maximal ideal.

Conversely, let us assume that every nonzero prime ideal is maximal. First, we prove that every prime ideal is principal. This is obvious for (0). If $\mathfrak{A}$ is a nonzero prime ideal, then, by hypothesis, it is maximal. Since $\mathfrak{A} \neq (0)$, there is some nonzero element $a \in \mathfrak{A}$. Since $\mathfrak{A}$ is maximal, a is not a unit, and since A is a UFD, there is a factorization $a = a_1 \cdots a_n$ of a into irreducible elements. Since $\mathfrak{A}$ is prime, we have $a_i \in \mathfrak{A}$ for some i. Now, by Proposition 32.3, since a_i is irreducible, the ideal (a_i) is prime, and so, by hypothesis, (a_i) is maximal. Since $(a_i) \subseteq \mathfrak{A}$ and (a_i) is maximal, we get $\mathfrak{A} = (a_i)$.

Next, assume that A is not a PID. Define the set, $\mathcal{F}$, by

$$\mathcal{F} = \{\mathfrak{A} \mid \mathfrak{A} \subseteq A, \quad \mathfrak{A} \text{ is not a principal ideal}\}.$$

Since A is not a PID, the set $\mathcal{F}$ is nonempty. Also, the reader will easily check that every chain in $\mathcal{F}$ is bounded in $\mathcal{F}$. Indeed, for any chain $(\mathfrak{A}_i)_{i\in I}$ of ideals in $\mathcal{F}$ it is not hard to verify that $\bigcup_{i\in I} \mathfrak{A}_i$ is an ideal which is not principal, so $\bigcup_{i\in I} \mathfrak{A}_i \in \mathcal{F}$. Then, by Zorn's lemma (Lemma C.1), the set $\mathcal{F}$ has some maximal element, $\mathfrak{A}$. Clearly, $\mathfrak{A} \neq (0)$ is a proper ideal (since $A = (1)$), and $\mathfrak{A}$ is not prime, since we just showed that prime ideals are principal. Then, by Theorem C.3, there is some maximal ideal, $\mathfrak{M}$, so that $\mathfrak{A} \subset \mathfrak{M}$. However, a maximal ideal is prime, and we have shown that a prime ideal is principal. Thus,

$$\mathfrak{A} \subseteq (p),$$

for some $p \in A$ that is not a unit. Moreover, by Proposition 32.1, the element p is irreducible. Define

$$\mathfrak{B} = \{a \in A \mid pa \in \mathfrak{A}\}.$$

Clearly, $\mathfrak{A} = p\mathfrak{B}$, $\mathfrak{B} \neq (0)$, $\mathfrak{A} \subseteq \mathfrak{B}$, and $\mathfrak{B}$ is a proper ideal. We claim that $\mathfrak{A} \neq \mathfrak{B}$. Indeed, if $\mathfrak{A} = \mathfrak{B}$ were true, then we would have $\mathfrak{A} = p\mathfrak{B} = \mathfrak{B}$, but this is impossible since p is irreducible, A is a UFD, and $\mathfrak{B} \neq (0)$ (we get $\mathfrak{B} = p^m\mathfrak{B}$ for all m, and every element of $\mathfrak{B}$ would be a multiple of p^m for arbitrarily large m, contradicting the fact that A is a UFD). Thus, we have $\mathfrak{A} \subset \mathfrak{B}$, and since $\mathfrak{A}$ is a maximal element of $\mathcal{F}$, we must have $\mathfrak{B} \notin \mathcal{F}$. However, $\mathfrak{B} \notin \mathcal{F}$ means that $\mathfrak{B}$ is a principal ideal, and thus, $\mathfrak{A} = p\mathfrak{B}$ is also a principal ideal, a contradiction. $\square$

Observe that the above proof shows that Proposition 32.13 also holds under the assumption that every prime ideal is principal.

32.2 The Chinese Remainder Theorem

In this section, which is a bit of an interlude, we prove a basic result about quotients of commutative rings by products of ideals that are pairwise relatively prime. This result has applications in number theory and in the structure theorem for finitely generated modules over a PID, which will be presented later.

Given two ideals $\mathfrak{a}$ and $\mathfrak{b}$ of a ring A, we define the ideal $\mathfrak{ab}$ as the set of all finite sums of the form

$$a_1b_1 + \cdots + a_kb_k, \quad a_i \in \mathfrak{a},\ b_i \in \mathfrak{b}.$$

The reader should check that $\mathfrak{ab}$ is indeed an ideal. Observe that $\mathfrak{ab} \subseteq \mathfrak{a}$ and $\mathfrak{ab} \subseteq \mathfrak{b}$, so that

$$\mathfrak{ab} \subseteq \mathfrak{a} \cap \mathfrak{b}.$$

In general equality does not hold. However if

$$\mathfrak{a} + \mathfrak{b} = A,$$

then we have

$$\mathfrak{ab} = \mathfrak{a} \cap \mathfrak{b}.$$

This is because there is some $a \in \mathfrak{a}$ and some $b \in \mathfrak{b}$ such that

$$a + b = 1,$$

so for every $x \in \mathfrak{a} \cap \mathfrak{b}$, we have

$$x = xa + xb,$$

which shows that $x \in \mathfrak{ab}$. Ideals $\mathfrak{a}$ and $\mathfrak{b}$ of A that satisfy the condition $\mathfrak{a} + \mathfrak{b} = A$ are sometimes said to be *comaximal*.

We define the homomorphism $\varphi\colon A \to A/\mathfrak{a} \times A/\mathfrak{b}$ by

$$\varphi(x) = (\overline{x}_{\mathfrak{a}}, \overline{x}_{\mathfrak{b}}),$$

where $\overline{x}_{\mathfrak{a}}$ is the equivalence class of x modulo $\mathfrak{a}$ (resp. $\overline{x}_{\mathfrak{b}}$ is the equivalence class of x modulo $\mathfrak{b}$). Recall that the ideal $\mathfrak{a}$ defines the equivalence relation $\equiv_{\mathfrak{a}}$ on A given by

$$x \equiv_{\mathfrak{a}} y \quad \text{iff} \quad x - y \in \mathfrak{a},$$

and that $A/\mathfrak{a}$ is the quotient ring of equivalence classes $\overline{x}_{\mathfrak{a}}$, where $x \in A$, and similarly for $A/\mathfrak{b}$. Sometimes, we also write $x \equiv y \pmod{\mathfrak{a}}$ for $x \equiv_{\mathfrak{a}} y$.

Clearly, the kernel of the homomorphism φ is $\mathfrak{a} \cap \mathfrak{b}$. If we assume that $\mathfrak{a} + \mathfrak{b} = A$, then $\mathrm{Ker}\,(\varphi) = \mathfrak{a} \cap \mathfrak{b} = \mathfrak{a}\mathfrak{b}$, and because φ has a constant value on the equivalence classes modulo $\mathfrak{a}\mathfrak{b}$, the map φ induces a quotient homomorphism

$$\theta\colon A/\mathfrak{a}\mathfrak{b} \to A/\mathfrak{a} \times A/\mathfrak{b}.$$

Because $\mathrm{Ker}\,(\varphi) = \mathfrak{a}\mathfrak{b}$, the homomorphism θ is injective. The Chinese Remainder Theorem says that θ is an isomorphism.

Theorem 32.14. *Given a commutative ring A, let $\mathfrak{a}$ and $\mathfrak{b}$ be any two ideals of A such that $\mathfrak{a} + \mathfrak{b} = A$. Then, the homomorphism $\theta\colon A/\mathfrak{a}\mathfrak{b} \to A/\mathfrak{a} \times A/\mathfrak{b}$ is an isomorphism.*

Proof. We already showed that θ is injective, so we need to prove that θ is surjective. We need to prove that for any $y, z \in A$, there is some $x \in A$ such that

$$\begin{aligned} x &\equiv y \pmod{\mathfrak{a}} \\ x &\equiv z \pmod{\mathfrak{b}}. \end{aligned}$$

Since $\mathfrak{a} + \mathfrak{b} = A$, there exist some $a \in \mathfrak{a}$ and some $b \in \mathfrak{b}$ such that

$$a + b = 1.$$

If we let

$$x = az + by,$$

then we have

$$x \equiv_{\mathfrak{a}} by \equiv_{\mathfrak{a}} (1 - a)y \equiv_{\mathfrak{a}} y - ay \equiv_{\mathfrak{a}} y,$$

and similarly

$$x \equiv_{\mathfrak{b}} az \equiv_{\mathfrak{b}} (1 - b)z \equiv_{\mathfrak{b}} z - bz \equiv_{\mathfrak{b}} z,$$

which shows that $x = az + by$ works. $\square$

Theorem 32.14 can be generalized to any (finite) number of ideals.

Theorem 32.15. *(Chinese Remainder Theorem) Given a commutative ring A, let $\mathfrak{a}_1, \ldots, \mathfrak{a}_n$ be any $n \geq 2$ ideals of A such that $\mathfrak{a}_i + \mathfrak{a}_j = A$ for all $i \neq j$. Then, the homomorphism $\theta\colon A/\mathfrak{a}_1 \cdots \mathfrak{a}_n \to A/\mathfrak{a}_1 \times \cdots \times A/\mathfrak{a}_n$ is an isomorphism.*

Proof. The map $\theta\colon A/\mathfrak{a}_1 \cap \cdots \cap \mathfrak{a}_n \to A/\mathfrak{a}_1 \times \cdots \times A/\mathfrak{a}_n$ is induced by the homomorphism $\varphi\colon A \to A/\mathfrak{a}_1 \times \cdots \times A/\mathfrak{a}_n$ given by

$$\varphi(x) = (\overline{x}_{\mathfrak{a}_1}, \ldots, \overline{x}_{\mathfrak{a}_n}).$$

Clearly, $\mathrm{Ker}\,(\varphi) = \mathfrak{a}_1 \cap \cdots \cap \mathfrak{a}_n$, so θ is well-defined and injective. We need to prove that

$$\mathfrak{a}_1 \cap \cdots \cap \mathfrak{a}_n = \mathfrak{a}_1 \cdots \mathfrak{a}_n$$

and that θ is surjective. We proceed by induction. The case $n = 2$ is Theorem 32.14. By induction, assume that

$$\mathfrak{a}_2 \cap \cdots \cap \mathfrak{a}_n = \mathfrak{a}_2 \cdots \mathfrak{a}_n.$$

We claim that

$$\mathfrak{a}_1 + \mathfrak{a}_2 \cdots \mathfrak{a}_n = A.$$

Indeed, since $\mathfrak{a}_1 + \mathfrak{a}_i = A$ for $i = 2, \ldots, n$, there exist some $a_i \in \mathfrak{a}_1$ and some $b_i \in \mathfrak{a}_i$ such that

$$a_i + b_i = 1, \quad i = 2, \ldots, n,$$

and by multiplying these equations, we get

$$a + b_2 \cdots b_n = 1,$$

where a is a sum of terms each containing some a_j as a factor, so $a \in \mathfrak{a}_1$ and $b_2 \cdots b_n \in \mathfrak{a}_2 \cdots \mathfrak{a}_n$, which shows that

$$\mathfrak{a}_1 + \mathfrak{a}_2 \cdots \mathfrak{a}_n = A,$$

as claimed. It follows that

$$\mathfrak{a}_1 \cap \mathfrak{a}_2 \cap \cdots \cap \mathfrak{a}_n = \mathfrak{a}_1 \cap (\mathfrak{a}_2 \cdots \mathfrak{a}_n) = \mathfrak{a}_1 \mathfrak{a}_2 \cdots \mathfrak{a}_n.$$

Let us now prove that θ is surjective by induction. The case $n = 2$ is Theorem 32.14. Let $x_1, \ldots, x_n$ be any $n \geq 3$ elements of A. First, applying Theorem 32.14 to $\mathfrak{a}_1$ and $\mathfrak{a}_2 \cdots \mathfrak{a}_n$, we can find $y_1 \in A$ such that

$$\begin{aligned} y_1 &\equiv 1 \pmod{\mathfrak{a}_1} \\ y_1 &\equiv 0 \pmod{\mathfrak{a}_2 \cdots \mathfrak{a}_n}. \end{aligned}$$

By the induction hypothesis, we can find $y_2, \ldots, y_n \in A$ such that for all i, j with $2 \leq i, j \leq n$,

$$\begin{aligned} y_i &\equiv 1 \pmod{\mathfrak{a}_i} \\ y_i &\equiv 0 \pmod{\mathfrak{a}_j}, \quad j \neq i. \end{aligned}$$

We claim that

$$x = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

works. Indeed, using the above congruences, for $i = 2, \ldots, n$, we get

$$x \equiv x_1y_1 + x_i \pmod{\mathfrak{a}_i}, \tag{$*$}$$

but since $\mathfrak{a}_2 \cdots \mathfrak{a}_n \subseteq \mathfrak{a}_i$ for $i = 2, \ldots, n$ and $y_1 \equiv 0 \pmod{\mathfrak{a}_2 \cdots \mathfrak{a}_n}$, we have

$$x_1y_1 \equiv 0 \pmod{\mathfrak{a}_i}, \quad i = 2, \ldots, n$$

and equation $(*)$ reduces to

$$x \equiv x_i \pmod{\mathfrak{a}_i}, \quad i = 2, \ldots, n.$$

For $i = 1$, we get

$$x \equiv x_1 \pmod{\mathfrak{a}_1},$$

therefore

$$x \equiv x_i \pmod{\mathfrak{a}_i}, \quad i = 1, \ldots, n.$$

proving surjectivity. $\square$

The classical version of the Chinese Remainder Theorem is the case where $A = \mathbb{Z}$ and where the ideals $\mathfrak{a}_i$ are defined by n pairwise relatively prime integers $m_1, \ldots, m_n$. By the Bezout identity, since m_i and m_j are relatively prime whenever $i \neq j$, there exist some $u_i, u_j \in \mathbb{Z}$ such that $u_im_i + u_jm_j = 1$, and so $m_i\mathbb{Z} + m_j\mathbb{Z} = \mathbb{Z}$. In this case, we get an isomorphism

$$\mathbb{Z}/(m_1 \cdots m_n)\mathbb{Z} \approx \prod_{i=1}^{n} \mathbb{Z}/m_i\mathbb{Z}.$$

In particular, if m is an integer greater than 1 and

$$m = \prod_i p_i^{r_i}$$

is its factorization into prime factors, then

$$\mathbb{Z}/m\mathbb{Z} \approx \prod_i \mathbb{Z}/p_i^{r_i}\mathbb{Z}.$$

In the previous situation where the integers $m_1, \ldots, m_n$ are pairwise relatively prime, if we write $m = m_1 \cdots m_n$ and $m_i' = m/m_i$ for $i = 1 \ldots, n$, then m_i and m_i' are relatively prime, and so m_i' has an inverse modulo m_i. If t_i is such an inverse, so that

$$m_i't_i \equiv 1 \pmod{m_i},$$

then it is not hard to show that for any $a_1, \dots, a_n \in \mathbb{Z}$,

$$x = a_1 t_1 m_1' + \cdots + a_n t_n m_n'$$

satisfies the congruences

$$x \equiv a_i \pmod{m_i}, \quad i = 1, \dots, n.$$

Theorem 32.15 can be used to characterize rings isomorphic to finite products of quotient rings. Such rings play a role in the structure theorem for torsion modules over a PID.

Given n rings $A_1, \dots, A_n$, recall that the product ring $A = A_1 \times \cdots \times A_n$ is the ring in which addition and multiplication are defined componenwise. That is,

$$\begin{aligned}(a_1, \dots, a_n) + (b_1, \dots, b_n) &= (a_1 + b_1, \dots, a_n + b_n)\\ (a_1, \dots, a_n) \cdot (b_1, \dots, b_n) &= (a_1 b_1, \dots, a_n b_n).\end{aligned}$$

The additive identity is $0_A = (0, \dots, 0)$ and the multiplicative identity is $1_A = (1, \dots, 1)$. Then, for $i = 1, \dots, n$, we can define the element $e_i \in A$ as follows:

$$e_i = (0, \dots, 0, 1, 0, \dots, 0),$$

where the 1 occurs in position i. Observe that the following properties hold for all $i, j = 1, \dots, n$:

$$\begin{aligned}e_i^2 &= e_i\\ e_i e_j &= 0, \quad i \neq j\\ e_1 + \cdots + e_n &= 1_A.\end{aligned}$$

Also, for any element $a = (a_1, \dots, a_n) \in A$, we have

$$e_i a = (0, \dots, 0, a_i, 0, \dots, 0) = pr_i(a),$$

where pr_i is the projection of A onto A_i. As a consequence

$$\mathrm{Ker}\,(pr_i) = (1_A - e_i)A.$$

Definition 32.3. Given a commutative ring A, a *direct decomposition* of A is a sequence $(\mathfrak{b}_1, \dots, \mathfrak{b}_n)$ of ideals in A such that there is an isomorphism $A \approx A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$.

The following theorem gives useful conditions characterizing direct decompositions of a ring.

Theorem 32.16. *Let A be a commutative ring and let $(\mathfrak{b}_1, \dots, \mathfrak{b}_n)$ be a sequence of ideals in A. The following conditions are equivalent:*

(a) The sequence $(\mathfrak{b}_1, \dots, \mathfrak{b}_n)$ is a direct decomposition of A.

(b) There exist some elements $e_1, \ldots, e_n$ of A such that

$$\begin{aligned} e_i^2 &= e_i \\ e_i e_j &= 0, \quad i \neq j \\ e_1 + \cdots + e_n &= 1_A, \end{aligned}$$

and $\mathfrak{b}_i = (1_A - e_i)A$, for $i, j = 1, \ldots, n$.

(c) We have $\mathfrak{b}_i + \mathfrak{b}_j = A$ for all $i \neq j$, and $\mathfrak{b}_1 \cdots \mathfrak{b}_n = (0)$.

(d) We have $\mathfrak{b}_i + \mathfrak{b}_j = A$ for all $i \neq j$, and $\mathfrak{b}_1 \cap \cdots \cap \mathfrak{b}_n = (0)$.

Proof. Assume (a). Since we have an isomorphism $A \approx A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$, we may identify A with $A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$, and $\mathfrak{b}_\mathfrak{i}$ with $\mathrm{Ker}\,(pr_i)$. Then, $e_1, \ldots, e_n$ are the elements defined just before Definition 32.3. As noted, $\mathfrak{b}_i = \mathrm{Ker}\,(pr_i) = (1_A - e_i)A$. This proves (b).

Assume (b). Since $\mathfrak{b}_i = (1_A - e_i)A$ and A is a ring with unit 1_A, we have $1_A - e_i \in \mathfrak{b}_i$ for $i = 1, \ldots, n$. For all $i \neq j$, we also have $e_i(1_A - e_j) = e_i - e_i e_j = e_i$, so (because $\mathfrak{b}_j$ is an ideal), $e_i \in \mathfrak{b}_j$, and thus, $1_A = 1_A - e_i + e_i \in \mathfrak{b}_\mathfrak{i} + \mathfrak{b}_j$, which shows that $\mathfrak{b}_i + \mathfrak{b}_j = A$ for all $i \neq j$. Furthermore, for any $x_i \in A$, with $1 \leq i \leq n$, we have

$$\begin{aligned} \prod_{i=1}^n x_i(1_A - e_i) &= \left(\prod_{i=1}^n x_i\right) \prod_{i=1}^n (1_A - e_i) \\ &= \left(\prod_{i=1}^n x_i\right)(1_A - \sum_{i=1}^n e_i) \\ &= 0, \end{aligned}$$

which proves that $\mathfrak{b}_1 \cdots \mathfrak{b}_n = (0)$. Thus, (c) holds.

The equivalence of (c) and (d) follows from the proof of Theorem 32.15.

The fact that (c) implies (a) is an immediate consequence of Theorem 32.15. $\square$

Here is example of Theorem 32.16. Take the commutative ring of residue classes mod 30, namely

$$A := \mathbb{Z}/30\mathbb{Z} = \{\bar{i}\}_{i=0}^{29}.$$

Let

$$\begin{aligned} \mathfrak{b}_1 &= 2\mathbb{Z}/30\mathbb{Z} := \{2\bar{i}\}_{i=0}^{14} \\ \mathfrak{b}_2 &= 3\mathbb{Z}/30\mathbb{Z} := \{3\bar{i}\}_{i=0}^{9} \\ \mathfrak{b}_3 &= 5\mathbb{Z}/30\mathbb{Z} := \{5\bar{i}\}_{i=0}^{5}. \end{aligned}$$

Each $\mathfrak{b}_i$ is an ideal in $\mathbb{Z}/30\mathbb{Z}$. Furthermore

$$\mathbb{Z}/30\mathbb{Z} = (\mathbb{Z}/30\mathbb{Z})/(2\mathbb{Z}/30\mathbb{Z}) \times (\mathbb{Z}/30\mathbb{Z})/(3\mathbb{Z}/30\mathbb{Z}) \times (\mathbb{Z}/30\mathbb{Z})/(5\mathbb{Z}/30\mathbb{Z}),$$

where

$$e_1 = (1,0,0) \to \overline{15}, \qquad e_2 = (0,1,0) \to \overline{10}, \qquad e_3 = (0,0,1) \to \overline{6},$$

since

$$\begin{gathered}\overline{15}^2 = \overline{15}, \qquad \overline{10}^2 = \overline{10}, \qquad \overline{6}^2 = \overline{6} \\ \overline{15}\,\overline{10} = \overline{15}\,\overline{6} = \overline{10}\,\overline{6} = 0, \qquad \overline{15} + \overline{10} + \overline{6} = \overline{1}.\end{gathered}$$

Note that $\overline{15}$ corresponds to $\overline{1} \in (\mathbb{Z}/30\mathbb{Z})/(2\mathbb{Z}/30\mathbb{Z})$, $\overline{10}$ corresponds to $\overline{1} \in (\mathbb{Z}/30\mathbb{Z})/(3\mathbb{Z}/30\mathbb{Z})$, while $\overline{6}$ corresponds to $\overline{1} \in (\mathbb{Z}/30\mathbb{Z})/(5\mathbb{Z}/30\mathbb{Z})$.

32.3 Noetherian Rings and Hilbert's Basis Theorem

Given a (commutative) ring A (with unit element 1), an ideal $\mathfrak{A} \subseteq A$ is said to be *finitely generated* if there exists a finite set $\{a_1, \ldots, a_n\}$ of elements from $\mathfrak{A}$ so that

$$\mathfrak{A} = (a_1, \ldots, a_n) = \{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \in A,\ 1 \leq i \leq n\}.$$

If K is a field, it turns out that every polynomial ideal $\mathfrak{A}$ in $K[X_1, \ldots, X_m]$ is finitely generated. This fact due to Hilbert and known as Hilbert's basis theorem, has very important consequences. For example, in algebraic geometry, one is interested in the zero locus of a set of polyomial equations, i.e., the set, $V(\mathcal{P})$, of n-tuples $(\lambda_1, \ldots, \lambda_n) \in K^n$ so that

$$P_i(\lambda_1, \ldots, \lambda_n) = 0$$

for all polynomials $P_i(X_1, \ldots, X_n)$ in some given family, $\mathcal{P} = (P_i)_{i \in I}$. However, it is clear that

$$V(\mathcal{P}) = V(\mathfrak{A}),$$

where $\mathfrak{A}$ is the ideal generated by $\mathcal{P}$. Then, Hilbert's basis theorem says that $V(\mathfrak{A})$ is actually defined by a *finite* number of polynomials (any set of generators of $\mathfrak{A}$), even if $\mathcal{P}$ is infinite.

The property that every ideal in a ring is finitely generated is equivalent to other natural properties, one of which is the so-called *ascending chain condition*, abbreviated *a.c.c.* Before proving Hilbert's basis theorem, we explore the equivalence of these conditions.

Definition 32.4. Let A be a commutative ring with unit 1. We say that A satisfies the *ascending chain condition*, for short, the *a.c.c*, if for every ascending chain of ideals

$$\mathfrak{A}_1 \subseteq \mathfrak{A}_2 \subseteq \cdots \subseteq \mathfrak{A}_i \subseteq \cdots,$$

there is some integer $n \geq 1$ so that

$$\mathfrak{A}_i = \mathfrak{A}_n \quad \text{for all} \quad i \geq n+1.$$

We say that A satisfies the *maximum condition* if every nonempty collection C of ideals in A has a maximal element, i.e., there is some ideal $\mathfrak{A} \in C$ which is not contained in any other ideal in C.

Proposition 32.17. *A ring A satisfies the a.c.c if and only if it satisfies the maximum condition.*

Proof. Suppose that A does not satisfy the a.c.c. Then, there is an infinite strictly ascending sequence of ideals

$$\mathfrak{A}_1 \subset \mathfrak{A}_2 \subset \cdots \subset \mathfrak{A}_i \subset \cdots,$$

and the collection $C = \{\mathfrak{A}_i\}$ has no maximal element.

Conversely, assume that A satisfies the a.c.c. Let C be a nonempty collection of ideals Since C is nonempty, we may pick some ideal $\mathfrak{A}_1$ in C. If $\mathfrak{A}_1$ is not maximal, then there is some ideal $\mathfrak{A}_2$ in C so that

$$\mathfrak{A}_1 \subset \mathfrak{A}_2.$$

Using this process, if C has no maximal element, we can define by induction an infinite strictly increasing sequence

$$\mathfrak{A}_1 \subset \mathfrak{A}_2 \subset \cdots \subset \mathfrak{A}_i \subset \cdots.$$

However, the a.c.c. implies that such a sequence cannot exist. Therefore, C has a maximal element. ☐

Having shown that the a.c.c. condition is equivalent to the maximal condition, we now prove that the a.c.c. condition is equivalent to the fact that every ideal is finitely generated.

Proposition 32.18. *A ring A satisfies the a.c.c if and only if every ideal is finitely generated.*

Proof. Assume that every ideal is finitely generated. Consider an ascending sequence of ideals

$$\mathfrak{A}_1 \subseteq \mathfrak{A}_2 \subseteq \cdots \subseteq \mathfrak{A}_i \subseteq \cdots.$$

Observe that $\mathfrak{A} = \bigcup_i \mathfrak{A}_i$ is also an ideal. By hypothesis, $\mathfrak{A}$ has a finite generating set $\{a_1, \ldots, a_n\}$. By definition of $\mathfrak{A}$, each a_i belongs to some $\mathfrak{A}_{j_i}$, and since the $\mathfrak{A}_i$ form an ascending chain, there is some m so that $a_i \in \mathfrak{A}_m$ for $i = 1, \ldots, n$. But then,

$$\mathfrak{A}_i = \mathfrak{A}_m$$

for all $i \geq m+1$, and the a.c.c. holds.

Conversely, assume that the a.c.c. holds. Let $\mathfrak{A}$ be any ideal in A and consider the family C of subideals of $\mathfrak{A}$ that are finitely generated. The family C is nonempty, since (0) is a subideal of $\mathfrak{A}$. By Proposition 32.17, the family C has some maximal element, say $\mathfrak{B}$. For any $a \in \mathfrak{A}$, the ideal $\mathfrak{B} + (a)$ (where $\mathfrak{B} + (a) = \{b + \lambda a \mid b \in \mathfrak{B},\ \lambda \in A\}$) is also finitely generated (since $\mathfrak{B}$ is finitely generated), and by maximality, we have

$$\mathfrak{B} = \mathfrak{B} + (a).$$

So, we get $a \in \mathfrak{B}$ for all $a \in \mathfrak{A}$, and thus, $\mathfrak{A} = \mathfrak{B}$, and $\mathfrak{A}$ is finitely generated. $\square$

Definition 32.5. A commutative ring A (with unit 1) is called *noetherian* if it satisfies the a.c.c. condition. A *noetherian domain* is a noetherian ring that is also a domain.

By Proposition 32.17 and Proposition 32.18, a noetherian ring can also be defined as a ring that either satisfies the maximal property or such that every ideal is finitely generated. The proof of Hilbert's basis theorem will make use the following lemma:

Lemma 32.19. *Let A be a (commutative) ring. For every ideal $\mathfrak{A}$ in $A[X]$, for every $i \geq 0$, let $L_i(\mathfrak{A})$ denote the set of elements of A consisting of 0 and of the coefficients of X^i in all the polynomials $f(X) \in \mathfrak{A}$ which are of degree i. Then, the $L_i(\mathfrak{A})$'s form an ascending chain of ideals in A. Furthermore, if $\mathfrak{B}$ is any ideal of $A[X]$ so that $\mathfrak{A} \subseteq \mathfrak{B}$ and if $L_i(\mathfrak{A}) = L_i(\mathfrak{B})$ for all $i \geq 0$, then $\mathfrak{A} = \mathfrak{B}$.*

Proof. That $L_i(\mathfrak{A})$ is an ideal and that $L_i(\mathfrak{A}) \subseteq L_{i+1}(\mathfrak{A})$ follows from the fact that if $f(X) \in \mathfrak{A}$ and $g(X) \in \mathfrak{A}$, then $f(X) + g(X)$, $\lambda f(X)$, and $Xf(X)$ all belong to $\mathfrak{A}$. Now, let $g(X)$ be any polynomial in $\mathfrak{B}$, and assume that $g(X)$ has degree n. Since $L_n(\mathfrak{A}) = L_n(\mathfrak{B})$, there is some polynomial $f_n(X)$ in $\mathfrak{A}$, of degree n, so that $g(X) - f_n(X)$ is of degree at most $n-1$. Now, since $\mathfrak{A} \subseteq \mathfrak{B}$, the polynomial $g(X) - f_n(X)$ belongs to $\mathfrak{B}$. Using this process, we can define by induction a sequence of polynomials $f_{n+i}(X) \in \mathfrak{A}$, so that each $f_{n+i}(X)$ is either zero or has degree $n - i$, and

$$g(X) - (f_n(X) + f_{n+1}(X) + \cdots + f_{n+i}(X))$$

is of degree at most $n - i - 1$. Note that this last polynomial must be zero when $i = n$, and thus, $g(X) \in \mathfrak{A}$. $\square$

We now prove Hilbert's basis theorem. The proof is substantially Hilbert's original proof. A slightly shorter proof can be given but it is not as transparent as Hilbert's proof (see the remark just after the proof of Theorem 32.20, and Zariski and Samuel [194], Chapter IV, Section 1, Theorem 1).

Theorem 32.20. *(Hilbert's basis theorem) If A is a noetherian ring, then $A[X]$ is also a noetherian ring.*

Proof. Let $\mathfrak{A}$ be any ideal in $A[X]$, and denote by $\mathcal{L}$ the set of elements of A consisting of 0 and of all the coefficients of the highest degree terms of all the polynomials in $\mathfrak{A}$. Observe that

$$\mathcal{L} = \bigcup_i L_i(\mathfrak{A}).$$

Thus, $\mathcal{L}$ is an ideal in A (this can also be proved directly). Since A is noetherian, $\mathcal{L}$ is finitely generated, and let $\{a_1, \ldots, a_n\}$ be a set of generators of $\mathcal{L}$. Let $f_1(X), \ldots, f_n(X)$ be polynomials in $\mathfrak{A}$ having respectively $a_1, \ldots, a_n$ as highest degree term coefficients. These polynomials generate an ideal $\mathfrak{B}$. Let q be the maximum of the degrees of the $f_i(X)$'s. Now, pick any polynomial $g(X) \in \mathfrak{A}$ of degree $d \geq q$, and let aX^d be its term of highest degree. Since $a \in \mathcal{L}$, we have

$$a = \lambda_1 a_1 + \cdots + \lambda_n a_n,$$

for some $\lambda_i \in A$. Consider the polynomial

$$g_1(X) = \sum_{i=1}^{n} \lambda_i f_i(X) X^{d-d_i},$$

where d_i is the degree of $f_i(X)$. Now, $g(X) - g_1(X)$ is a polynomial in $\mathfrak{A}$ of degree at most $d-1$. By repeating this procedure, we get a sequence of polynomials $g_i(X)$ in $\mathfrak{B}$, having strictly decreasing degrees, and such that the polynomial

$$g(X) - (g_1(X) + \cdots + g_i(X))$$

is of degree at most $d - i$. This polynomial must be of degree at most $q - 1$ as soon as $i = d - q + 1$. Thus, we proved that every polynomial in $\mathfrak{A}$ of degree $d \geq q$ belongs to $\mathfrak{B}$.

It remains to take care of the polynomials in $\mathfrak{A}$ of degree at most $q - 1$. Since A is noetherian, each ideal $L_i(\mathfrak{A})$ is finitely generated, and let $\{a_{i1}, \ldots, a_{in_i}\}$ be a set of generators for $L_i(\mathfrak{A})$ (for $i = 0, \ldots, q-1$). Let $f_{ij}(X)$ be a polynomial in $\mathfrak{A}$ having $a_{ij}X^i$ as its highest degree term. Given any polynomial $g(X) \in \mathfrak{A}$ of degree $d \leq q - 1$, if we denote its term of highest degree by aX^d, then, as in the previous argument, we can write

$$a = \lambda_1 a_{d1} + \cdots + \lambda_{n_d} a_{dn_d},$$

and we define

$$g_1(X) = \sum_{i=1}^{n_d} \lambda_i f_{di}(X) X^{d-d_i},$$

where d_i is the degree of $f_{di}(X)$. Then, $g(X) - g_1(X)$ is a polynomial in $\mathfrak{A}$ of degree at most $d-1$, and by repeating this procedure at most q times, we get an element of $\mathfrak{A}$ of degree 0, and the latter is a linear combination of the f_{0i}'s. This proves that every polynomial in $\mathfrak{A}$ of degree at most $q-1$ is a combination of the polynomials $f_{ij}(X)$, for $0 \leq i \leq q-1$ and $1 \leq j \leq n_i$. Therefore, $\mathfrak{A}$ is generated by the $f_k(X)$'s and the $f_{ij}(X)$'s, a finite number of polynomials. $\square$

Remark: Only a small part of Lemma 32.19 was used in the above proof, namely, the fact that $L_i(\mathfrak{A})$ is an ideal. A shorter proof of Theorem 32.21 making full use of Lemma 32.19 can be given as follows:

Proof. (Second proof) Let $(\mathfrak{A}_i)_{i\geq 1}$ be an ascending sequence of ideals in $A[X]$. Consider the doubly indexed family $(L_i(\mathfrak{A}_j))$ of ideals in A. Since A is noetherian, by the maximal property, this family has a maximal element $L_p(\mathfrak{A}_q)$. Since the $L_i(\mathfrak{A}_j)$'s form an ascending sequence when either i or j is fixed, we have $L_i(\mathfrak{A}_j) = L_p(\mathfrak{A}_q)$ for all i and j with $i \geq p$ and $j \geq q$, and thus, $L_i(\mathfrak{A}_q) = L_i(\mathfrak{A}_j)$ for all i and j with $i \geq p$ and $j \geq q$. On the other hand, for any fixed i, the a.c.c. shows that there exists some integer $n(i)$ so that $L_i(\mathfrak{A}_j) = L_i(\mathfrak{A}_{n(i)})$ for all $j \geq n(i)$. Since $L_i(\mathfrak{A}_q) = L_i(\mathfrak{A}_j)$ when $i \geq p$ and $j \geq q$, we may take $n(i) = q$ if $i \geq p$. This shows that there is some n_0 so that $n(i) \leq n_0$ for all $i \geq 0$, and thus, we have $L_i(\mathfrak{A}_j) = L_i(\mathfrak{A}_{n(0)})$ for every i and for every $j \geq n(0)$. By Lemma 32.19, we get $\mathfrak{A}_j = \mathfrak{A}_{n(0)}$ for every $j \geq n(0)$, establishing the fact that $A[X]$ satisfies the a.c.c. ☐

Using induction, we immediately obtain the following important result.

Corollary 32.21. *If A is a noetherian ring, then $A[X_1, \ldots, X_n]$ is also a noetherian ring.*

Since a field K is obviously noetherian (since it has only two ideals, (0) and K), we also have:

Corollary 32.22. *If K is a field, then $K[X_1, \ldots, X_n]$ is a noetherian ring.*

32.4 Futher Readings

The material of this Chapter is thoroughly covered in Lang [109], Artin [7], Mac Lane and Birkhoff [118], Bourbaki [25, 26], Malliavin [119], Zariski and Samuel [194], and Van Der Waerden [179].

Chapter 33

Tensor Algebras and Symmetric Algebras

Tensors are creatures that we would prefer did not exist but keep showing up whenever multilinearity manifests itself.

One of the goals of differential geometry is to be able to generalize "calculus on $\mathbb{R}^n$" to spaces more general than $\mathbb{R}^n$, namely manifolds. We would like to differentiate functions $f\colon M \to \mathbb{R}$ defined on a manifold, optimize functions (find their minima or maxima), but also to integrate such functions, as well as compute areas and volumes of subspaces of our manifold.

The suitable notion of differentiation is the notion of tangent map, a linear notion. One of the main discoveries made at the beginning of the twentieth century by Poincaré and Élie Cartan, is that the "right" approach to integration is to integrate *differential forms*, and not functions. To integrate a function f, we integrate the form $f\omega$, where ω is a *volume form* on the manifold M. The formalism of differential forms takes care of the process of the change of variables quite automatically, and allows for a very clean statement of *Stokes' formula*.

Differential forms can be combined using a notion of product called the wedge product, but what really gives power to the formalism of differential forms is the magical operation d of *exterior differentiation*. Given a form ω, we obtain another form $d\omega$, and remarkably, the following equation holds

$$dd\omega = 0.$$

As silly as it looks, the above equation lies at the core of the notion of cohomology, a powerful algebraic tool to understand the topology of manifolds, and more generally of topological spaces.

Élie Cartan had many of the intuitions that lead to the cohomology of differential forms, but it was George de Rham who defined it rigorously and proved some important theorems about it. It turns out that the notion of Laplacian can also be defined on differential forms using a device due to Hodge, and some important theorems can be obtained: the Hodge decomposition theorem, and Hodge's theorem about the isomorphism between the de Rham cohomology groups and the spaces of harmonic forms.

To understand all this, one needs to learn about differential forms, which turn out to be certain kinds of skew-symmetric (also called alternating) tensors.

If one's only goal is to define differential forms, then it is possible to take some short cuts and to avoid introducing the general notion of a tensor. However, tensors that are not necessarily skew-symmetric arise naturally, such as the curvature tensor, and in the theory of vector bundles, general tensor products are needed.

Consequently, we made the (perhaps painful) decision to provide a fairly detailed exposition of tensors, starting with arbitrary tensors, and then specializing to symmetric and alternating tensors. In particular, we explain rather carefully the process of taking the dual of a tensor (of all three flavors).

We refrained from following the approach in which a tensor is defined as a multilinear map defined on a product of dual spaces, because it seems very artificial and confusing (certainly to us). This approach relies on duality results that only hold in finite dimension, and consequently unecessarily restricts the theory of tensors to finite dimensional spaces. We also feel that it is important to begin with a coordinate-free approach. Bases can be chosen for computations, but tensor algebra should not be reduced to raising or lowering indices.

Readers who feel that they are familiar with tensors should probably skip this chapter and the next. They can come back to them "by need."

We begin by defining tensor products of vector spaces over a field and then we investigate some basic properties of these tensors, in particular the existence of bases and duality. After this we investigate special kinds of tensors, namely symmetric tensors and skew-symmetric tensors. Tensor products of modules over a commutative ring with identity will be discussed very briefly. They show up naturally when we consider the space of sections of a tensor product of vector bundles.

Given a linear map $f\colon E \to F$ (where E and F are two vector spaces over a field K), we know that if we have a basis $(u_i)_{i\in I}$ for E, then f is completely determined by its values $f(u_i)$ on the basis vectors. For a multilinear map $f\colon E^n \to F$, we don't know if there is such a nice property but it would certainly be very useful.

In many respects tensor products allow us to define multilinear maps in terms of their action on a suitable basis. The crucial idea is to *linearize*, that is, to create a new vector space $E^{\otimes n}$ such that the multilinear map $f\colon E^n \to F$ is turned into a *linear map* $f_\otimes\colon E^{\otimes n} \to F$ which is equivalent to f in a strong sense. If in addition, f is symmetric, then we can define a symmetric tensor power $\mathrm{Sym}^n(E)$, and every symmetric multilinear map $f\colon E^n \to F$ is turned into a *linear map* $f_\odot\colon \mathrm{Sym}^n(E) \to F$ which is equivalent to f in a strong sense. Similarly, if f is alternating, then we can define a skew-symmetric tensor power $\bigwedge^n(E)$, and every alternating multilinear map is turned into a *linear map* $f_\wedge\colon \bigwedge^n(E) \to F$ which is equivalent to f in a strong sense.

Tensor products can be defined in various ways, some more abstract than others. We try to stay down to earth, without excess.

Before proceeding any further, we review some facts about dual spaces and pairings. Pairings will be used to deal with dual spaces of tensors.

33.1 Linear Algebra Preliminaries: Dual Spaces and Pairings

We assume that we are dealing with vector spaces over a field K. As usual the *dual space* E^* of a vector space E is defined by $E^* = \mathrm{Hom}(E, K)$. The dual space E^* is the vector space consisting of all linear maps $\omega\colon E \to K$ with values in the field K.

A problem that comes up often is to decide when a space E is isomorphic to the dual F^* of some other space F (possibly equal to E). The notion of pairing due to Pontrjagin provides a very clean criterion.

Definition 33.1. Given two vector spaces E and F over a field K, a map $\langle -, - \rangle\colon E \times F \to K$ is a *nondegenerate pairing* iff it is bilinear and iff $\langle u, v \rangle = 0$ for all $v \in F$ implies $u = 0$, and $\langle u, v \rangle = 0$ for all $u \in E$ implies $v = 0$. A nondegenerate pairing induces two linear maps $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ defined such that for all for all $u \in E$ and all $v \in F$, $\varphi(u)$ is the linear form in F^* and $\psi(v)$ is the linear form in E^* given by

$$\begin{aligned} \varphi(u)(y) &= \langle u, y \rangle \quad \text{for all } y \in F \\ \psi(v)(x) &= \langle x, v \rangle \quad \text{for all } x \in E. \end{aligned}$$

Schematically, $\varphi(u) = \langle u, - \rangle$ and $\psi(v) = \langle -, v \rangle$.

Proposition 33.1. *For every nondegenerate pairing $\langle -, - \rangle\colon E \times F \to K$, the induced maps $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are linear and injective. Furthermore, if E and F are finite dimensional, then $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are bijective.*

Proof. The maps $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are linear because $u, v \mapsto \langle u, v \rangle$ is bilinear. Assume that $\varphi(u) = 0$. This means that $\varphi(u)(y) = \langle u, y \rangle = 0$ for all $y \in F$, and as our pairing is nondegenerate, we must have $u = 0$. Similarly, ψ is injective. If E and F are finite dimensional, then $\dim(E) = \dim(E^*)$ and $\dim(F) = \dim(F^*)$. However, the injectivity of φ and ψ implies that that $\dim(E) \leq \dim(F^*)$ and $\dim(F) \leq \dim(E^*)$. Consequently $\dim(E) \leq \dim(F)$ and $\dim(F) \leq \dim(E)$, so $\dim(E) = \dim(F)$. Therefore, $\dim(E) = \dim(F^*)$ and φ is bijective (and similarly $\dim(F) = \dim(E^*)$ and ψ is bijective). $\square$

Proposition 33.1 shows that when E and F are finite dimensional, a nondegenerate pairing induces *canonical isomorphims* $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$; that is, isomorphisms that do not depend on the choice of bases. An important special case is the case where $E = F$ and we have an inner product (a symmetric, positive definite bilinear form) on E.

Remark: When we use the term "canonical isomorphism," we mean that such an isomorphism is defined independently of any choice of bases. For example, if E is a finite dimensional vector space and $(e_1, \ldots, e_n)$ is any basis of E, we have the dual basis $(e_1^*, \ldots, e_n^*)$ of E^* (where, $e_i^*(e_j) = \delta_{i\,j}$), and thus the map $e_i \mapsto e_i^*$ is an isomorphism between E and E^*. This isomorphism is *not* canonical.

On the other hand, if $\langle -, - \rangle$ is an inner product on E, then Proposition 33.1 shows that the nondegenerate pairing $\langle -, - \rangle$ on $E \times E$ induces a canonical isomorphism between E and E^*. This isomorphism is often denoted $\flat \colon E \to E^*$, and we usually write $u^\flat$ for $\flat(u)$, with $u \in E$. Schematically, $u^\flat = \langle u, - \rangle$. The inverse of $\flat$ is denoted $\sharp \colon E^* \to E$, and given any linear form $\omega \in E^*$, we usually write $\omega^\sharp$ for $\sharp(\omega)$. Schematically, $\omega = \langle \omega^\sharp, - \rangle$.

Given any basis, $(e_1, \ldots, e_n)$ of E (not necessarily orthonormal), let (g_{ij}) be the $n \times n$-matrix given by $g_{ij} = \langle e_i, e_j \rangle$ (the *Gram* matrix of the inner product). Recall that the *dual basis* $(e_1^*, \ldots, e_n^*)$ of E^* consists of the coordinate forms $e_i^* \in E^*$, which are characterized by the following properties:

$$e_i^*(e_j) = \delta_{ij}, \quad 1 \leq i, j \leq n.$$

The inverse of the Gram matrix (g_{ij}) is often denoted by (g^{ij}) (by raising the indices).

The tradition of raising and lowering indices is pervasive in the literature on tensors. It is indeed useful to have some notational convention to distinguish between vectors and linear forms (also called *one-forms* or *covectors*). The usual convention is that coordinates of vectors are written using superscripts, as in $u = \sum_{i=1}^n u^i e_i$, and coordinates of one-forms are written using subscripts, as in $\omega = \sum_{i=1}^n \omega_i e_i^*$. Actually, since vectors are indexed with subscripts, one-forms are indexed with superscripts, so e_i^* should be written as e^i.

The motivation is that summation signs can then be omitted, according to the *Einstein summation convention*. According to this convention, whenever a summation variable (such as i) appears both as a subscript and a superscript in an expression, it is assumed that it is involved in a summation. For example the sum $\sum_{i=1}^n u^i e_i$ is abbreviated as

$$u^i e_i,$$

and the sum $\sum_{i=1}^n \omega_i e^i$ is abbreviated as

$$\omega_i e^i.$$

In this text we will not use the Einstein summation convention, which we find somewhat confusing, and we will also write e_i^* instead of e^i.

The maps $\flat$ and $\sharp$ can be described explicitly in terms of the Gram matrix of the inner product and its inverse.

Proposition 33.2. *For any vector space E, given a basis $(e_1, \ldots, e_n)$ for E and its dual basis $(e_1^*, \ldots, e_n^*)$ for E^*, for any inner product $\langle -, - \rangle$ on E, if (g_{ij}) is its Gram matrix, with*

$g_{ij} = \langle e_i, e_j \rangle$, and (g^{ij}) is its inverse, then for every vector $u = \sum_{j=1}^n u^j e_j \in E$ and every one-form $\omega = \sum_{i=1}^n \omega_i e_i^* \in E^*$, we have

$$u^\flat = \sum_{i=1}^n \omega_i e_i^*, \quad \textit{with} \quad \omega_i = \sum_{j=1}^n g_{ij} u^j,$$

and

$$\omega^\sharp = \sum_{j=1}^n (\omega^\sharp)^j e_j, \quad \textit{with} \quad (\omega^\sharp)^i = \sum_{j=1}^n g^{ij} \omega_j.$$

Proof. For every $u = \sum_{j=1}^n u^j e_j$, since $u^\flat(v) = \langle u, v \rangle$ for all $v \in E$, we have

$$u^\flat(e_i) = \langle u, e_i \rangle = \left\langle \sum_{j=1}^n u^j e_j, e_i \right\rangle = \sum_{j=1}^n u^j \langle e_j, e_i \rangle = \sum_{j=1}^n g_{ij} u^j,$$

so we get

$$u^\flat = \sum_{i=1}^n \omega_i e_i^*, \quad \text{with} \quad \omega_i = \sum_{j=1}^n g_{ij} u^j.$$

If we write $\omega \in E^*$ as $\omega = \sum_{i=1}^n \omega_i e_i^*$ and $\omega^\sharp \in E$ as $\omega^\sharp = \sum_{j=1}^n (\omega^\sharp)^j e_j$, since

$$\omega_i = \omega(e_i) = \langle \omega^\sharp, e_i \rangle = \sum_{j=1}^n (\omega^\sharp)^j g_{ij}, \qquad 1 \leq i \leq n,$$

we get

$$(\omega^\sharp)^i = \sum_{j=1}^n g^{ij} \omega_j,$$

where (g^{ij}) is the inverse of the matrix (g_{ij}). ∎

The map $\flat$ has the effect of lowering (flattening!) indices, and the map $\sharp$ has the effect of raising (sharpening!) indices.

Here is an explicit example of Proposition 33.2. Let (e_1, e_2) be a basis of E such that

$$\langle e_1, e_1 \rangle = 1, \qquad \langle e_1, e_2 \rangle = 2, \qquad \langle e_2, e_2 \rangle = 5.$$

Then

$$g = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}, \qquad g^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}.$$

Set $u = u^1 e_1 + u^2 e_2$ and observe that

$$\begin{aligned} u^\flat(e_1) &= \langle u^1 e_1 + u^2 e_2, e_1 \rangle = \langle e_1, e_1 \rangle u^1 + \langle e_2, e_1 \rangle u^2 = g_{11} u^1 + g_{12} u^2 = u^1 + 2u^2 \\ u^\flat(e_2) &= \langle u^1 e_1 + u^2 e_2, e_2 \rangle = \langle e_1, e_2 \rangle u^1 + \langle e_2, e_2 \rangle u^2 = g_{21} u^1 + g_{22} u^2 = 2u^1 + 5u^2, \end{aligned}$$

which in turn implies that

$$u^\flat = \omega_1 e_1^* + \omega_2 e_2^* = u^\flat(e_1)e_1^* + u^\flat(e_2)e_2^* = (u^1 + 2u^2)e_1^* + (2u^1 + 5u^2)e_2^*.$$

Given $\omega = \omega_1 e_1^* + \omega_2 e_2^*$, we calculate $\omega^\sharp = (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2$ from the following two linear equalities:

$$\begin{aligned}
\omega_1 = \omega(e_1) &= \langle \omega^\sharp, e_1\rangle = \langle (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2, e_1\rangle \\
&= \langle e_1, e_1\rangle(\omega^\sharp)^1 + \langle e_2, e_1\rangle(\omega^\sharp)^2 = (\omega^\sharp)^1 + 2(\omega^\sharp)^2 = g_{11}(\omega^\sharp)^1 + g_{12}(\omega^\sharp)^2 \\
\omega_2 = \omega(e_2) &= \langle \omega^\sharp, e_2\rangle = \langle (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2, e_2\rangle \\
&= \langle e_1, e_2\rangle(\omega^\sharp)^1 + \langle e_2, e_2\rangle(\omega^\sharp)^2 = 2(\omega^\sharp)^1 + 5(\omega^\sharp)^2 = g_{21}(\omega^\sharp)^1 + g_{22}(\omega^\sharp)^2.
\end{aligned}$$

These equalities are concisely written as

$$\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} (\omega^\sharp)^1 \\ (\omega^\sharp)^2 \end{pmatrix} = g \begin{pmatrix} (\omega^\sharp)^1 \\ (\omega^\sharp)^2 \end{pmatrix}.$$

Then

$$\begin{pmatrix} (\omega^\sharp)^1 \\ (\omega^\sharp)^2 \end{pmatrix} = g^{-1} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix},$$

which in turn implies

$$(\omega^\sharp)^1 = 5\omega_1 - 2\omega_2, \qquad (\omega^\sharp)^2 = -2\omega_1 + \omega_2,$$

i.e.

$$\omega^\sharp = (5\omega_1 - 2\omega_2)e_1 + (-2\omega_1 + \omega_2)e_2.$$

The inner product $\langle -, -\rangle$ on E induces an inner product on E^* denoted $\langle -, -\rangle_{E^*}$, and given by

$$\langle \omega_1, \omega_2\rangle_{E^*} = \langle \omega_1^\sharp, \omega_2^\sharp\rangle, \quad \text{for all } \omega_1, \omega_2 \in E^*.$$

Then we have

$$\langle u^\flat, v^\flat\rangle_{E^*} = \langle (u^\flat)^\sharp, (v^\flat)^\sharp\rangle = \langle u, v\rangle \quad \text{for all} \quad u, v \in E.$$

If $(e_1, \ldots, e_n)$ is a basis of E and $g_{ij} = \langle e_i, e_j\rangle$, as

$$(e_i^*)^\sharp = \sum_{k=1}^n g^{ik} e_k,$$

an easy computation shows that

$$\langle e_i^*, e_j^*\rangle_{E^*} = \langle (e_i^*)^\sharp, (e_j^*)^\sharp\rangle = g^{ij};$$

that is, in the basis $(e_1^*, \ldots, e_n^*)$, the inner product on E^* is represented by the matrix (g^{ij}), the inverse of the matrix (g_{ij}).

The inner product on a finite vector space also yields a canonical isomorphism between the space $\mathrm{Hom}(E, E; K)$ of bilinear forms on E, and the space $\mathrm{Hom}(E, E)$ of linear maps from E to itself. Using this isomorphism, we can define the trace of a bilinear form in an intrinsic manner. This technique is used in differential geometry, for example, to define the divergence of a differential one-form.

Proposition 33.3. *If $\langle -, - \rangle$ is an inner product on a finite vector space E (over a field, K), then for every bilinear form $f \colon E \times E \to K$, there is a unique linear map $f^\natural \colon E \to E$ such that*

$$f(u, v) = \langle f^\natural(u), v \rangle, \quad \text{for all } u, v \in E.$$

The map $f \mapsto f^\natural$ is a linear isomorphism between $\mathrm{Hom}(E, E; K)$ *and* $\mathrm{Hom}(E, E)$.

Proof. For every $g \in \mathrm{Hom}(E, E)$, the map given by

$$f(u, v) = \langle g(u), v \rangle, \quad u, v \in E,$$

is clearly bilinear. It is also clear that the above defines a linear map from $\mathrm{Hom}(E, E)$ to $\mathrm{Hom}(E, E; K)$. This map is injective, because if $f(u, v) = 0$ for all $u, v \in E$, as $\langle -, - \rangle$ is an inner product, we get $g(u) = 0$ for all $u \in E$. Furthermore, both spaces $\mathrm{Hom}(E, E)$ and $\mathrm{Hom}(E, E; K)$ have the same dimension, so our linear map is an isomorphism. $\square$

If $(e_1, \ldots, e_n)$ is an orthonormal basis of E, then we check immediately that the trace of a linear map g (which is independent of the choice of a basis) is given by

$$\mathrm{tr}(g) = \sum_{i=1}^{n} \langle g(e_i), e_i \rangle,$$

where $n = \dim(E)$.

Definition 33.2. We define the *trace of the bilinear form f* by

$$\mathrm{tr}(f) = \mathrm{tr}(f^\natural).$$

From Proposition 33.3, $\mathrm{tr}(f)$ is given by

$$\mathrm{tr}(f) = \sum_{i=1}^{n} f(e_i, e_i),$$

for any orthonormal basis $(e_1, \ldots, e_n)$ of E. We can also check directly that the above expression is independent of the choice of an orthonormal basis.

We demonstrate how to calculate $\mathrm{tr}(f)$ where $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ with $f((x_1, y_1), (x_2, y_2)) = x_1x_2 + 2x_2y_1 + 3x_1y_2 - y_1y_2$. Under the standard basis for $\mathbb{R}^2$, the bilinear form f is represented as

$$\begin{pmatrix} x_1 & y_1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}.$$

This matrix representation shows that

$$f^{\natural} = \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix}^{\top} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix},$$

and hence

$$\mathrm{tr}(f) = \mathrm{tr}(f^{\natural}) = \mathrm{tr} \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} = 0.$$

We will also need the following proposition to show that various families are linearly independent.

Proposition 33.4. *Let E and F be two nontrivial vector spaces and let $(u_i)_{i\in I}$ be any family of vectors $u_i \in E$. The family $(u_i)_{i\in I}$ is linearly independent iff for every family $(v_i)_{i\in I}$ of vectors $v_i \in F$, there is some linear map $f \colon E \to F$ so that $f(u_i) = v_i$ for all $i \in I$.*

Proof. Left as an exercise. ☐

33.2 Tensors Products

First we define tensor products, and then we prove their existence and uniqueness up to isomorphism.

Definition 33.3. Let K be a given field, and let $E_1, \ldots, E_n$ be $n \geq 2$ given vector spaces. For any vector space F, a map $f \colon E_1 \times \cdots \times E_n \to F$ is *multilinear* iff it is linear in each of its argument; that is,

$$\begin{aligned} f(u_1, \ldots u_{i_1}, v + w, u_{i+1}, \ldots, u_n) &= f(u_1, \ldots u_{i_1}, v, u_{i+1}, \ldots, u_n) \\ &\quad + f(u_1, \ldots u_{i_1}, w, u_{i+1}, \ldots, u_n) \\ f(u_1, \ldots u_{i_1}, \lambda v, u_{i+1}, \ldots, u_n) &= \lambda f(u_1, \ldots u_{i_1}, v, u_{i+1}, \ldots, u_n), \end{aligned}$$

for all $u_j \in E_j$ $(j \neq i)$, all $v, w \in E_i$ and all $\lambda \in K$, for $i = 1 \ldots, n$.

The set of multilinear maps as above forms a vector space denoted $\mathrm{L}(E_1, \ldots, E_n; F)$ or $\mathrm{Hom}(E_1, \ldots, E_n; F)$. When $n = 1$, we have the vector space of linear maps $\mathrm{L}(E, F)$ (also denoted $\mathrm{Hom}(E, F)$). (To be very precise, we write $\mathrm{Hom}_K(E_1, \ldots, E_n; F)$ and $\mathrm{Hom}_K(E, F)$.)

Definition 33.4. A *tensor product* of $n \geq 2$ vector spaces $E_1, \ldots, E_n$ is a vector space T together with a multilinear map $\varphi\colon E_1 \times \cdots \times E_n \to T$, such that for every vector space F and for every multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, there is a unique linear map $f_\otimes\colon T \to F$ with

$$f(u_1, \ldots, u_n) = f_\otimes(\varphi(u_1, \ldots, u_n)),$$

for all $u_1 \in E_1, \ldots, u_n \in E_n$, or for short

$$f = f_\otimes \circ \varphi.$$

Equivalently, there is a unique linear map $f_\otimes$ such that the following diagram commutes.

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\varphi"] \arrow[dr, "f"'] & T \arrow[d, "f_\otimes"] \\
& F
\end{tikzcd}$$

The above property is called the *universal mapping property* of the tensor product (T, φ).

We show that any two tensor products (T_1, φ_1) and (T_2, φ_2) for $E_1, \ldots, E_n$, are isomorphic.

Proposition 33.5. *Given any two tensor products (T_1, φ_1) and (T_2, φ_2) for $E_1, \ldots, E_n$, there is an isomorphism $h\colon T_1 \to T_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Focusing on (T_1, φ_1), we have a multilinear map $\varphi_2\colon E_1 \times \cdots \times E_n \to T_2$, and thus there is a unique linear map $(\varphi_2)_\otimes\colon T_1 \to T_2$ with

$$\varphi_2 = (\varphi_2)_\otimes \circ \varphi_1$$

as illustrated by the following commutative diagram.

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_2"'] & T_1 \arrow[d, "(\varphi_2)_\otimes"] \\
& T_2
\end{tikzcd}$$

Similarly, focusing now on on (T_2, φ_2), we have a multilinear map $\varphi_1\colon E_1 \times \cdots \times E_n \to T_1$, and thus there is a unique linear map $(\varphi_1)_\otimes\colon T_2 \to T_1$ with

$$\varphi_1 = (\varphi_1)_\otimes \circ \varphi_2$$

as illustrated by the following commutative diagram.

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_1"'] & T_2 \arrow[d, "(\varphi_1)_\otimes"] \\
& T_1
\end{tikzcd}$$

Putting these diagrams together, we obtain the commutative diagrams

$$\begin{tikzcd}
& T_1 \arrow[d, "(\varphi_2)_\otimes"] \\
E_1 \times \cdots \times E_n \arrow[ur, "\varphi_1"] \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_1"'] & T_2 \arrow[d, "(\varphi_1)_\otimes"] \\
& T_1
\end{tikzcd}$$

and

$$\begin{tikzcd}
& T_2 \arrow[d, "(\varphi_1)_\otimes"] \\
E_1 \times \cdots \times E_n \arrow[ur, "\varphi_2"] \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_2"'] & T_1 \arrow[d, "(\varphi_2)_\otimes"] \\
& T_2,
\end{tikzcd}$$

which means that

$$\varphi_1 = (\varphi_1)_\otimes \circ (\varphi_2)_\otimes \circ \varphi_1 \quad \text{and} \quad \varphi_2 = (\varphi_2)_\otimes \circ (\varphi_1)_\otimes \circ \varphi_2.$$

On the other hand, focusing on (T_1, φ_1), we have a multilinear map $\varphi_1 \colon E_1 \times \cdots \times E_n \to T_1$, but the unique linear map $h \colon T_1 \to T_1$ with

$$\varphi_1 = h \circ \varphi_1$$

is $h = \mathrm{id}$, as illustrated by the following commutative diagram

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_1"'] & T_1 \arrow[d, "\mathrm{id}"] \\
& T_1,
\end{tikzcd}$$

and since $(\varphi_1)_\otimes \circ (\varphi_2)_\otimes$ is linear as a composition of linear maps, we must have

$$(\varphi_1)_\otimes \circ (\varphi_2)_\otimes = \mathrm{id}.$$

Similarly, we have the commutative diagram

$$\begin{tikzcd} E_1 \times \cdots \times E_n \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_2"'] & T_2 \arrow[d, "\mathrm{id}"] \\ & T_2, \end{tikzcd}$$

and we must have

$$(\varphi_2)_\otimes \circ (\varphi_1)_\otimes = \mathrm{id}.$$

This shows that $(\varphi_1)_\otimes$ and $(\varphi_2)_\otimes$ are inverse linear maps, and thus, $(\varphi_2)_\otimes \colon T_1 \to T_2$ is an isomorphism between T_1 and T_2. $\square$

Now that we have shown that tensor products are unique up to isomorphism, we give a construction that produces them. Tensor products are obtained from free vector spaces by a quotient process, so let us begin by describing the construction of the free vector space generated by a set.

For simplicity assume that our set I is finite, say

$$I = \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}.$$

The construction works for any field K (and in fact for any commutative ring A, in which case we obtain the free A-module generated by I). Assume that $K = \mathbb{R}$. The *free vector space generated by* I is the set of all formal linear combinations of the form

$$a\heartsuit + b\diamondsuit + c\spadesuit + d\clubsuit,$$

with $a, b, c, d \in \mathbb{R}$. It is assumed that the order of the terms does not matter. For example,

$$2\heartsuit - 5\diamondsuit + 3\spadesuit = -5\diamondsuit + 2\heartsuit + 3\spadesuit.$$

Addition and multiplication by a scalar are are defined as follows:

$$\begin{aligned} &(a_1\heartsuit + b_1\diamondsuit + c_1\spadesuit + d_1\clubsuit) + (a_2\heartsuit + b_2\diamondsuit + c_2\spadesuit + d_2\clubsuit) \\ &\qquad = (a_1 + a_2)\heartsuit + (b_1 + b_2)\diamondsuit + (c_1 + c_2)\spadesuit + (d_1 + d_2)\clubsuit, \end{aligned}$$

and

$$\alpha \cdot (a\heartsuit + b\diamondsuit + c\spadesuit + d\clubsuit) = \alpha a\heartsuit + \alpha b\diamondsuit + \alpha c\spadesuit + \alpha d\clubsuit,$$

for all $a, b, c, d, \alpha \in \mathbb{R}$. With these operations, it is immediately verified that we obtain a vector space denoted $\mathbb{R}^{(I)}$. The set I can be viewed as embedded in $\mathbb{R}^{(I)}$ by the injection ι given by

$$\iota(\heartsuit) = 1\heartsuit, \quad \iota(\diamondsuit) = 1\diamondsuit, \quad \iota(\spadesuit) = 1\spadesuit, \quad \iota(\clubsuit) = 1\clubsuit.$$

Thus, $\mathbb{R}^{(I)}$ can be viewed as the vector space with the special basis $I = \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}$. In our case, $\mathbb{R}^{(I)}$ is isomorophic to $\mathbb{R}^4$.

The exact same construction works for any field K, and we obtain a vector space denoted by $K^{(I)}$ and an injection $\iota\colon I \to K^{(I)}$.

The main reason why the free vector space $K^{(I)}$ over a set I is interesting is that it satisfies a *universal mapping property*. This means that for every vector space F (over the field K), any function $h\colon I \to F$, where F is *considered just a set*, has a unique linear extension $\overline{h}\colon K^{(I)} \to F$. By extension, we mean that $\overline{h}(i) = h(i)$ for all $i \in I$, or more rigorously that $h = \overline{h} \circ \iota$.

For example, if $I = \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}$, $K = \mathbb{R}$, and $F = \mathbb{R}^3$, the function h given by

$$h(\heartsuit) = (1,1,1), \quad h(\diamondsuit) = (1,1,0), \quad h(\spadesuit) = (1,0,0), \quad h(\clubsuit) = (0,0-1)$$

has a unique linear extension $\overline{h}\colon \mathbb{R}^{(I)} \to \mathbb{R}^3$ to the free vector space $\mathbb{R}^{(I)}$, given by

$$\begin{aligned}\overline{h}(a\heartsuit + b\diamondsuit + c\spadesuit + d\clubsuit) &= a\overline{h}(\heartsuit) + b\overline{h}(\diamondsuit) + c\overline{h}(\spadesuit) + d\overline{h}(\clubsuit)\\ &= ah(\heartsuit) + bh(\diamondsuit) + ch(\spadesuit) + dh(\clubsuit)\\ &= a(1,1,1) + b(1,1,0) + c(1,0,0) + d(0,0,-1)\\ &= (a+b+c, a+b, a-d).\end{aligned}$$

To generalize the construction of a free vector space to infinite sets I, we observe that the formal linear combination $a\heartsuit + b\diamondsuit + c\spadesuit + d\clubsuit$ can be viewed as the function $f\colon I \to \mathbb{R}$ given by

$$f(\heartsuit) = a, \quad f(\diamondsuit) = b, \quad f(\spadesuit) = c, \quad f(\clubsuit) = d,$$

where $a, b, c, d \in \mathbb{R}$. More generally, we can replace $\mathbb{R}$ by any field K. If I is finite, then the set of all such functions is a vector space under pointwise addition and pointwise scalar multiplication. If I is infinite, since addition and scalar multiplication only makes sense for finite vectors, we require that our functions $f\colon I \to K$ take the value 0 except for possibly finitely many arguments. We can think of such functions as an infinite sequences $(f_i)_{i\in I}$ of elements f_i of K indexed by I, with only finitely many nonzero f_i. The formalization of this construction goes as follows.

Given any set I viewed as an index set, let $K^{(I)}$ be the set of all functions $f\colon I \to K$ such that $f(i) \neq 0$ only for finitely many $i \in I$. As usual, denote such a function by $(f_i)_{i\in I}$; it is a family of finite support. We make $K^{(I)}$ into a vector space by defining addition and scalar multiplication by

$$\begin{aligned}(f_i) + (g_i) &= (f_i + g_i)\\ \lambda(f_i) &= (\lambda f_i).\end{aligned}$$

The family $(e_i)_{i\in I}$ is defined such that $(e_i)_j = 0$ if $j \neq i$ and $(e_i)_i = 1$. It is a basis of the vector space $K^{(I)}$, so that every $w \in K^{(I)}$ can be uniquely written as a finite linear combination of the e_i. There is also an injection $\iota\colon I \to K^{(I)}$ such that $\iota(i) = e_i$ for every $i \in I$. Furthermore, it is easy to show that for any vector space F, and for any function

$h\colon I \to F$, there is a unique linear map $\overline{h}\colon K^{(I)} \to F$ such that $h = \overline{h} \circ \iota$, as in the following diagram.

$$\begin{tikzcd} I \arrow[r, "\iota"] \arrow[dr, "h"'] & K^{(I)} \arrow[d, "\overline{h}"] \\ & F \end{tikzcd}$$

Definition 33.5. The vector space $(K^{(I)}, \iota)$ constructed as above from a set I is called the *free vector space generated by* I (or over I). The commutativity of the above diagram is called the *universal mapping property* of the free vector space $(K^{(I)}, \iota)$ over I.

Using the proof technique of Proposition 33.5, it is not hard to prove that any two vector spaces satisfying the above universal mapping property are isomorphic.

We can now return to the construction of tensor products. For simplicity consider two vector spaces E_1 and E_2. Whatever $E_1 \otimes E_2$ and $\varphi\colon E_1 \times E_2 \to E_1 \otimes E_2$ are, since φ is supposed to be bilinear, we must have

$$\begin{aligned} \varphi(u_1 + u_2, v_1) &= \varphi(u_1, v_1) + \varphi(u_2, v_1) \\ \varphi(u_1, v_1 + v_2) &= \varphi(u_1, v_1) + \varphi(u_1, v_2) \\ \varphi(\lambda u_1, v_1) &= \lambda\varphi(u_1, v_1) \\ \varphi(u_1, \mu v_1) &= \mu\varphi(u_1, v_1) \end{aligned}$$

for all $u_1, u_2 \in E_1$, all $v_1, v_2 \in E_2$, and all $\lambda, \mu \in K$. Since $E_1 \otimes E_2$ must satisfy the universal mapping property of Definition 33.4, we may want to define $E_1 \otimes E_2$ as the free vector space $K^{(E_1 \times E_2)}$ generated by $I = E_1 \times E_2$ and let φ be the injection of $E_1 \times E_2$ into $K^{(E_1 \times E_2)}$. The problem is that in $K^{(E_1 \times E_2)}$, vectors such that

$$(u_1 + u_2, v_1) \quad \textit{and} \quad (u_1, v_1) + (u_2, v_2)$$

are different, when they should really be the same, since φ is bilinear. Since $K^{(E_1 \times E_2)}$ is free, there are no relations among the generators and this vector space is too big for our purpose.

The remedy is simple: take the quotient of the free vector space $K^{(E_1 \times E_2)}$ by the subspace N generated by the vectors of the form

$$\begin{aligned} &(u_1 + u_2, v_1) - (u_1, v_1) - (u_2, v_1) \\ &(u_1, v_1 + v_2) - (u_1, v_1) - (u_1, v_2) \\ &(\lambda u_1, v_1) - \lambda(u_1, v_1) \\ &(u_1, \mu v_1) - \mu(u_1, v_1). \end{aligned}$$

Then, if we let $E_1 \otimes E_2$ be the quotient space $K^{(E_1 \times E_2)}/N$ and let φ be the quotient map, this forces φ to be bilinear. Checking that $(K^{(E_1 \times E_2)}/N, \varphi)$ satisfies the universal mapping property is straightforward. Here is the detailed construction.

Theorem 33.6. *Given $n \geq 2$ vector spaces $E_1, \ldots, E_n$, a tensor product $(E_1 \otimes \cdots \otimes E_n, \varphi)$ for $E_1, \ldots, E_n$ can be constructed. Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the tensor product $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, where $u_1 \in E_1, \ldots, u_n \in E_n$, and for every multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, the unique linear map $f_\otimes\colon E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$ is defined by*

$$f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n)$$

on the generators $u_1 \otimes \cdots \otimes u_n$ of $E_1 \otimes \cdots \otimes E_n$.

Proof. First we apply the construction of a free vector space to the cartesian product $I = E_1 \times \cdots \times E_n$, obtaining the free vector space $M = K^{(I)}$ on $I = E_1 \times \cdots \times E_n$. Since every basis generator $e_i \in M$ is uniquely associated with some n-tuple $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$, we denote e_i by $(u_1, \ldots, u_n)$.

Next let N be the subspace of M generated by the vectors of the following type:

$$\begin{aligned}
&(u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n),\\
&(u_1, \ldots, \lambda u_i, \ldots, u_n) - \lambda(u_1, \ldots, u_i, \ldots, u_n).
\end{aligned}$$

We let $E_1 \otimes \cdots \otimes E_n$ be the quotient M/N of the free vector space M by N, $\pi\colon M \to M/N$ be the quotient map, and set

$$\varphi = \pi \circ \iota.$$

By construction, φ is multilinear, and since π is surjective and the $\iota(i) = e_i$ generate M, the fact that each i is of the form $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$ implies that $\varphi(u_1, \ldots, u_n)$ generate M/N. Thus, if we denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the space $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, with $u_i \in E_i$.

It remains to show that $(E_1 \otimes \cdots \otimes E_n, \varphi)$ satisfies the universal mapping property. To this end, we begin by proving there is a map h such that $f = h \circ \varphi$. Since $M = K^{(E_1 \times \cdots \times E_n)}$ is free on $I = E_1 \times \cdots \times E_n$, there is a unique linear map $\overline{f}\colon K^{(E_1 \times \cdots \times E_n)} \to F$, such that

$$f = \overline{f} \circ \iota,$$

as in the diagram below.

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\iota"] \arrow[dr, "f"'] & K^{(E_1 \times \cdots \times E_n)} = M \arrow[d, "\overline{f}"] \\
& F
\end{tikzcd}$$

Because f is multilinear, note that we must have $\overline{f}(w) = 0$ for every $w \in N$; for example, on the generator

$$(u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n)$$

we have

$$\begin{aligned}
&\overline{f}((u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n)) \\
&\qquad = f(u_1, \ldots, u_i + v_i, \ldots, u_n) - f(u_1, \ldots, u_i, \ldots, u_n) - f(u_1, \ldots, v_i, \ldots, u_n) \\
&\qquad = f(u_1, \ldots, u_i, \ldots, u_n) + f(u_1, \ldots, v_i, \ldots, u_n) - f(u_1, \ldots, u_i, \ldots, u_n) \\
&\qquad\quad - f(u_1, \ldots, v_i, \ldots, u_n) \\
&\qquad = 0.
\end{aligned}$$

But then, $\overline{f}\colon M \to F$ factors through M/N, which means that there is a unique linear map $h\colon M/N \to F$ such that $\overline{f} = h \circ \pi$ making the following diagram commute

$$\begin{tikzcd}
M \arrow[r, "\pi"] \arrow[dr, "\overline{f}"'] & M/N \arrow[d, "h"] \\
& F,
\end{tikzcd}$$

by defining $h([z]) = \overline{f}(z)$ for every $z \in M$, where $[z]$ denotes the equivalence class in M/N of $z \in M$. Indeed, the fact that $\overline{f}$ vanishes on N insures that h is well defined on M/N, and it is clearly linear by definition. Since $f = \overline{f} \circ \iota$, from the equation $\overline{f} = h \circ \pi$, by composing on the right with ι, we obtain

$$f = \overline{f} \circ \iota = h \circ \pi \circ \iota = h \circ \varphi,$$

as in the following commutative diagram.

$$\begin{tikzcd}
& K^{(E_1 \times \cdots \times E_n)} \arrow[dr, "\pi"] \arrow[dd, "\overline{f}"] & \\
E_1 \times \cdots \times E_n \arrow[ur, "\iota"] \arrow[dr, "f"'] & & K^{(E_1 \times \cdots \times E_n)}/N \arrow[dl, "h"] \\
& F &
\end{tikzcd}$$

We now prove the uniqueness of h. For any linear map $f_\otimes\colon E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$, since the vectors $u_1 \otimes \cdots \otimes u_n$ generate $E_1 \otimes \cdots \otimes E_n$ and since $\varphi(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n$, the map $f_\otimes$ is uniquely defined by

$$f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n).$$

Since $f = h \circ \varphi$, the map h is unique, and we let $f_\otimes = h$. $\square$

The map φ from $E_1 \times \cdots \times E_n$ to $E_1 \otimes \cdots \otimes E_n$ is often denoted by $\iota_\otimes$, so that

$$\iota_\otimes(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n.$$

What is important about Theorem 33.6 is not so much the construction itself but the fact that it produces a tensor product with the universal mapping property with respect to multilinear maps. Indeed, Theorem 33.6 yields a canonical isomorphism

$$\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F) \cong \mathrm{L}(E_1, \ldots, E_n; F)$$

between the vector space of linear maps $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$, and the vector space of multilinear maps $\mathcal{L}(E_1, \ldots, E_n; F)$, *via* the linear map $- \circ \varphi$ defined by

$$h \mapsto h \circ \varphi,$$

where $h \in \mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$. Indeed, $h \circ \varphi$ is clearly multilinear, and since by Theorem 33.6, for every multilinear map $f \in \mathcal{L}(E_1, \ldots, E_n; F)$, there is a unique linear map $f_\otimes \in \mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$ such that $f = f_\otimes \circ \varphi$, the map $- \circ \varphi$ is bijective. As a matter of fact, its inverse is the map

$$f \mapsto f_\otimes.$$

We record this fact as the following proposition.

Proposition 33.7. *Given a tensor product $(E_1 \otimes \cdots \otimes E_n, \varphi)$, the linear map $h \mapsto h \circ \varphi$ is a canonical isomorphism*

$$\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F) \cong \mathrm{L}(E_1, \ldots, E_n; F)$$

between the vector space of linear maps $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$, and the vector space of multilinear maps $\mathcal{L}(E_1, \ldots, E_n; F)$.

Using the "Hom" notation, the above canonical isomorphism is written

$$\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, F) \cong \mathrm{Hom}(E_1, \ldots, E_n; F).$$

Remarks:

(1) To be very precise, since the tensor product depends on the field K, we should subscript the symbol $\otimes$ with K and write

$$E_1 \otimes_K \cdots \otimes_K E_n.$$

However, we often omit the subscript K unless confusion may arise.

(2) For $F = K$, the base field, Proposition 33.7 yields a canonical isomorphism between the vector space $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, K)$, and the vector space of multilinear forms $\mathcal{L}(E_1, \ldots, E_n; K)$. However, $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, K)$ is the dual space $(E_1 \otimes \cdots \otimes E_n)^*$, and thus the vector space of multilinear forms $\mathcal{L}(E_1, \ldots, E_n; K)$ is canonically isomorphic to $(E_1 \otimes \cdots \otimes E_n)^*$.

Since this isomorphism is used often, we record it as the following proposition.

Proposition 33.8. *Given a tensor product $E_1 \otimes \cdots \otimes E_n$,, there is a canonical isomorphism*

$$\mathrm{L}(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*$$

between the vector space of multilinear maps $\mathcal{L}(E_1, \ldots, E_n; K)$ and the dual $(E_1 \otimes \cdots \otimes E_n)^*$ of the tensor product $E_1 \otimes \cdots \otimes E_n$.

The fact that the map $\varphi \colon E_1 \times \cdots \times E_n \to E_1 \otimes \cdots \otimes E_n$ is multilinear, can also be expressed as follows:

$$\begin{aligned} u_1 \otimes \cdots \otimes (v_i + w_i) \otimes \cdots \otimes u_n &= (u_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes u_n) + (u_1 \otimes \cdots \otimes w_i \otimes \cdots \otimes u_n), \\ u_1 \otimes \cdots \otimes (\lambda u_i) \otimes \cdots \otimes u_n &= \lambda(u_1 \otimes \cdots \otimes u_i \otimes \cdots \otimes u_n). \end{aligned}$$

Of course, this is just what we wanted!

Definition 33.6. Tensors in $E_1 \otimes \cdots \otimes E_n$ are called *n-tensors*, and tensors of the form $u_1 \otimes \cdots \otimes u_n$, where $u_i \in E_i$ are called *simple (or decomposable) n-tensors*. Those n-tensors that are not simple are often called *compound n-tensors*.

Not only do tensor products act on spaces, but they also act on linear maps (they are functors).

Proposition 33.9. *Given two linear maps $f \colon E \to E'$ and $g \colon F \to F'$, there is a unique linear map*

$$f \otimes g \colon E \otimes F \to E' \otimes F'$$

such that

$$(f \otimes g)(u \otimes v) = f(u) \otimes g(v),$$

for all $u \in E$ and all $v \in F$.

Proof. We can define $h \colon E \times F \to E' \otimes F'$ by

$$h(u, v) = f(u) \otimes g(v).$$

It is immediately verified that h is bilinear, and thus it induces a unique linear map

$$f \otimes g \colon E \otimes F \to E' \otimes F'$$

making the following diagram commutes

$$\begin{tikzcd} E \times F \arrow[r, "\iota_\otimes"] \arrow[dr, "h"'] & E \otimes F \arrow[d, "f \otimes g"] \\ & E' \otimes F', \end{tikzcd}$$

such that $(f \otimes g)(u \otimes v) = f(u) \otimes g(v)$, for all $u \in E$ and all $v \in F$. $\square$

Definition 33.7. The linear map $f \otimes g\colon E \otimes F \to E' \otimes F'$ given by Proposition 33.9 is called the *tensor product* of $f\colon E \to E'$ and $g\colon F \to F'$.

Another way to define $f \otimes g$ proceeds as follows. Given two linear maps $f\colon E \to E'$ and $g\colon F \to F'$, the map $f \times g$ is the linear map from $E \times F$ to $E' \times F'$ given by

$$(f \times g)(u, v) = (f(u), g(v)), \quad \text{for all } u \in E \text{ and all } v \in F.$$

Then the map h in the proof of Proposition 33.9 is given by $h = \iota'_\otimes \circ (f \times g)$, and $f \otimes g$ is the unique linear map making the following diagram commute.

$$\begin{tikzcd}
E \times F \arrow[r, "\iota_\otimes"] \arrow[d, "f\times g"'] & E \otimes F \arrow[d, "f\otimes g"] \\
E' \times F' \arrow[r, "\iota'_\otimes"'] & E' \otimes F'
\end{tikzcd}$$

Remark: The notation $f \otimes g$ is potentially ambiguous, because $\mathrm{Hom}(E, F)$ and $\mathrm{Hom}(E', F')$ are vector spaces, so we can form the tensor product $\mathrm{Hom}(E, F) \otimes \mathrm{Hom}(E', F')$ which contains elements also denoted $f \otimes g$. To avoid confusion, the first kind of tensor product of linear maps defined in Proposition 33.9 (which yields a linear map in $\mathrm{Hom}(E \otimes F, E' \otimes F')$) can be denoted by $T(f, g)$. If we denote the tensor product $E \otimes F$ by $T(E, F)$, this notation makes it clearer that T is a bifunctor. If E, E' and F, F' are finite dimensional, by picking bases it is not hard to show that the map induced by $f \otimes g \mapsto T(f, g)$ is an isomorphism

$$\mathrm{Hom}(E, F) \otimes \mathrm{Hom}(E', F') \cong \mathrm{Hom}(E \otimes F, E' \otimes F').$$

Proposition 33.10. *Suppose we have linear maps $f\colon E \to E'$, $g\colon F \to F'$, $f'\colon E' \to E''$ and $g'\colon F' \to F''$. Then the following identity holds:*

$$(f' \circ f) \otimes (g' \circ g) = (f' \otimes g') \circ (f \otimes g). \tag{$*$}$$

Proof. We have the commutative diagram

$$\begin{tikzcd}
E \times F \arrow[r, "\iota_\otimes"] \arrow[d, "f\times g"'] & E \otimes F \arrow[d, "f\otimes g"] \\
E' \times F' \arrow[r, "\iota'_\otimes"] \arrow[d, "f'\times g'"'] & E' \otimes F' \arrow[d, "f'\otimes g'"] \\
E'' \times F'' \arrow[r, "\iota''_\otimes"'] & E'' \otimes F'',
\end{tikzcd}$$

and thus the commutative diagram.

$$\begin{tikzcd}
E \times F \arrow[r, "\iota_\otimes"] \arrow[d, "(f'\times g')\circ(f\times g)"'] & E \otimes F \arrow[d, "(f'\otimes g')\circ(f\otimes g)"] \\
E'' \times F'' \arrow[r, "\iota''_\otimes"'] & E'' \otimes F''
\end{tikzcd}$$

We also have the commutative diagram.

$$\begin{tikzcd}
E \times F \arrow[r, "\iota_\otimes"] \arrow[d, "(f' \circ f) \times (g' \circ g)"'] & E \otimes F \arrow[d, "(f' \circ f) \otimes (g' \circ g)"] \\
E'' \times F'' \arrow[r, "\iota''_\otimes"'] & E'' \otimes F''.
\end{tikzcd}$$

Since we immediately verify that

$$(f' \circ f) \times (g' \circ g) = (f' \times g') \circ (f \times g),$$

by uniqueness of the map between $E \otimes F$ and $E'' \otimes F''$ in the above diagram, we conclude that

$$(f' \circ f) \otimes (g' \circ g) = (f' \otimes g') \circ (f \otimes g),$$

as claimed. $\square$

The above formula $(*)$ yields the following useful fact.

Proposition 33.11. *If $f\colon E \to E'$ and $g\colon F \to F'$ are isomorphims, then $f \otimes g\colon E \otimes F \to E' \otimes F'$ is also an isomorphism.*

Proof. If $f^{-1}\colon E' \to E$ is the inverse of $f\colon E \to E'$ and $g^{-1}\colon F' \to F$ is the inverse of $g\colon F \to F'$, then $f^{-1} \otimes g^{-1}\colon E' \otimes F' \to E \otimes F$ is the inverse of $f \otimes g\colon E \otimes F \to E' \otimes F'$, which is shown as follows:

$$\begin{aligned}
(f \otimes g) \circ (f^{-1} \otimes g^{-1}) &= (f \circ f^{-1}) \otimes (g \circ g^{-1}) \\
&= \mathrm{id}_{E'} \otimes \mathrm{id}_{F'} \\
&= \mathrm{id}_{E' \otimes F'},
\end{aligned}$$

and

$$\begin{aligned}
(f^{-1} \otimes g^{-1}) \circ (f \otimes g) &= (f^{-1} \circ f) \otimes (g^{-1} \circ g) \\
&= \mathrm{id}_{E} \otimes \mathrm{id}_{F} \\
&= \mathrm{id}_{E \otimes F}.
\end{aligned}$$

Therefore, $f \otimes g\colon E \otimes F \to E' \otimes F'$ is an isomorphism. $\square$

The generalization to the tensor product $f_1 \otimes \cdots \otimes f_n$ of $n \geq 3$ linear maps $f_i\colon E_i \to F_i$ is immediate, and left to the reader.

33.3 Bases of Tensor Products

We showed that $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors of the form $u_1 \otimes \cdots \otimes u_n$. However, these vectors are not linearly independent. This situation can be fixed when considering bases.

To explain the idea of the proof, consider the case when we have two spaces E and F both of dimension 3. Given a basis (e_1, e_2, e_3) of E and a basis (f_1, f_2, f_3) of F, we would like to prove that

$$e_1 \otimes f_1, \quad e_1 \otimes f_2, \quad e_1 \otimes f_3, \quad e_2 \otimes f_1, \quad e_2 \otimes f_2, \quad e_2 \otimes f_3, \quad e_3 \otimes f_1, \quad e_3 \otimes f_2, \quad e_3 \otimes f_3$$

are linearly independent. To prove this, it suffices to show that for any vector space G, if $w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33}$ are any vectors in G, then there is a bilinear map $h\colon E \times F \to G$ such that

$$h(e_i, e_j) = w_{ij}, \quad 1 \leq i, j \leq 3.$$

Because h yields a unique linear map $h_\otimes\colon E \otimes F \to G$ such that

$$h_\otimes(e_i \otimes e_j) = w_{ij}, \quad 1 \leq i, j \leq 3,$$

and by Proposition 33.4, the vectors

$$e_1 \otimes f_1, \quad e_1 \otimes f_2, \quad e_1 \otimes f_3, \quad e_2 \otimes f_1, \quad e_2 \otimes f_2, \quad e_2 \otimes f_3, \quad e_3 \otimes f_1, \quad e_3 \otimes f_2, \quad e_3 \otimes f_3$$

are linearly independent. This suggests understanding how a bilinear function $f\colon E \times F \to G$ is expressed in terms of its values $f(e_i, f_j)$ on the basis vectors (e_1, e_2, e_3) and (f_1, f_2, f_3), and this can be done easily. Using bilinearity we obtain

$$\begin{aligned} f(u_1e_1 + u_2e_2 + u_3e_3, v_1f_1 + v_2f_2 + v_3f_3) &= u_1v_1f(e_1, f_1) + u_1v_2f(e_1, f_2) + u_1v_3f(e_1, f_3) \\ &\quad + u_2v_1f(e_2, f_1) + u_2v_2f(e_2, f_2) + u_2v_3f(e_2, f_3) \\ &\quad + u_3v_1f(e_3, f_1) + u_3v_2f(e_3, f_2) + u_3v_3f(e_3, f_3). \end{aligned}$$

Therefore, given $w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33} \in G$, the function h given by

$$\begin{aligned} h(u_1e_1 + u_2e_2 + u_3e_3, v_1f_1 + v_2f_2 + v_3f_3) &= u_1v_1w_{11} + u_1v_2w_{12} + u_1v_3w_{13} \\ &\quad + u_2v_1w_{21} + u_2v_2w_{22} + u_2v_3w_{23} \\ &\quad + u_3v_1w_{31} + u_3v_2w_{33} + u_3v_3w_{33} \end{aligned}$$

is clearly bilinear, and by construction $h(e_i, f_j) = w_{ij}$, so it does the job.

The generalization of this argument to any number of vector spaces of any dimension (even infinite) is straightforward.

Proposition 33.12. *Given $n \geq 2$ vector spaces $E_1, \ldots, E_n$, if $(u^k_i)_{i\in I_k}$ is a basis for E_k, $1 \leq k \leq n$, then the family of vectors*

$$(u^1_{i_1} \otimes \cdots \otimes u^n_{i_n})_{(i_1,\ldots,i_n)\in I_1\times\ldots\times I_n}$$

is a basis of the tensor product $E_1 \otimes \cdots \otimes E_n$.

Proof. For each k, $1 \leq k \leq n$, every $v^k \in E_k$ can be written uniquely as

$$v^k = \sum_{j \in I_k} v^k_j u^k_j,$$

for some family of scalars $(v^k_j)_{j \in I_k}$. Let F be any nontrivial vector space. We show that for every family

$$(w_{i_1,\ldots,i_n})_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n},$$

of vectors in F, there is some linear map $h \colon E_1 \otimes \cdots \otimes E_n \to F$ such that

$$h(u^1_{i_1} \otimes \cdots \otimes u^n_{i_n}) = w_{i_1,\ldots,i_n}.$$

Then by Proposition 33.4, it follows that

$$(u^1_{i_1} \otimes \cdots \otimes u^n_{i_n})_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n}$$

is linearly independent. However, since $(u^k_i)_{i \in I_k}$ is a basis for E_k, the $u^1_{i_1} \otimes \cdots \otimes u^n_{i_n}$ also generate $E_1 \otimes \cdots \otimes E_n$, and thus, they form a basis of $E_1 \otimes \cdots \otimes E_n$.

We define the function $f \colon E_1 \times \cdots \times E_n \to F$ as follows: For any n nonempty finite subsets $J_1, \ldots, J_n$ such that $J_k \subseteq I_k$ for $k = 1, \ldots, n$,

$$f(\sum_{j_1 \in J_1} v^1_{j_1} u^1_{j_1}, \ldots, \sum_{j_n \in J_n} v^n_{j_n} u^n_{j_n}) = \sum_{j_1 \in J_1, \ldots, j_n \in J_n} v^1_{j_1} \cdots v^n_{j_n} \, w_{j_1,\ldots,j_n}.$$

It is immediately verified that f is multilinear. By the universal mapping property of the tensor product, the linear map $f_\otimes \colon E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$, is the desired map h. $\square$

In particular, when each I_k is finite and of size $m_k = \dim(E_k)$, we see that the dimension of the tensor product $E_1 \otimes \cdots \otimes E_n$ is $m_1 \cdots m_n$. As a corollary of Proposition 33.12, if $(u^k_i)_{i \in I_k}$ is a basis for E_k, $1 \leq k \leq n$, then every tensor $z \in E_1 \otimes \cdots \otimes E_n$ can be written in a unique way as

$$z = \sum_{(i_1,\ldots,i_n) \, \in \, I_1 \times \ldots \times I_n} \lambda_{i_1,\ldots,i_n} \, u^1_{i_1} \otimes \cdots \otimes u^n_{i_n},$$

for some unique family of scalars $\lambda_{i_1,\ldots,i_n} \in K$, all zero except for a finite number.

33.4 Some Useful Isomorphisms for Tensor Products

Proposition 33.13. *Given three vector spaces E, F, G, there exists unique canonical isomorphisms*

(1) $E \otimes F \cong F \otimes E$

(2) $(E \otimes F) \otimes G \cong E \otimes (F \otimes G) \cong E \otimes F \otimes G$

(3) $(E \oplus F) \otimes G \cong (E \otimes G) \oplus (F \otimes G)$

(4) $K \otimes E \cong E$

such that respectively

(a) $u \otimes v \mapsto v \otimes u$
(b) $(u \otimes v) \otimes w \mapsto u \otimes (v \otimes w) \mapsto u \otimes v \otimes w$
(c) $(u,\, v) \otimes w \mapsto (u \otimes w,\, v \otimes w)$
(d) $\lambda \otimes u \mapsto \lambda u$.

Proof. Except for (3), these isomorphisms are proved using the universal mapping property of tensor products.

(1) The map from $E \times F$ to $F \otimes E$ given by $(u, v) \mapsto v \otimes u$ is clearly bilinear, thus it induces a unique linear $\alpha\colon E \otimes F \to F \otimes E$ making the following diagram commute

$$\begin{tikzcd} E \times F \arrow[r, "\iota_\otimes"] \arrow[dr] & E \otimes F \arrow[d, "\alpha"] \\ & F \otimes E, \end{tikzcd}$$

such that

$$\alpha(u \otimes v) = v \otimes u, \quad \text{for all } u \in E \text{ and all } v \in F.$$

Similarly, the map from $F \times E$ to $E \otimes F$ given by $(v, u) \mapsto u \otimes v$ is clearly bilinear, thus it induces a unique linear $\beta\colon F \otimes E \to E \otimes F$ making the following diagram commute

$$\begin{tikzcd} F \times E \arrow[r, "\iota_\otimes"] \arrow[dr] & F \otimes E \arrow[d, "\beta"] \\ & E \otimes F, \end{tikzcd}$$

such that

$$\beta(v \otimes u) = u \otimes v, \quad \text{for all } u \in E \text{ and all } v \in F.$$

It is immediately verified that

$$(\beta \circ \alpha)(u \otimes v) = u \otimes v \quad \text{and} \quad (\alpha \circ \beta)(v \otimes u) = v \otimes u$$

for all $u \in E$ and all $v \in F$. Since the tensors of the form $u \otimes v$ span $E \otimes F$ and similarly the tensors of the form $v \otimes u$ span $F \otimes E$, the map $\beta \circ \alpha$ is actually the identity on $E \otimes F$, and similarly $\alpha \circ \beta$ is the identity on $F \otimes E$, so α and β are isomorphisms.

(2) Fix some $w \in G$. The map

$$(u,\, v) \mapsto u \otimes v \otimes w$$

from $E \times F$ to $E \otimes F \otimes G$ is bilinear, and thus there is a linear map $f_w \colon E \otimes F \to E \otimes F \otimes G$ making the following diagram commute

$$\begin{tikzcd}
E \times F \arrow[r, "\iota_\otimes"] \arrow[dr] & E \otimes F \arrow[d, "f_w"] \\
& E \otimes F \otimes G,
\end{tikzcd}$$

with $f_w(u \otimes v) = u \otimes v \otimes w$.

Next consider the map

$$(z,\, w) \mapsto f_w(z),$$

from $(E \otimes F) \times G$ into $E \otimes F \otimes G$. It is easily seen to be bilinear, and thus it induces a linear map $f \colon (E \otimes F) \otimes G \to E \otimes F \otimes G$ making the following diagram commute

$$\begin{tikzcd}
(E \otimes F) \times G \arrow[r, "\iota_\otimes"] \arrow[dr] & (E \otimes F) \otimes G \arrow[d, "f"] \\
& E \otimes F \otimes G,
\end{tikzcd}$$

with $f((u \otimes v) \otimes w) = u \otimes v \otimes w$.

Also consider the map

$$(u,\, v, w) \mapsto (u \otimes v) \otimes w$$

from $E \times F \times G$ to $(E \otimes F) \otimes G$. It is trilinear, and thus there is a linear map $g \colon E \otimes F \otimes G \to (E \otimes F) \otimes G$ making the following diagram commute

$$\begin{tikzcd}
E \times F \times G \arrow[r, "\iota_\otimes"] \arrow[dr] & E \otimes F \otimes G \arrow[d, "g"] \\
& (E \otimes F) \otimes G,
\end{tikzcd}$$

with $g(u \otimes v \otimes w) = (u \otimes v) \otimes w$. Clearly, $f \circ g$ and $g \circ f$ are identity maps, and thus f and g are isomorphisms. The other case is similar.

(3) Given a fixed vector space G, for any two vector spaces M and N and every linear map $f \colon M \to N$, let $\tau_G(f) = f \otimes \mathrm{id}_G$ be the unique linear map making the following diagram commute.

$$\begin{tikzcd}
M \times G \arrow[r, "\iota_{M\otimes}"] \arrow[d, "f \times \mathrm{id}_G"'] & M \otimes G \arrow[d, "f \otimes \mathrm{id}_G"] \\
N \times G \arrow[r, "\iota_{N\otimes}"'] & N \otimes G
\end{tikzcd}$$

The identity $(*)$ proved in Proposition 33.10 shows that if $g\colon N \to P$ is another linear map, then

$$\tau_G(g) \circ \tau_G(f) = (g \otimes \mathrm{id}_G) \circ (f \otimes \mathrm{id}_G) = (g \circ f) \otimes (\mathrm{id}_G \circ \mathrm{id}_G) = (g \circ f) \otimes \mathrm{id}_G = \tau_G(g \circ f).$$

Clearly, $\tau_G(0) = 0$, and a direct computation on generators also shows that

$$\tau_G(\mathrm{id}_M) = (\mathrm{id}_M \otimes \mathrm{id}_G) = \mathrm{id}_{M \otimes G},$$

and that if $f'\colon M \to N$ is another linear map, then

$$\tau_G(f + f') = \tau_G(f) + \tau_G(f').$$

In fancy terms, τ_G is a functor. Now, if $E \oplus F$ is a direct sum, it is a standard fact of linear algebra that if $\pi_E\colon E \oplus F \to E$ and $\pi_F\colon E \oplus F \to F$ are the projection maps, then

$$\pi_E \circ \pi_E = \pi_E \qquad \pi_F \circ \pi_F = \pi_F \qquad \pi_E \circ \pi_F = 0 \qquad \pi_F \circ \pi_E = 0 \qquad \pi_E + \pi_F = \mathrm{id}_{E \oplus F}.$$

If we apply τ_G to these identites, we get

$$\begin{array}{lll} \tau_G(\pi_E) \circ \tau_G(\pi_E) = \tau_G(\pi_E) & \tau_G(\pi_F) \circ \tau_G(\pi_F) = \tau_G(\pi_F) & \\ \tau_G(\pi_E) \circ \tau_G(\pi_F) = 0 & \tau_G(\pi_F) \circ \tau_G(\pi_E) = 0 & \tau_G(\pi_E) + \tau_G(\pi_F) = \mathrm{id}_{(E \oplus F) \otimes G}. \end{array}$$

Observe that $\tau_G(\pi_E) = \pi_E \otimes \mathrm{id}_G$ is a map from $(E \oplus F) \otimes G$ onto $E \otimes G$ and that $\tau_G(\pi_F) = \pi_F \otimes \mathrm{id}_G$ is a map from $(E \oplus F) \otimes G$ onto $F \otimes G$, and by linear algebra, the above equations mean that we have a direct sum

$$(E \otimes G) \oplus (F \otimes G) \cong (E \oplus F) \otimes G.$$

(4) We have the linear map $\epsilon\colon E \to K \otimes E$ given by

$$\epsilon(u) = 1 \otimes u, \quad \text{for all } u \in E.$$

The map $(\lambda, u) \mapsto \lambda u$ from $K \times E$ to E is bilinear, so it induces a unique linear map $\eta\colon K \otimes E \to E$ making the following diagram commute

$$\begin{tikzcd} K \times E \arrow[r, "\iota_\otimes"] \arrow[dr] & K \otimes E \arrow[d, "\eta"] \\ & E, \end{tikzcd}$$

such that $\eta(\lambda \otimes u) = \lambda u$, for all $\lambda \in K$ and all $u \in E$. We have

$$(\eta \circ \epsilon)(u) = \eta(1 \otimes u) = 1u = u,$$

and

$$(\epsilon \circ \eta)(\lambda \otimes u) = \epsilon(\lambda u) = 1 \otimes (\lambda u) = \lambda(1 \otimes u) = \lambda \otimes u,$$

which shows that both $\epsilon \circ \eta$ and $\eta \circ \epsilon$ are the identity, so ϵ and η are isomorphisms. $\square$

Remark: The isomorphism (3) can be generalized to finite and even arbitrary direct sums $\bigoplus_{i\in I} E_i$ of vector spaces (where I is an arbitrary nonempty index set). We have an isomorphism

$$\left(\bigoplus_{i\in I} E_i\right) \otimes G \cong \bigoplus_{i\in I} (E_i \otimes G).$$

This isomorphism (with isomorphism (1)) can be used to give another proof of Proposition 33.12 (see Bertin [15], Chapter 4, Section 1) or Lang [109], Chapter XVI, Section 2).

Proposition 33.14. *Given any three vector spaces E, F, G, we have the canonical isomorphism*

$$\mathrm{Hom}(E, F; G) \cong \mathrm{Hom}(E, \mathrm{Hom}(F, G)).$$

Proof. Any bilinear map $f\colon E \times F \to G$ gives the linear map $\varphi(f) \in \mathrm{Hom}(E, \mathrm{Hom}(F, G))$, where $\varphi(f)(u)$ is the linear map in $\mathrm{Hom}(F, G)$ given by

$$\varphi(f)(u)(v) = f(u, v).$$

Conversely, given a linear map $g \in \mathrm{Hom}(E, \mathrm{Hom}(F, G))$, we get the bilinear map $\psi(g)$ given by

$$\psi(g)(u, v) = g(u)(v),$$

and it is clear that φ and ψ and mutual inverses. $\square$

Since by Proposition 33.7 there is a canonical isomorphism

$$\mathrm{Hom}(E \otimes F, G) \cong \mathrm{Hom}(E, F; G),$$

together with the isomorphism

$$\mathrm{Hom}(E, F; G) \cong \mathrm{Hom}(E, \mathrm{Hom}(F, G))$$

given by Proposition 33.14, we obtain the important corollary:

Proposition 33.15. *For any three vector spaces E, F, G, we have the canonical isomorphism*

$$\mathrm{Hom}(E \otimes F, G) \cong \mathrm{Hom}(E, \mathrm{Hom}(F, G)).$$

33.5 Duality for Tensor Products

In this section all vector spaces are assumed to have *finite dimension*, unless specified otherwise. Let us now see how tensor products behave under duality. For this, we define a pairing between $E_1^* \otimes \cdots \otimes E_n^*$ and $E_1 \otimes \cdots \otimes E_n$ as follows: For any fixed $(v_1^*, \ldots, v_n^*) \in E_1^* \times \cdots \times E_n^*$, we have the multilinear map

$$l_{v_1^*,\ldots,v_n^*}\colon (u_1, \ldots, u_n) \mapsto v_1^*(u_1) \cdots v_n^*(u_n)$$

from $E_1 \times \cdots \times E_n$ to K. The map $l_{v_1^*,\ldots,v_n^*}$ extends uniquely to a linear map $L_{v_1^*,\ldots,v_n^*} \colon E_1 \otimes \cdots \otimes E_n \longrightarrow K$ making the following diagram commute.

$$\begin{tikzcd}
E_1 \times \cdots \times E_n \arrow[r, "\iota_\otimes"] \arrow[dr, "l_{v_1^*,\ldots,v_n^*}"'] & E_1 \otimes \cdots \otimes E_n \arrow[d, "L_{v_1^*,\ldots,v_n^*}"] \\
& K
\end{tikzcd}$$

We also have the multilinear map

$$(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*,\ldots,v_n^*}$$

from $E_1^* \times \cdots \times E_n^*$ to $\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, K)$, which extends to a unique linear map L from $E_1^* \otimes \cdots \otimes E_n^*$ to $\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, K)$ making the following diagram commute.

$$\begin{tikzcd}
E_1^* \times \cdots \times E_n^* \arrow[r, "\iota_\otimes"] \arrow[dr, "L_{v_1^*,\ldots,v_n^*}"'] & E_1^* \otimes \cdots \otimes E_n^* \arrow[d, "L"] \\
& \mathrm{Hom}(E_1 \otimes \cdots \otimes E_n; K)
\end{tikzcd}$$

However, in view of the isomorphism

$$\mathrm{Hom}(U \otimes V, W) \cong \mathrm{Hom}(U, \mathrm{Hom}(V, W))$$

given by Proposition 33.15, with $U = E_1^* \otimes \cdots \otimes E_n^*$, $V = E_1 \otimes \cdots \otimes E_n$ and $W = K$, we can view L as a linear map

$$L \colon (E_1^* \otimes \cdots \otimes E_n^*) \otimes (E_1 \otimes \cdots \otimes E_n) \to K,$$

which corresponds to a bilinear map

$$\langle -, - \rangle \colon (E_1^* \otimes \cdots \otimes E_n^*) \times (E_1 \otimes \cdots \otimes E_n) \longrightarrow K, \qquad (\dagger\dagger)$$

via the isomorphism $(U \otimes V)^* \cong \mathrm{Hom}(U, V; K)$ given by Proposition 33.8. This pairing is given explicitly on generators by

$$\langle v_1^* \otimes \cdots \otimes v_n^*, u_1 \ldots, u_n \rangle = v_1^*(u_1) \cdots v_n^*(u_n).$$

This pairing is nondegenerate, as proved below.

Proof. If $(e_1^1, \ldots, e_{m_1}^1), \ldots, (e_1^n, \ldots, e_{m_n}^n)$ are bases for $E_1, \ldots, E_n$, then for every basis element $(e_{i_1}^1)^* \otimes \cdots \otimes (e_{i_n}^n)^*$ of $E_1^* \otimes \cdots \otimes E_n^*$, and any basis element $e_{j_1}^1 \otimes \cdots \otimes e_{j_n}^n$ of $E_1 \otimes \cdots \otimes E_n$, we have

$$\langle (e_{i_1}^1)^* \otimes \cdots \otimes (e_{i_n}^n)^*, e_{j_1}^1 \otimes \cdots \otimes e_{j_n}^n \rangle = \delta_{i_1\, j_1} \cdots \delta_{i_n\, j_n},$$

where $\delta_{i\, j}$ is *Kronecker delta*, defined such that $\delta_{i\, j} = 1$ if $i = j$, and 0 otherwise. Given any $\alpha \in E_1^* \otimes \cdots \otimes E_n^*$, assume that $\langle \alpha, \beta \rangle = 0$ for all $\beta \in E_1 \otimes \cdots \otimes E_n$. The vector α is a finite linear combination $\alpha = \sum \lambda_{i_1,\ldots,i_n} (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*$, for some unique $\lambda_{i_1,\ldots,i_n} \in K$. If we choose $\beta = e^1_{i_1} \otimes \cdots \otimes e^n_{i_n}$, then we get

$$\begin{aligned} 0 = \langle \alpha, e^1_{i_1} \otimes \cdots \otimes e^n_{i_n} \rangle &= \Big\langle \sum \lambda_{i_1,\ldots,i_n} (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*, e^1_{i_1} \otimes \cdots \otimes e^n_{i_n} \Big\rangle \\ &= \sum \lambda_{i_1,\ldots,i_n} \langle (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*, e^1_{i_1} \otimes \cdots \otimes e^n_{i_n} \rangle \\ &= \lambda_{i_1,\ldots,i_n}. \end{aligned}$$

Therefore, $\alpha = 0$,

Conversely, given any $\beta \in E_1 \otimes \cdots \otimes E_n$, assume that $\langle \alpha, \beta \rangle = 0$, for all $\alpha \in E_1^* \otimes \cdots \otimes E_n^*$. The vector β is a finite linear combination $\beta = \sum \lambda_{i_1,\ldots,i_n} e^1_{i_1} \otimes \cdots \otimes e^n_{i_n}$, for some unique $\lambda_{i_1,\ldots,i_n} \in K$. If we choose $\alpha = (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*$, then we get

$$\begin{aligned} 0 = \langle (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*, \beta \rangle &= \Big\langle (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*, \sum \lambda_{i_1,\ldots,i_n} e^1_{i_1} \otimes \cdots \otimes e^n_{i_n} \Big\rangle \\ &= \sum \lambda_{i_1,\ldots,i_n} \langle (e^1_{i_1})^* \otimes \cdots \otimes (e^n_{i_n})^*, e^1_{i_1} \otimes \cdots \otimes e^n_{i_n} \rangle \\ &= \lambda_{i_1,\ldots,i_n}. \end{aligned}$$

Therefore, $\beta = 0$. ☐

By Proposition 33.1,[1] we have a canonical isomorphism

$$(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*.$$

Here is our main proposition about duality of tensor products.

Proposition 33.16. *We have canonical isomorphisms*

$$(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*,$$

and

$$\mu\colon E_1^* \otimes \cdots \otimes E_n^* \cong \mathrm{Hom}(E_1, \ldots, E_n; K).$$

Proof. The second isomorphism follows from the isomorphism $(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*$ together with the isomorphism $\mathrm{Hom}(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*$ given by Proposition 33.8. ☐

Remarks:

1. The isomorphism $\mu\colon E_1^* \otimes \cdots \otimes E_n^* \cong \mathrm{Hom}(E_1, \ldots, E_n; K)$ can be described explicitly as the linear extension to $E_1^* \otimes \cdots \otimes E_n^*$ of the map given by

$$\mu(v_1^* \otimes \cdots \otimes v_n^*)(u_1 \ldots, u_n) = v_1^*(u_1) \cdots v_n^*(u_n).$$

[1]This is where the assumption that our spaces are finite-dimensional is used.

2. The canonical isomorphism of Proposition 33.16 holds under more general conditions. Namely, that K is a commutative ring with identity and that the E_i are finitely-generated projective K-modules (see Definition 35.7). See Bourbaki, [25] (Chapter III, §11, Section 5, Proposition 7).

We prove another useful canonical isomorphism that allows us to treat linear maps as tensors.

Let E and F be two vector spaces and let $\alpha\colon E^* \times F \to \mathrm{Hom}(E, F)$ be the map defined such that

$$\alpha(u^*, f)(x) = u^*(x)f,$$

for all $u^* \in E^*$, $f \in F$, and $x \in E$. This map is clearly bilinear, and thus it induces a linear map $\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F)$ making the following diagram commute

$$\begin{tikzcd} E^* \times F \arrow[r, "\iota_\otimes"] \arrow[dr, "\alpha"'] & E^* \otimes F \arrow[d, "\alpha_\otimes"] \\ & \mathrm{Hom}(E, F), \end{tikzcd}$$

such that

$$\alpha_\otimes(u^* \otimes f)(x) = u^*(x)f.$$

Proposition 33.17. *If E and F are vector spaces (not necessarily finite dimensional), then the following properties hold:*

(1) The linear map $\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F)$ is injective.

(2) If E is finite-dimensional, then $\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F)$ is a canonical isomorphism.

(3) If F is finite-dimensional, then $\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F)$ is a canonical isomorphism.

Proof. (1) Let $(e_i^*)_{i \in I}$ be a basis of E^* and let $(f_j)_{j \in J}$ be a basis of F. Then we know that $(e_i^* \otimes f_j)_{i \in I, j \in J}$ is a basis of $E^* \otimes F$. To prove that $\alpha_\otimes$ is injective, let us show that its kernel is reduced to (0). For any vector

$$\omega = \sum_{i \in I', j \in J'} \lambda_{ij}\, e_i^* \otimes f_j$$

in $E^* \otimes F$, with I' and J' some finite sets, assume that $\alpha_\otimes(\omega) = 0$. This means that for every $x \in E$, we have $\alpha_\otimes(\omega)(x) = 0$; that is,

$$\sum_{i \in I', j \in J'} \alpha_\otimes(\lambda_{ij}\, e_i^* \otimes f_j)(x) = \sum_{j \in J'} \left(\sum_{i \in I'} \lambda_{ij} e_i^*(x) \right) f_j = 0.$$

Since $(f_j)_{j\in J}$ is a basis of F, for every $j \in J'$, we must have

$$\sum_{i\in I'} \lambda_{ij} e_i^*(x) = 0, \quad \text{for all } x \in E.$$

But then $(e_i^*)_{i\in I'}$ would be linearly dependent, contradicting the fact that $(e_i^*)_{i\in I}$ is a basis of E^*, so we must have

$$\lambda_{ij} = 0, \quad \text{for all } i \in I' \text{ and all } j \in J',$$

which shows that $\omega = 0$. Therefore, $\alpha_\otimes$ is injective.

(2) Let $(e_j)_{1\leq j\leq n}$ be a finite basis of E, and as usual, let $e_j^* \in E^*$ be the linear form defined by

$$e_j^*(e_k) = \delta_{j,k},$$

where $\delta_{j,k} = 1$ iff $j = k$ and 0 otherwise. We know that $(e_j^*)_{1\leq j\leq n}$ is a basis of E^* (this is where we use the finite dimension of E). For any linear map $f \in \mathrm{Hom}(E, F)$, for every $x = x_1e_1 + \cdots + x_ne_n \in E$, we have

$$f(x) = f(x_1e_1 + \cdots + x_ne_n) = x_1f(e_1) + \cdots + x_nf(e_n) = e_1^*(x)f(e_1) + \cdots + e_n^*(x)f(e_n).$$

Consequently, every linear map $f \in \mathrm{Hom}(E, F)$ can be expressed as

$$f(x) = e_1^*(x)f_1 + \cdots + e_n^*(x)f_n,$$

for some $f_i \in F$. Furthermore, if we apply f to e_i, we get $f(e_i) = f_i$, so the f_i are unique. Observe that

$$(\alpha_\otimes(e_1^* \otimes f_1 + \cdots + e_n^* \otimes f_n))(x) = \sum_{i=1}^{n} (\alpha_\otimes(e_i^* \otimes f_i))(x) = \sum_{i=1}^{n} e_i^*(x)f_i.$$

Thus, $\alpha_\otimes$ is surjective, so $\alpha_\otimes$ is a bijection.

(3) Let $(f_1, \ldots, f_m)$ be a finite basis of F, and let $(f_1^*, \ldots, f_m^*)$ be its dual basis. Given any linear map $h\colon E \to F$, for all $u \in E$, since $f_i^*(f_j) = \delta_{ij}$, we have

$$h(u) = \sum_{i=1}^{m} f_i^*(h(u))f_i.$$

If

$$h(u) = \sum_{j=1}^{m} v_j^*(u)f_j \quad \text{for all } u \in E \tag{$*$}$$

for some linear forms $(v_1^*, \ldots, v_m^*) \in (E^*)^m$, then

$$f_i^*(h(u)) = \sum_{j=1}^{m} v_j^*(u)f_i^*(f_j) = v_i^*(u) \quad \text{for all } u \in E,$$

which shows that $v_i^* = f_i^* \circ h$ for $i = 1, \ldots, m$. This means that h has a unique expression in terms of linear forms as in $(*)$. Define the map α from $(E^*)^m$ to $\mathrm{Hom}(E, F)$ by

$$\alpha(v_1^*, \ldots, v_m^*)(u) = \sum_{j=1}^{m} v_j^*(u) f_j \quad \text{for all } u \in E.$$

This map is linear. For any $h \in \mathrm{Hom}(E, F)$, we showed earlier that the expression of h in $(*)$ is unique, thus α is an isomorphism. Similarly, $E^* \otimes F$ is isomorphic to $(E^*)^m$. Any tensor $\omega \in E^* \otimes F$ can be written as a linear combination

$$\sum_{k=1}^{p} u_k^* \otimes y_k$$

for some $u_k^* \in E^*$ and some $y_k \in F$, and since $(f_1, \ldots, f_m)$ is a basis of F, each y_k can be written as a linear combination of $(f_1, \ldots, f_m)$, so ω can be expressed as

$$\omega = \sum_{i=1}^{m} v_i^* \otimes f_i, \tag{$\dagger$}$$

for some linear forms $v_i^* \in E^*$ which are linear combinations of the u_k^*. If we pick a basis $(w_i^*)_{i \in I}$ for E^*, then we know that the family $(w_i^* \otimes f_j)_{i \in I, 1 \leq j \leq m}$ is a basis of $E^* \otimes F$, and this implies that the v_i^* in $(\dagger)$ are unique. Define the linear map β from $(E^*)^m$ to $E^* \otimes F$ by

$$\beta(v_1^*, \ldots, v_m^*) = \sum_{i=1}^{m} v_i^* \otimes f_i.$$

Since every tensor $\omega \in E^* \otimes F$ can be written in a unique way as in $(\dagger)$, this map is an isomorphism. $\square$

Note that in Proposition 33.17, we have an isomorphism if either E or F has finite dimension. The following proposition allows us to view a multilinear as a tensor product.

Proposition 33.18. *If the $E_1, \ldots E_n$ are finite-dimensional vector spaces and F is any vector space, then we have the canonical isomorphism*

$$\mathrm{Hom}(E_1, \ldots, E_n; F) \cong E_1^* \otimes \cdots \otimes E_n^* \otimes F.$$

Proof. In view of the canonical isomorphism

$$\mathrm{Hom}(E_1, \ldots, E_n; F) \cong \mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, F)$$

given by Proposition 33.7 and the canonical isomorphism $(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*$ given by Proposition 33.16, if the E_i's are finite-dimensional, then Proposition 33.17 yields the canonical isomorphism

$$\mathrm{Hom}(E_1, \ldots, E_n; F) \cong E_1^* \otimes \cdots \otimes E_n^* \otimes F,$$

as claimed. $\square$

33.6 Tensor Algebras

Our goal is to define a vector space $T(V)$ obtained by taking the direct sum of the tensor products

$$\underbrace{V \otimes \cdots \otimes V}_{m},$$

and to define a multiplication operation on $T(V)$ which makes $T(V)$ into an algebraic structure called an algebra. The algebra $T(V)$ satisfies a universal property stated in Proposition 33.19, which makes it the "free algebra" generated by the vector space V.

Definition 33.8. The tensor product

$$\underbrace{V \otimes \cdots \otimes V}_{m}$$

is also denoted as

$$\bigotimes^{m} V \quad \text{or} \quad V^{\otimes m}$$

and is called the *m-th tensor power of* V (with $V^{\otimes 1} = V$, and $V^{\otimes 0} = K$).

We can pack all the tensor powers of V into the "big" vector space

$$T(V) = \bigoplus_{m \geq 0} V^{\otimes m},$$

denoted $T^{\bullet}(V)$ or $\bigotimes V$ to avoid confusion with the tangent bundle.

This is an interesting object because we can define a multiplication operation on it which makes it into an *algebra*.

When V is of finite dimension n, we can pick some basis $(e_1 \ldots, e_n)$ of V, and then every tensor $\omega \in T(V)$ can be expressed as a linear combination of terms of the form $e_{i_1} \otimes \cdots \otimes e_{i_k}$, where $(i_1, \ldots, i_k)$ is any sequence of elements from the set $\{1, \ldots, n\}$. We can think of the tensors $e_{i_1} \otimes \cdots \otimes e_{i_k}$ as monomials in the noncommuting variables $e_1, \ldots, e_n$. Thus the space $T(V)$ corresponds to the algebra of polynomials with coefficients in K in n *noncommuting variables.*

Let us review the definition of an algebra over a field. Let K denote any (commutative) field, although for our purposes, we may assume that $K = \mathbb{R}$ (and occasionally, $K = \mathbb{C}$). Since we will only be dealing with associative algebras with a multiplicative unit, we only define algebras of this kind.

Definition 33.9. Given a field K, a K*-algebra* is a K-vector space A together with a bilinear operation $\cdot \colon A \times A \to A$, called *multiplication*, which makes A into a ring with unity 1 (or 1_A, when we want to be very precise). This means that $\cdot$ is associative and that there is a multiplicative identity element 1 so that $1 \cdot a = a \cdot 1 = a$, for all $a \in A$. Given two

K-algebras A and B, a *K-algebra homomorphism* $h\colon A \to B$ is a linear map that is also a ring homomorphism, with $h(1_A) = 1_B$; that is,

$$\begin{aligned} h(a_1 \cdot a_2) &= h(a_1) \cdot h(a_2) \quad \text{for all } a_1, a_2 \in A \\ h(1_A) &= 1_B. \end{aligned}$$

The set of K-algebra homomorphisms between A and B is denoted $\mathrm{Hom}_{\mathrm{alg}}(A, B)$.

For example, the ring $\mathrm{M}_n(K)$ of all $n \times n$ matrices over a field K is a K-algebra.

There is an obvious notion of ideal of a K-algebra.

Definition 33.10. Let A be a K-algebra. An *ideal* $\mathfrak{A} \subseteq A$ is a linear subspace of A that is also a two-sided ideal with respect to multiplication in A; this means that for all $a \in \mathfrak{A}$ and all $\alpha, \beta \in A$, we have $\alpha a \beta \in \mathfrak{A}$.

If the field K is understood, we usually simply say an algebra instead of a K-algebra.

We would like to define a multiplication operation on $T(V)$ which makes it into a K-algebra. As

$$T(V) = \bigoplus_{i \geq 0} V^{\otimes i},$$

for every $i \geq 0$, there is a natural injection $\iota_n\colon V^{\otimes n} \to T(V)$, and in particular, an injection $\iota_0\colon K \to T(V)$. The multiplicative unit $\mathbf{1}$ of $T(V)$ is the image $\iota_0(1)$ in $T(V)$ of the unit 1 of the field K. Since every $v \in T(V)$ can be expressed as a finite sum

$$v = \iota_{n_1}(v_1) + \cdots + \iota_{n_k}(v_k),$$

where $v_i \in V^{\otimes n_i}$ and the n_i are natural numbers with $n_i \neq n_j$ if $i \neq j$, to define multiplication in $T(V)$, using bilinearity, it is enough to define multiplication operations $\cdot\colon V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes (m+n)}$, which, using the isomorphisms $V^{\otimes n} \cong \iota_n(V^{\otimes n})$, yield multiplication operations $\cdot\colon \iota_m(V^{\otimes m}) \times \iota_n(V^{\otimes n}) \longrightarrow \iota_{m+n}(V^{\otimes (m+n)})$. First, for $\omega_1 \in V^{\otimes m}$ and $\omega_2 \in V^{\otimes n}$, we let

$$\omega_1 \cdot \omega_2 = \omega_1 \otimes \omega_2.$$

This defines a bilinear map so it defines a multiplication $V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes m} \otimes V^{\otimes n}$. This is not quite what we want, but there is a canonical isomorphism

$$V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes (m+n)}$$

which yields the desired multiplication $\cdot\colon V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes (m+n)}$.

The isomorphism $V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes (m+n)}$ can be established by induction using the isomorphism $(E \otimes F) \otimes G \cong E \otimes F \otimes G$. First we prove by induction on $m \geq 2$ that

$$V^{\otimes (m-1)} \otimes V \cong V^{\otimes m},$$

and then by induction on $n \geq 1$ than

$$V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes(m+n)}.$$

In summary the multiplication $V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes(m+n)}$ is defined so that

$$(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.$$

(This has to be made rigorous by using isomorphisms involving the associativity of tensor products, for details, see Jacobson [99], Section 3.9, or Bertin [15], Chapter 4, Section 2.)

Definition 33.11. Given a K-vector space V (not necessarily finite dimensional), the vector space

$$T(V) = \bigoplus_{m \geq 0} V^{\otimes m}$$

denoted $T^\bullet(V)$ or $\bigotimes V$ equipped with the multiplication operations $V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes(m+n)}$ defined above is called the *tensor algebra of* V.

Remark: It is important to note that multiplication in $T(V)$ is **not** commutative. Also, in all rigor, the unit $\mathbf{1}$ of $T(V)$ is **not equal** to 1, the unit of the field K. However, in view of the injection $\iota_0 \colon K \to T(V)$, for the sake of notational simplicity, we will denote $\mathbf{1}$ by 1. More generally, in view of the injections $\iota_n \colon V^{\otimes n} \to T(V)$, we identify elements of $V^{\otimes n}$ with their images in $T(V)$.

The algebra $T(V)$ satisfies a universal mapping property which shows that it is unique up to isomorphism. For simplicity of notation, let $i \colon V \to T(V)$ be the natural injection of V into $T(V)$.

Proposition 33.19. *Given any K-algebra A, for any linear map $f \colon V \to A$, there is a unique K-algebra homomorphism $\overline{f} \colon T(V) \to A$ so that*

$$f = \overline{f} \circ i,$$

as in the diagram below.

$$\begin{tikzcd} V \arrow[r, "i"] \arrow[dr, "f"'] & T(V) \arrow[d, "\overline{f}"] \\ & A \end{tikzcd}$$

Proof. Left an an exercise (use Theorem 33.6). A proof can be found in Knapp [104] (Appendix A, Proposition A.14) or Bertin [15] (Chapter 4, Theorem 2.4). $\square$

Proposition 33.19 implies that there is a natural isomorphism

$$\mathrm{Hom}_{\mathrm{alg}}(T(V), A) \cong \mathrm{Hom}(V, A),$$

where the algebra A on the right-hand side is viewed as a vector space. Proposition 33.19 also has the following corollary.

Proposition 33.20. *Given a linear map $h\colon V_1 \to V_2$ between two vectors spaces V_1, V_2 over a field K, there is a unique K-algebra homomorphism $\otimes h\colon T(V_1) \to T(V_2)$ making the following diagram commute.*

$$
\begin{tikzcd}
V_1 \arrow[r, "i_1"] \arrow[d, "h"'] & T(V_1) \arrow[d, "\otimes h"] \\
V_2 \arrow[r, "i_2"] & T(V_2).
\end{tikzcd}
$$

Most algebras of interest arise as well-chosen quotients of the tensor algebra $T(V)$. This is true for the *exterior algebra* $\bigwedge(V)$ (also called *Grassmann algebra*), where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v$, where $v \in V$,and for the *symmetric algebra* $\mathrm{Sym}(V)$, where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes w - w \otimes v$, where $v, w \in V$.

Algebras such as $T(V)$ are graded in the sense that there is a sequence of subspaces $V^{\otimes n} \subseteq T(V)$ such that

$$T(V) = \bigoplus_{k \geq 0} V^{\otimes n},$$

and the multiplication $\otimes$ behaves well w.r.t. the grading, *i.e.*, $\otimes\colon V^{\otimes m} \times V^{\otimes n} \to V^{\otimes (m+n)}$.

Definition 33.12. A K-algebra E is said to be a *graded algebra* iff there is a sequence of subspaces $E^n \subseteq E$ such that

$$E = \bigoplus_{k \geq 0} E^n,$$

(with $E^0 = K$) and the multiplication $\cdot$ respects the grading; that is, $\cdot\colon E^m \times E^n \to E^{m+n}$. Elements in E^n are called *homogeneous elements of rank (or degree) n*.

In differential geometry and in physics it is necessary to consider slightly more general tensors.

Definition 33.13. Given a vector space V, for any pair of nonnegative integers (r, s), the *tensor space $T^{r,s}(V)$ of type (r, s)* is the tensor product

$$T^{r,s}(V) = V^{\otimes r} \otimes (V^*)^{\otimes s} = \underbrace{V \otimes \cdots \otimes V}_{r} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{s},$$

with $T^{0,0}(V) = K$. We also define the *tensor algebra $T^{\bullet,\bullet}(V)$* as the direct sum (coproduct)

$$T^{\bullet,\bullet}(V) = \bigoplus_{r,s \geq 0} T^{r,s}(V).$$

Tensors in $T^{r,s}(V)$ are called *homogeneous of degree (r, s)*.

Note that tensors in $T^{r,0}(V)$ are just our "old tensors" in $V^{\otimes r}$. We make $T^{\bullet,\bullet}(V)$ into an algebra by defining multiplication operations

$$T^{r_1,s_1}(V) \times T^{r_2,s_2}(V) \longrightarrow T^{r_1+r_2,s_1+s_2}(V)$$

in the usual way, namely: For $u = u_1 \otimes \cdots \otimes u_{r_1} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^*$ and
$v = v_1 \otimes \cdots \otimes v_{r_2} \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*$, let

$$u \otimes v = u_1 \otimes \cdots \otimes u_{r_1} \otimes v_1 \otimes \cdots \otimes v_{r_2} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^* \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*.$$

Denote by $\mathrm{Hom}(V^r, (V^*)^s; W)$ the vector space of all multilinear maps from $V^r \times (V^*)^s$ to W. Then we have the universal mapping property which asserts that there is a canonical isomorphism

$$\mathrm{Hom}(T^{r,s}(V), W) \cong \mathrm{Hom}(V^r, (V^*)^s; W).$$

In particular,

$$(T^{r,s}(V))^* \cong \mathrm{Hom}(V^r, (V^*)^s; K).$$

For finite dimensional vector spaces, the duality of Section 33.5 is also easily extended to the tensor spaces $T^{r,s}(V)$. We define the pairing

$$T^{r,s}(V^*) \times T^{r,s}(V) \longrightarrow K$$

as follows: if

$$v^* = v_1^* \otimes \cdots \otimes v_r^* \otimes u_{r+1} \otimes \cdots \otimes u_{r+s} \in T^{r,s}(V^*)$$

and

$$u = u_1 \otimes \cdots \otimes u_r \otimes v_{r+1}^* \otimes \cdots \otimes v_{r+s}^* \in T^{r,s}(V),$$

then

$$(v^*, u) = v_1^*(u_1) \cdots v_{r+s}^*(u_{r+s}).$$

This is a nondegenerate pairing, and thus we get a canonical isomorphism

$$(T^{r,s}(V))^* \cong T^{r,s}(V^*).$$

Consequently, we get a canonical isomorphism

$$T^{r,s}(V^*) \cong \mathrm{Hom}(V^r, (V^*)^s; K).$$

We summarize these results in the following proposition.

Proposition 33.21. *Let V be a vector space and let*

$$T^{r,s}(V) = V^{\otimes r} \otimes (V^*)^{\otimes s} = \underbrace{V \otimes \cdots \otimes V}_{r} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{s}.$$

We have the canonical isomorphisms

$$(T^{r,s}(V))^* \cong T^{r,s}(V^*),$$

and

$$T^{r,s}(V^*) \cong \mathrm{Hom}(V^r, (V^*)^s; K).$$

Remark: The tensor spaces, $T^{r,s}(V)$ are also denoted $T^r_s(V)$. A tensor $\alpha \in T^{r,s}(V)$ is said to be *contravariant* in the first r arguments and *covariant* in the last s arguments. This terminology refers to the way tensors behave under coordinate changes. Given a basis $(e_1, \ldots, e_n)$ of V, if $(e_1^*, \ldots, e_n^*)$ denotes the dual basis, then every tensor $\alpha \in T^{r,s}(V)$ is given by an expression of the form

$$\alpha = \sum_{\substack{i_1,\ldots,i_r\\ j_1,\ldots,j_s}} a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^*_{j_1} \otimes \cdots \otimes e^*_{j_s}.$$

The tradition in classical tensor notation is to use lower indices on vectors and upper indices on linear forms and in accordance to *Einstein summation convention* (or *Einstein notation*) the position of the indices on the coefficients is reversed. *Einstein summation convention* (already encountered in Section 33.1) is to assume that a summation is performed for all values of every index that appears simultaneously once as an upper index and once as a lower index. According to this convention, the tensor α above is written

$$\alpha = a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s}.$$

An older view of tensors is that they are multidimensional arrays of coefficients,

$$\left(a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s}\right),$$

subject to the rules for changes of bases.

Another operation on general tensors, contraction, is useful in differential geometry.

Definition 33.14. For all $r, s \geq 1$, the *contraction* $c_{i,j} \colon T^{r,s}(V) \to T^{r-1,s-1}(V)$, with $1 \leq i \leq r$ and $1 \leq j \leq s$, is the linear map defined on generators by

$$\begin{aligned} &c_{i,j}(u_1 \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes v_s^*) \\ &\qquad = v_j^*(u_i)\, u_1 \otimes \cdots \otimes \widehat{u_i} \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes \widehat{v_j^*} \otimes \cdots \otimes v_s^*, \end{aligned}$$

where the hat over an argument means that it should be omitted.

Let us figure our what is $c_{1,1} \colon T^{1,1}(V) \to \mathbb{R}$, that is $c_{1,1} \colon V \otimes V^* \to \mathbb{R}$. If $(e_1, \ldots, e_n)$ is a basis of V and $(e_1^*, \ldots, e_n^*)$ is the dual basis, by Proposition 33.17 every $h \in V \otimes V^* \cong \mathrm{Hom}(V, V)$ can be expressed as

$$h = \sum_{i,j=1}^n a_{ij}\, e_i \otimes e_j^*.$$

As

$$c_{1,1}(e_i \otimes e_j^*) = \delta_{i,j},$$

we get

$$c_{1,1}(h) = \sum_{i=1}^n a_{ii} = \mathrm{tr}(h),$$

where $\mathrm{tr}(h)$ is the *trace* of h, where h is viewed as the linear map given by the matrix, (a_{ij}). Actually, since $c_{1,1}$ is defined independently of any basis, $c_{1,1}$ provides an intrinsic definition of the trace of a linear map $h \in \mathrm{Hom}(V, V)$.

Remark: Using the Einstein summation convention, if

$$\alpha = a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},$$

then

$$c_{k,l}(\alpha) = a^{i_1,\ldots,i_{k-1},i_{k+1}\ldots,i_r}_{j_1,\ldots,j_{l-1},j_{l+1},\ldots,j_s} e_{i_1} \otimes \cdots \otimes \widehat{e_{i_k}} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes \widehat{e^{j_l}} \otimes \cdots \otimes e^{j_s}.$$

If E and F are two K-algebras, we know that their tensor product $E \otimes F$ exists as a vector space. We can make $E \otimes F$ into an algebra as well. Indeed, we have the multilinear map

$$E \times F \times E \times F \longrightarrow E \otimes F$$

given by $(a, b, c, d) \mapsto (ac) \otimes (bd)$, where ac is the product of a and c in E and bd is the product of b and d in F. By the universal mapping property, we get a linear map,

$$E \otimes F \otimes E \otimes F \longrightarrow E \otimes F.$$

Using the isomorphism

$$E \otimes F \otimes E \otimes F \cong (E \otimes F) \otimes (E \otimes F),$$

we get a linear map

$$(E \otimes F) \otimes (E \otimes F) \longrightarrow E \otimes F,$$

and thus a bilinear map,

$$(E \otimes F) \times (E \otimes F) \longrightarrow E \otimes F$$

which is our multiplication operation in $E \otimes F$. This multiplication is determined by

$$(a \otimes b) \cdot (c \otimes d) = (ac) \otimes (bd).$$

In summary we have the following proposition.

Proposition 33.22. *Given two K-algebra E and F, the operation on $E \otimes F$ defined on generators by*

$$(a \otimes b) \cdot (c \otimes d) = (ac) \otimes (bd)$$

makes $E \otimes F$ into a K-algebra.

We now turn to symmetric tensors.

33.7 Symmetric Tensor Powers

Our goal is to come up with a notion of tensor product that will allow us to treat symmetric multilinear maps as linear maps. Note that we have to restrict ourselves to a *single* vector space E, rather then n vector spaces $E_1, \ldots, E_n$, so that symmetry makes sense.

Definition 33.15. A multilinear map $f\colon E^n \to F$ is *symmetric* iff

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = f(u_1, \ldots, u_n),$$

for all $u_i \in E$ and all permutations, $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. The group of permutations on $\{1, \ldots, n\}$ (the *symmetric group*) is denoted $\mathfrak{S}_n$. The vector space of all symmetric multilinear maps $f\colon E^n \to F$ is denoted by $\mathrm{Sym}^n(E; F)$ or $\mathrm{Hom}_{\mathrm{symlin}}(E^n, F)$. Note that $\mathrm{Sym}^1(E; F) = \mathrm{Hom}(E, F)$.

We could proceed directly as in Theorem 33.6 and construct symmetric tensor products from scratch. However, since we already have the notion of a tensor product, there is a more economical method. First we define symmetric tensor powers.

Definition 33.16. An *n-th symmetric tensor power* of a vector space E, where $n \geq 1$, is a vector space S together with a symmetric multilinear map $\varphi\colon E^n \to S$ such that, for every vector space F and for every symmetric multilinear map $f\colon E^n \to F$, there is a unique linear map $f_\odot\colon S \to F$, with

$$f(u_1, \ldots, u_n) = f_\odot(\varphi(u_1, \ldots, u_n)),$$

for all $u_1, \ldots, u_n \in E$, or for short

$$f = f_\odot \circ \varphi.$$

Equivalently, there is a unique linear map $f_\odot$ such that the following diagram commutes.

$$\begin{tikzcd} E^n \arrow[r, "\varphi"] \arrow[dr, "f"'] & S \arrow[d, "f_\odot"] \\ & F \end{tikzcd}$$

The above property is called the *universal mapping property* of the symmetric tensor power (S, φ).

We next show that any two symmetric n-th tensor powers (S_1, φ_1) and (S_2, φ_2) for E are isomorphic.

Proposition 33.23. *Given any two symmetric n-th tensor powers (S_1, φ_1) and (S_2, φ_2) for E, there is an isomorphism $h\colon S_1 \to S_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Replace tensor product by n-th symmetric tensor power in the proof of Proposition 33.5. $\square$

We now give a construction that produces a symmetric n-th tensor power of a vector space E.

Theorem 33.24. *Given a vector space E, a symmetric n-th tensor power $(\mathrm{S}^n(E), \varphi)$ for E can be constructed ($n \geq 1$). Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$, the symmetric tensor power $\mathrm{S}^n(E)$ is generated by the vectors $u_1 \odot \cdots \odot u_n$, where $u_1, \ldots, u_n \in E$, and for every symmetric multilinear map $f \colon E^n \to F$, the unique linear map $f_\odot \colon \mathrm{S}^n(E) \to F$ such that $f = f_\odot \circ \varphi$ is defined by*

$$f_\odot(u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n)$$

on the generators $u_1 \odot \cdots \odot u_n$ of $\mathrm{S}^n(E)$.

Proof. The tensor power $E^{\otimes n}$ is too big, and thus we define an appropriate quotient. Let C be the subspace of $E^{\otimes n}$ generated by the vectors of the form

$$u_1 \otimes \cdots \otimes u_n - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)},$$

for all $u_i \in E$, and all permutations $\sigma \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. We claim that the quotient space $(E^{\otimes n})/C$ does the job.

Let $p \colon E^{\otimes n} \to (E^{\otimes n})/C$ be the quotient map, and let $\varphi \colon E^n \to (E^{\otimes n})/C$ be the map given by

$$\varphi = p \circ \varphi_0,$$

where $\varphi_0 \colon E^n \to E^{\otimes n}$ is the injection given by $\varphi_0(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n$.

Let us denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$. It is clear that φ is symmetric. Since the vectors $u_1 \otimes \cdots \otimes u_n$ generate $E^{\otimes n}$, and p is surjective, the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^{\otimes n})/C$.

It remains to show that $((E^{\otimes n})/C, \varphi)$ satisfies the universal mapping property. To this end we begin by proving that there is a map h such that $f = h \circ \varphi$. Given any symmetric multilinear map $f \colon E^n \to F$, by Theorem 33.6 there is a linear map $f_\otimes \colon E^{\otimes n} \to F$ such that $f = f_\otimes \circ \varphi_0$, as in the diagram below.

$$\begin{tikzcd} E^n \arrow[r, "\varphi_0"] \arrow[dr, "f"'] & E^{\otimes n} \arrow[d, "f_\otimes"] \\ & F \end{tikzcd}$$

However, since f is symmetric, we have $f_{\otimes}(z) = 0$ for every $z \in C$. Thus, we get an induced linear map $h\colon (E^{\otimes n})/C \to F$ making the following diagram commute.

$$\begin{tikzcd} & E^{\otimes n} \arrow[dr, "p"] \arrow[dd, "f_{\otimes}"] & \\ E^n \arrow[ur, "\varphi_0"] \arrow[dr, "f"'] & & (E^{\otimes n})/C \arrow[dl, "h"] \\ & F & \end{tikzcd}$$

If we define $h([z]) = f_{\otimes}(z)$ for every $z \in E^{\otimes n}$, where $[z]$ is the equivalence class in $(E^{\otimes n})/C$ of $z \in E^{\otimes n}$, the above diagram shows that $f = h \circ p \circ \varphi_0 = h \circ \varphi$. We now prove the uniqueness of h. For any linear map $f_{\odot}\colon (E^{\otimes n})/C \to F$ such that $f = f_{\odot} \circ \varphi$, since $\varphi(u_1, \ldots, u_n) = u_1 \odot \cdots \odot u_n$ and the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^{\otimes n})/C$, the map $f_{\odot}$ is uniquely defined by

$$f_{\odot}(u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n).$$

Since $f = h \circ \varphi$, the map h is unique, and we let $f_{\odot} = h$. Thus, $\mathrm{S}^n(E) = (E^{\otimes n})/C$ and φ constitute a symmetric n-th tensor power of E. $\square$

The map φ from E^n to $\mathrm{S}^n(E)$ is often denoted $\iota_{\odot}$, so that

$$\iota_{\odot}(u_1, \ldots, u_n) = u_1 \odot \cdots \odot u_n.$$

Again, the actual construction is not important. What is important is that the symmetric n-th power has the universal mapping property with respect to symmetric multilinear maps.

Remark: The notation $\odot$ for the commutative multiplication of symmetric tensor powers is not standard. Another notation commonly used is $\cdot$. We often abbreviate "symmetric tensor power" as "symmetric power." The symmetric power $\mathrm{S}^n(E)$ is also denoted $\mathrm{Sym}^n E$ but we prefer to use the notation Sym to denote spaces of symmetric multilinear maps. To be consistent with the use of $\odot$, we could have used the notation $\bigodot^n E$. Clearly, $\mathrm{S}^1(E) \cong E$ and it is convenient to set $\mathrm{S}^0(E) = K$.

The fact that the map $\varphi\colon E^n \to \mathrm{S}^n(E)$ is symmetric and multilinear can also be expressed as follows:

$$\begin{aligned} u_1 \odot \cdots \odot (v_i + w_i) \odot \cdots \odot u_n &= (u_1 \odot \cdots \odot v_i \odot \cdots \odot u_n) + (u_1 \odot \cdots \odot w_i \odot \cdots \odot u_n), \\ u_1 \odot \cdots \odot (\lambda u_i) \odot \cdots \odot u_n &= \lambda(u_1 \odot \cdots \odot u_i \odot \cdots \odot u_n), \\ u_{\sigma(1)} \odot \cdots \odot u_{\sigma(n)} &= u_1 \odot \cdots \odot u_n, \end{aligned}$$

for all permutations $\sigma \in \mathfrak{S}_n$.

The last identity shows that the "operation" $\odot$ is commutative. This allows us to view the symmetric tensor $u_1 \odot \cdots \odot u_n$ as an object called a multiset.

Given a set A, a multiset with elements from A is a generalization of the concept of a set that allows multiple instances of elements from A to occur. For example, if $A = \{a, b, c, d\}$, the following are multisets:

$$M_1 = \{a, a, b\}, \;\; M_2 = \{a, a, b, b, c\}, \;\; M_3 = \{a, a, b, b, c, d, d, d\}.$$

Here is another way to represent multisets as tables showing the multiplicities of the elements in the multiset:

$$M_1 = \begin{pmatrix} a & b & c & d \\ 2 & 1 & 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} a & b & c & d \\ 2 & 2 & 1 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} a & b & c & d \\ 2 & 2 & 1 & 3 \end{pmatrix}.$$

The above are just graphs of functions from the set $A = \{a, b, c, d\}$ to $\mathbb{N}$. This suggests the following definition.

Definition 33.17. A finite *multiset* M over a set A is a function $M \colon A \to \mathbb{N}$ such that $M(a) \neq 0$ for finitely many $a \in A$. The *multiplicity* of an element $a \in A$ in M is $M(a)$. The set of all multisets over A is denoted by $\mathbb{N}^{(A)}$, and we let $\mathrm{dom}(M) = \{a \in A \mid M(a) \neq 0\}$, which is a finite set. The set $\mathrm{dom}(M)$ is the set of elements in A that actually occur in M. For any multiset $M \in \mathbb{N}^{(A)}$, note that $\sum_{a \in A} M(a)$ makes sense, since $\sum_{a\in A} M(a) = \sum_{a \in \mathrm{dom}(A)} M(a)$, and $\mathrm{dom}(M)$ is finite; this sum is the total number of elements in the multiset A and is called the *size* of M. Let $|M| = \sum_{a\in A} M(a)$.

Going back to our symmetric tensors, we can view the tensors of the form $u_1 \odot \cdots \odot u_n$ as multisets of size n over the set E.

Theorem 33.24 implies the following proposition.

Proposition 33.25. *There is a canonical isomorphism*

$$\mathrm{Hom}(\mathrm{S}^n(E), F) \cong \mathrm{Sym}^n(E; F),$$

between the vector space of linear maps $\mathrm{Hom}(\mathrm{S}^n(E), F)$ *and the vector space of symmetric multilinear maps* $\mathrm{Sym}^n(E; F)$ *given by the linear map* $- \circ \varphi$ *defined by* $h \mapsto h \circ \varphi$*, with* $h \in \mathrm{Hom}(\mathrm{S}^n(E), F)$*.*

Proof. The map $h \circ \varphi$ is clearly symmetric multilinear. By Theorem 33.24, for every symmetric multilinear map $f \in \mathrm{Sym}^n(E; F)$ there is a unique linear map $f_\odot \in \mathrm{Hom}(\mathrm{S}^n(E), F)$ such that $f = f_\odot \circ \varphi$, so the map $- \circ \varphi$ is bijective. Its inverse is the map $f \mapsto f_\odot$. $\square$

In particular, when $F = K$, we get the following important fact.

Proposition 33.26. *There is a canonical isomorphism*

$$(\mathrm{S}^n(E))^* \cong \mathrm{Sym}^n(E; K).$$

Definition 33.18. Symmetric tensors in $\mathrm{S}^n(E)$ are called *symmetric n-tensors*, and tensors of the form $u_1 \odot \cdots \odot u_n$, where $u_i \in E$, are called *simple (or decomposable) symmetric n-tensors*. Those symmetric n-tensors that are not simple are often called *compound symmetric n-tensors*.

Given linear map $f\colon E \to E'$, since the map $\iota'_{\odot} \circ (f \times f)$ is bilinear and symmetric, there is a unique linear map $f \odot f\colon \mathrm{S}^2(E) \to \mathrm{S}^2(E')$ making the following diagram commute.

$$\begin{tikzcd}
E^2 \arrow[r, "\iota_{\odot}"] \arrow[d, "f\times f"'] & \mathrm{S}^2(E) \arrow[d, "f\odot f"] \\
(E')^2 \arrow[r, "\iota'_{\odot}"'] & \mathrm{S}^2(E').
\end{tikzcd}$$

Observe that $f \odot g$ is determined by

$$(f \odot f)(u \odot v) = f(u) \odot f(v).$$

Proposition 33.27. *Given any two linear maps $f\colon E \to E'$ and $f'\colon E' \to E''$, we have*

$$(f' \circ f) \odot (f' \circ f) = (f' \odot f') \circ (f \odot f).$$

The generalization to the symmetric tensor product $f \odot \cdots \odot f$ of $n \geq 3$ copies of the linear map $f\colon E \to E'$ is immediate, and left to the reader.

33.8 Bases of Symmetric Powers

The vectors $u_1 \odot \cdots \odot u_m$ where $u_1, \ldots, u_m \in E$ generate $\mathrm{S}^m(E)$, but they are not linearly independent. We will prove a version of Proposition 33.12 for symmetric tensor powers using multisets.

Recall that a (finite) multiset over a set I is a function $M\colon I \to \mathbb{N}$, such that $M(i) \neq 0$ for finitely many $i \in I$. The set of all multisets over I is denoted as $\mathbb{N}^{(I)}$ and we let $\mathrm{dom}(M) = \{i \in I \mid M(i) \neq 0\}$, the finite set of elements in I that actually occur in M. The size of the multiset M is $|M| = \sum_{a\in A} M(a)$.

To explain the idea of the proof, consider the case when $m = 2$ and E has dimension 3. Given a basis (e_1, e_2, e_3) of E, we would like to prove that

$$e_1 \odot e_1, \quad e_1 \odot e_2, \quad e_1 \odot e_3, \quad e_2 \odot e_2, \quad e_2 \odot e_3, \quad e_3 \odot e_3$$

are linearly independent. To prove this, it suffices to show that for any vector space F, if $w_{11}, w_{12}, w_{13}, w_{22}, w_{23}, w_{33}$ are any vectors in F, then there is a symmetric bilinear map $h\colon E^2 \to F$ such that

$$h(e_i, e_j) = w_{ij}, \quad 1 \leq i \leq j \leq 3.$$

Because h yields a unique linear map $h_\odot\colon \mathrm{S}^2(E) \to F$ such that

$$h_\odot(e_i \odot e_j) = w_{ij}, \quad 1 \leq i \leq j \leq 3,$$

by Proposition 33.4, the vectors

$$e_1 \odot e_1, \quad e_1 \odot e_2, \quad e_1 \odot e_3, \quad e_2 \odot e_2, \quad e_2 \odot e_3, \quad e_3 \odot e_3$$

are linearly independent. This suggests understanding how a symmetric bilinear function $f\colon E^2 \to F$ is expressed in terms of its values $f(e_i, e_j)$ on the basis vectors (e_1, e_2, e_3), and this can be done easily. Using bilinearity and symmetry, we obtain

$$\begin{aligned} f(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) &= u_1v_1f(e_1, e_1) + (u_1v_2 + u_2v_1)f(e_1, e_2) \\ &\quad + (u_1v_3 + u_3v_1)f(e_1, e_3) + u_2v_2f(e_2, e_2) \\ &\quad + (u_2v_3 + u_3v_2)f(e_2, e_3) + u_3v_3f(e_3, e_3). \end{aligned}$$

Therefore, given $w_{11}, w_{12}, w_{13}, w_{22}, w_{23}, w_{33} \in F$, the function h given by

$$\begin{aligned} h(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) &= u_1v_1w_{11} + (u_1v_2 + u_2v_1)w_{12} \\ &\quad + (u_1v_3 + u_3v_1)w_{13} + u_2v_2w_{22} \\ &\quad + (u_2v_3 + u_3v_2)w_{23} + u_3v_3w_{33} \end{aligned}$$

is clearly bilinear symmetric, and by construction $h(e_i, e_j) = w_{ij}$, so it does the job.

The generalization of this argument to any $m \geq 2$ and to a space E of any dimension (even infinite) is conceptually clear, but notationally messy. If $\dim(E) = n$ and if $(e_1, \ldots, e_n)$ is a basis of E, for any m vectors $v_j = \sum_{i=1}^n u_{i,j}e_i$ in E, for any symmetric multilinear map $f\colon E^m \to F$, we have

$$\begin{aligned} &f(v_1, \ldots, v_m) \\ &= \sum_{k_1+\cdots+k_n=m} \left(\sum_{\substack{I_1\cup\cdots\cup I_n=\{1,\ldots,m\} \\ I_i\cap I_j=\emptyset,\, i\neq j,\, |I_j|=k_j}} \left(\prod_{i_1\in I_1} u_{1,i_1} \right) \cdots \left(\prod_{i_n\in I_n} u_{n,i_n} \right) \right) f(\underbrace{e_1, \ldots, e_1}_{k_1}, \ldots, \underbrace{e_n, \ldots, e_n}_{k_n}). \end{aligned}$$

Definition 33.19. Given any set J of $n \geq 1$ elements, say $J = \{j_1, \ldots, j_n\}$, and given any $m \geq 2$, for any sequence $(k_1 \ldots, k_n)$ of natural numbers $k_i \in \mathbb{N}$ such that $k_1 + \cdots + k_n = m$, the multiset M of size m

$$M = \{\underbrace{j_1, \ldots, j_1}_{k_1}, \underbrace{j_2, \ldots, j_2}_{k_2}, \ldots, \underbrace{j_n, \ldots, j_n}_{k_n}\}$$

is denoted by $M(m, J, k_1, \ldots, k_n)$. Note that $M(j_i) = k_i$, for $i = 1, \ldots, n$. Given any $k \geq 1$, and any $u \in E$, we denote $\underbrace{u \odot \cdots \odot u}_{k}$ as $u^{\odot k}$.

We can now prove the following proposition.

Proposition 33.28. *Given a vector space E, if $(e_i)_{i\in I}$ is a basis for E, then the family of vectors*

$$\left(e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)}\right)_{\substack{M\in\mathbb{N}^{(I)},\, |M|=m,\\ \{i_1,\ldots,i_k\}=\mathrm{dom}(M)}}$$

is a basis of the symmetric m-th tensor power $\mathrm{S}^m(E)$.

Proof. The proof is very similar to that of Proposition 33.12. First assume that E has finite dimension n. In this case $I = \{1,\ldots,n\}$, and any multiset $M \in \mathbb{N}^{(I)}$ of size $|M| = m$ is of the form $M(m,\{1,\ldots,n\},k_1,\ldots,k_n)$, with $k_i = M(i)$ and $k_1+\cdots+k_n = m$.

For any nontrivial vector space F, for any family of vectors

$$(w_M)_{M\in\mathbb{N}^{(I)},\,|M|=m},$$

we show the existence of a symmetric multilinear map $h\colon \mathrm{S}^m(E) \to F$, such that for every $M \in \mathbb{N}^{(I)}$ with $|M| = m$, we have

$$h(e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)}) = w_M,$$

where $\{i_1,\ldots,i_k\} = \mathrm{dom}(M)$. We define the map $f\colon E^m \to F$ as follows: for any m vectors $v_1,\ldots,v_m \in E$ we can write $v_k = \sum_{i=1}^n u_{i,k}e_i$ for $k = 1,\ldots,m$ and we set

$$\begin{aligned} &f(v_1,\ldots,v_m)\\ &\quad = \sum_{k_1+\cdots+k_n=m}\left(\sum_{\substack{I_1\cup\cdots\cup I_n=\{1,\ldots,m\}\\ I_i\cap I_j=\emptyset,\, i\neq j,\, |I_j|=k_j}}\left(\prod_{i_1\in I_1} u_{1,i_1}\right)\cdots\left(\prod_{i_n\in I_n} u_{n,i_n}\right)\right)w_{M(m,\{1,\ldots,n\},k_1,\ldots,k_n)}. \end{aligned}$$

It is not difficult to verify that f is symmetric and multilinear. By the universal mapping property of the symmetric tensor product, the linear map $f_\odot\colon \mathrm{S}^m(E) \to F$ such that $f = f_\odot \circ \varphi$, is the desired map h. Then by Proposition 33.4, it follows that the family

$$\left(e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)}\right)_{\substack{M\in\mathbb{N}^{(I)},\, |M|=m,\\ \{i_1,\ldots,i_k\}=\mathrm{dom}(M)}}$$

is linearly independent. Using the commutativity of $\odot$, we can also show that these vectors generate $\mathrm{S}^m(E)$, and thus, they form a basis for $\mathrm{S}^m(E)$.

If I is infinite dimensional, then for any m vectors $v_1,\ldots,v_m \in F$ there is a finite subset J of I such that $v_k = \sum_{j\in J} u_{j,k}e_j$ for $k = 1,\ldots,m$, and if we write $n = |J|$, then the formula for $f(v_1,\ldots,v_m)$ is obtained by replacing the set $\{1,\ldots,n\}$ by J. The details are left as an exercise. $\square$

As a consequence, when I is finite, say of size $p = \dim(E)$, the dimension of $\mathrm{S}^m(E)$ is the number of finite multisets $(j_1, \ldots, j_p)$, such that $j_1 + \cdots + j_p = m$, $j_k \geq 0$. We leave as an exercise to show that this number is $\binom{p+m-1}{m}$. Thus, if $\dim(E) = p$, then the dimension of $\mathrm{S}^m(E)$ is $\binom{p+m-1}{m}$. Compare with the dimension of $E^{\otimes m}$, which is p^m. In particular, when $p = 2$, the dimension of $\mathrm{S}^m(E)$ is $m+1$. This can also be seen directly.

Remark: The number $\binom{p+m-1}{m}$ is also the number of homogeneous monomials

$$X_1^{j_1} \cdots X_p^{j_p}$$

of total degree m in p variables (we have $j_1 + \cdots + j_p = m$). This is not a coincidence! Given a vector space E and a basis $(e_i)_{i \in I}$ for E, Proposition 33.28 shows that every symmetric tensor $z \in \mathrm{S}^m(E)$ can be written in a unique way as

$$z = \sum_{\substack{M \in \mathbb{N}^{(I)} \\ \sum_{i \in I} M(i) = m \\ \{i_1, \ldots, i_k\} = \mathrm{dom}(M)}} \lambda_M \, e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)},$$

for some unique family of scalars $\lambda_M \in K$, all zero except for a finite number.

This looks like a homogeneous polynomial of total degree m, where the monomials of total degree m are the symmetric tensors

$$e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)}$$

in the "indeterminates" e_i, where $i \in I$ (recall that $M(i_1) + \cdots + M(i_k) = m$) and implies that polynomials can be defined in terms of symmetric tensors.

33.9 Some Useful Isomorphisms for Symmetric Powers

We can show the following property of the symmetric tensor product, using the proof technique of Proposition 33.13 (3).

Proposition 33.29. *We have the following isomorphism:*

$$\mathrm{S}^n(E \oplus F) \cong \bigoplus_{k=0}^{n} \mathrm{S}^k(E) \otimes \mathrm{S}^{n-k}(F).$$

33.10 Duality for Symmetric Powers

In this section all vector spaces are assumed to have *finite dimension over a field of characteristic zero*. We define a nondegenerate pairing $\mathrm{S}^n(E^*) \times \mathrm{S}^n(E) \longrightarrow K$ as follows: Consider the multilinear map

$$(E^*)^n \times E^n \longrightarrow K$$

given by

$$(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).$$

Note that the expression on the right-hand side is "almost" the determinant $\det(v_j^*(u_i))$, except that the sign $\mathrm{sgn}(\sigma)$ is missing (where $\mathrm{sgn}(\sigma)$ is the signature of the permutation σ; that is, the parity of the number of transpositions into which σ can be factored). Such an expression is called a *permanent*.

It can be verified that this expression is symmetric w.r.t. the u_i's and also w.r.t. the v_j^*. For any fixed $(v_1^*, \ldots, v_n^*) \in (E^*)^n$, we get a symmetric multilinear map

$$l_{v_1^*, \ldots, v_n^*} \colon (u_1, \ldots, u_n) \mapsto \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n)$$

from E^n to K. The map $l_{v_1^*, \ldots, v_n^*}$ extends uniquely to a linear map $L_{v_1^*, \ldots, v_n^*} \colon \mathrm{S}^n(E) \to K$ making the following diagram commute:

$$\begin{tikzcd} E^n \arrow[r, "\iota_{\odot}"] \arrow[dr, "l_{v_1^*, \ldots, v_n^*}"'] & \mathrm{S}^n(E) \arrow[d, "L_{v_1^*, \ldots, v_n^*}"] \\ & K. \end{tikzcd}$$

We also have the symmetric multilinear map

$$(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*, \ldots, v_n^*}$$

from $(E^*)^n$ to $\mathrm{Hom}(\mathrm{S}^n(E), K)$, which extends to a linear map L from $\mathrm{S}^n(E^*)$ to $\mathrm{Hom}(\mathrm{S}^n(E), K)$ making the following diagram commute:

$$\begin{tikzcd} (E^*)^n \arrow[r, "\iota_{\odot^*}"] \arrow[dr] & \mathrm{S}^n(E^*) \arrow[d, "L"] \\ & \mathrm{Hom}(\mathrm{S}^n(E), K). \end{tikzcd}$$

However, in view of the isomorphism

$$\mathrm{Hom}(U \otimes V, W) \cong \mathrm{Hom}(U, \mathrm{Hom}(V, W)),$$

with $U = \mathrm{S}^n(E^*)$, $V = \mathrm{S}^n(E)$ and $W = K$, we can view L as a linear map

$$L \colon \mathrm{S}^n(E^*) \otimes \mathrm{S}^n(E) \longrightarrow K,$$

which by Proposition 33.8 corresponds to a bilinear map

$$\langle -, - \rangle \colon \mathrm{S}^n(E^*) \times \mathrm{S}^n(E) \longrightarrow K. \tag{$*$}$$

This pairing is given explicitly on generators by

$$\langle v_1^* \odot \cdots \odot v_n^*, u_1, \ldots, u_n \rangle = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).$$

Now this pairing in nondegenerate. This can be shown using bases.[2] If $(e_1, \ldots, e_m)$ is a basis of E, then for every basis element $(e_{i_1}^*)^{\odot n_1} \odot \cdots \odot (e_{i_k}^*)^{\odot n_k}$ of $\mathrm{S}^n(E^*)$, with $n_1 + \cdots + n_k = n$, we have

$$\langle (e_{i_1}^*)^{\odot n_1} \odot \cdots \odot (e_{i_k}^*)^{\odot n_k}, e_{i_1}^{\odot n_1} \odot \cdots \odot e_{i_k}^{\odot n_k} \rangle = n_1! \cdots n_k!,$$

and

$$\langle (e_{i_1}^*)^{\odot n_1} \odot \cdots \odot (e_{i_k}^*)^{\odot n_k}, e_{j_1} \odot \cdots \odot e_{j_n} \rangle = 0$$

if $(j_1 \ldots, j_n) \neq (\underbrace{i_1, \ldots, i_1}_{n_1}, \ldots, \underbrace{i_k, \ldots, i_k}_{n_k})$.

If the field K has characteristic zero, then $n_1! \cdots n_k! \neq 0$. We leave the details as an exercise to the reader. Therefore we get a canonical isomorphism

$$(\mathrm{S}^n(E))^* \cong \mathrm{S}^n(E^*).$$

The following proposition summarizes the duality properties of symmetric powers.

Proposition 33.30. *Assume the field K has characteristic zero. We have the canonical isomorphisms*

$$(\mathrm{S}^n(E))^* \cong \mathrm{S}^n(E^*)$$

and

$$\mathrm{S}^n(E^*) \cong \mathrm{Sym}^n(E; K) = \mathrm{Hom}_{\mathrm{symlin}}(E^n, K),$$

which allows us to interpret symmetric tensors over E^* as symmetric multilinear maps.

Proof. The isomorphism

$$\mu \colon \mathrm{S}^n(E^*) \cong \mathrm{Sym}^n(E; K)$$

follows from the isomorphisms $(\mathrm{S}^n(E))^* \cong \mathrm{S}^n(E^*)$ and $(\mathrm{S}^n(E))^* \cong \mathrm{Sym}^n(E; K)$ given by Proposition 33.26. $\square$

Remarks:

1. The isomorphism $\mu \colon \mathrm{S}^n(E^*) \cong \mathrm{Sym}^n(E; K)$ discussed above can be described explicitly as the linear extension of the map given by

$$\mu(v_1^* \odot \cdots \odot v_n^*)(u_1, \ldots, u_n) = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).$$

[2]This is where the assumption that we are in finite dimension and that the field has characteristic zero are used.

If $(e_1, \ldots, e_m)$ is a basis of E, then for every basis element $(e_{i_1}^*)^{\odot n_1} \odot \cdots \odot (e_{i_k}^*)^{\odot n_k}$ of $\mathrm{S}^n(E^*)$, with $n_1 + \cdots + n_k = n$, we have

$$\mu((e_{i_1}^*)^{\odot n_1} \odot \cdots \odot (e_{i_k}^*)^{\odot n_k})(\underbrace{e_{i_1}, \ldots, e_{i_1}}_{n_1} \ldots, \underbrace{e_{i_k}, \ldots, e_{i_k}}_{n_k}) = n_1! \cdots n_k!,$$

If the field K has positive characteristic, then it is possible that $n_1! \cdots n_k! = 0$, and this is why we required K to be of characteristic 0 in order for Proposition 33.30 to hold.

2. The canonical isomorphism of Proposition 33.30 holds under more general conditions. Namely, that K is a commutative algebra with identity over $\mathbb{Q}$, and that the E is a finitely-generated projective K-module (see Definition 35.7). See Bourbaki, [25] (Chapter III, §11, Section 5, Proposition 8).

The map from E^n to $\mathrm{S}^n(E)$ given by $(u_1, \ldots, u_n) \mapsto u_1 \odot \cdots \odot u_n$ yields a surjection $\pi\colon E^{\otimes n} \to \mathrm{S}^n(E)$. Because we are dealing with vector spaces, this map has some section; that is, there is some injection $\eta\colon \mathrm{S}^n(E) \to E^{\otimes n}$ with $\pi \circ \eta = \mathrm{id}$. Since our field K has characteristic 0, there is a special section having a natural definition involving a symmetrization process defined as follows: For every permutation σ, we have the map $r_\sigma\colon E^n \to E^{\otimes n}$ given by

$$r_\sigma(u_1, \ldots, u_n) = u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$

As r_σ is clearly multilinear, r_σ extends to a linear map $(r_\sigma)_\otimes\colon E^{\otimes n} \to E^{\otimes n}$ making the following diagram commute

$$\begin{tikzcd} E^n \arrow[r, "\iota_\otimes"] \arrow[dr, "r_\sigma"'] & E^{\otimes n} \arrow[d, "(r_\sigma)_\otimes"] \\ & E^{\otimes n}, \end{tikzcd}$$

and we get a map $\mathfrak{S}_n \times E^{\otimes n} \longrightarrow E^{\otimes n}$, namely

$$\sigma \cdot z = (r_\sigma)_\otimes(z).$$

It is immediately checked that this is a left action of the symmetric group $\mathfrak{S}_n$ on $E^{\otimes n}$, and the tensors $z \in E^{\otimes n}$ such that

$$\sigma \cdot z = z, \quad \text{for all} \quad \sigma \in \mathfrak{S}_n$$

are called *symmetrized* tensors.

We define the map $\eta\colon E^n \to E^{\otimes n}$ by

$$\eta(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \sigma \cdot (u_1 \otimes \cdots \otimes u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$

As the right hand side is clearly symmetric, we get a linear map $\eta_\odot \colon \mathrm{S}^n(E) \to E^{\otimes n}$ making the following diagram commute.

$$\begin{tikzcd} E^n \arrow[r, "\iota_\odot"] \arrow[dr, "\eta"'] & \mathrm{S}^n(E) \arrow[d, "\eta_\odot"] \\ & E^{\otimes n} \end{tikzcd}$$

Clearly, $\eta_\odot(\mathrm{S}^n(E))$ is the set of symmetrized tensors in $E^{\otimes n}$. If we consider the map $S = \eta_\odot \circ \pi \colon E^{\otimes n} \longrightarrow E^{\otimes n}$ where π is the surjection $\pi \colon E^{\otimes n} \to \mathrm{S}^n(E)$, it is easy to check that $S \circ S = S$. Therefore, S is a projection, and by linear algebra, we know that

$$E^{\otimes n} = S(E^{\otimes n}) \oplus \mathrm{Ker}\, S = \eta_\odot(\mathrm{S}^n(E)) \oplus \mathrm{Ker}\, S.$$

It turns out that $\mathrm{Ker}\, S = E^{\otimes n} \cap \mathfrak{I} = \mathrm{Ker}\, \pi$, where $\mathfrak{I}$ is the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes v - v \otimes u \in E^{\otimes 2}$ (for example, see Knapp [104], Appendix A). Therefore, $\eta_\odot$ is injective,

$$E^{\otimes n} = \eta_\odot(\mathrm{S}^n(E)) \oplus (E^{\otimes n} \cap \mathfrak{I}) = \eta_\odot(\mathrm{S}^n(E)) \oplus \mathrm{Ker}\, \pi,$$

and the symmetric tensor power $\mathrm{S}^n(E)$ is naturally embedded into $E^{\otimes n}$.

33.11 Symmetric Algebras

As in the case of tensors, we can pack together all the symmetric powers $\mathrm{S}^n(V)$ into an algebra.

Definition 33.20. Given a vector space V, the space

$$\mathrm{S}(V) = \bigoplus_{m \geq 0} \mathrm{S}^m(V),$$

is called the *symmetric tensor algebra of V*.

We could adapt what we did in Section 33.6 for general tensor powers to symmetric tensors but since we already have the algebra $T(V)$, we can proceed faster. If $\mathfrak{I}$ is the two-sided ideal generated by all tensors of the form $u \otimes v - v \otimes u \in V^{\otimes 2}$, we set

$$\mathrm{S}^\bullet(V) = T(V)/\mathfrak{I}.$$

Observe that since the ideal $\mathfrak{I}$ is generated by elements in $V^{\otimes 2}$, every tensor in $\mathfrak{I}$ is a linear combination of tensors of the form $\omega_1 \otimes (u \otimes v - v \otimes u) \otimes \omega_2$, with $\omega_1 \in V^{\otimes n_1}$ and $\omega_2 \in V^{\otimes n_2}$ for some $n_1, n_2 \in \mathbb{N}$, which implies that

$$\mathfrak{I} = \bigoplus_{m \geq 0} (\mathfrak{I} \cap V^{\otimes m}).$$

Then, $\mathrm{S}^\bullet(V)$ automatically inherits a multiplication operation which is commutative, and since $T(V)$ is graded, that is

$$T(V) = \bigoplus_{m\geq 0} V^{\otimes m},$$

we have

$$\mathrm{S}^\bullet(V) = \bigoplus_{m\geq 0} V^{\otimes m}/(\mathfrak{I}\cap V^{\otimes m}).$$

However, it is easy to check that

$$\mathrm{S}^m(V) \cong V^{\otimes m}/(\mathfrak{I}\cap V^{\otimes m}),$$

so

$$\mathrm{S}^\bullet(V) \cong \mathrm{S}(V).$$

When V is of finite dimension n, $\mathrm{S}(V)$ corresponds to *the algebra of polynomials with coefficients in K in n variables* (this can be seen from Proposition 33.28). When V is of infinite dimension and $(u_i)_{i\in I}$ is a basis of V, the algebra $\mathrm{S}(V)$ corresponds to the algebra of polynomials in infinitely many variables in I. What's nice about the symmetric tensor algebra $\mathrm{S}(V)$ is that it provides an intrinsic definition of a polynomial algebra in any set of I variables.

It is also easy to see that $\mathrm{S}(V)$ satisfies the following universal mapping property.

Proposition 33.31. *Given any commutative K-algebra A, for any linear map $f\colon V\to A$, there is a unique K-algebra homomorphism $\overline{f}\colon \mathrm{S}(V)\to A$ so that*

$$f = \overline{f}\circ i,$$

as in the diagram below.

$$\begin{tikzcd} V \arrow[r, "i"] \arrow[dr, "f"'] & \mathrm{S}(V) \arrow[d, "\overline{f}"] \\ & A \end{tikzcd}$$

Remark: If E is finite-dimensional, recall the isomorphism $\mu\colon \mathrm{S}^n(E^*) \longrightarrow \mathrm{Sym}^n(E;K)$ defined as the linear extension of the map given by

$$\mu(v_1^*\odot\cdots\odot v_n^*)(u_1,\ldots,u_n) = \sum_{\sigma\in\mathfrak{S}_n} v^*_{\sigma(1)}(u_1)\cdots v^*_{\sigma(n)}(u_n).$$

Now we have also a multiplication operation $\mathrm{S}^m(E^*)\times\mathrm{S}^n(E^*)\longrightarrow\mathrm{S}^{m+n}(E^*)$. The following question then arises:

Can we define a multiplication $\mathrm{Sym}^m(E;K) \times \mathrm{Sym}^n(E;K) \longrightarrow \mathrm{Sym}^{m+n}(E;K)$ directly on symmetric multilinear forms, so that the following diagram commutes?

$$\begin{tikzcd} \mathrm{S}^m(E^*) \times \mathrm{S}^n(E^*) \arrow[r, "\odot"] \arrow[d, "\mu_m \times \mu_n"] & \mathrm{S}^{m+n}(E^*) \arrow[d, "\mu_{m+n}"] \\ \mathrm{Sym}^m(E;K) \times \mathrm{Sym}^n(E;K) \arrow[r, "\cdot"] & \mathrm{Sym}^{m+n}(E;K) \end{tikzcd}$$

The answer is *yes*! The solution is to define this multiplication such that for $f \in \mathrm{Sym}^m(E;K)$ and $g \in \mathrm{Sym}^n(E;K)$,

$$(f \cdot g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}), \qquad (*)$$

where shuffle(m,n) consists of all (m,n)-"shuffles;" that is, permutations σ of $\{1, \ldots m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$. Observe that a (m,n)-shuffle is completely determined by the sequence $\sigma(1) < \cdots < \sigma(m)$.

For example, suppose $m = 2$ and $n = 1$. Given $v_1^*, v_2^*, v_3^* \in E^*$, the multiplication structure on $\mathrm{S}(E^*)$ implies that $(v_1^* \odot v_2^*) \cdot v_3^* = v_1^* \odot v_2^* \odot v_3^* \in \mathrm{S}^3(E^*)$. Furthermore, for $u_1, u_2, u_3, \in E$,

$$\begin{aligned} \mu_3(v_1^* \odot v_2^* \odot v_3^*)(u_1, u_2, u_3) &= \sum_{\sigma \in \mathfrak{S}_3} v_{\sigma(1)}^*(u_1) v_{\sigma(2)}^*(u_2) v_{\sigma(3)}^*(u_3) \\ &= v_1^*(u_1)v_2^*(u_2)v_3^*(u_3) + v_1^*(u_1)v_3^*(u_2)v_2^*(u_3) \\ &\quad + v_2^*(u_1)v_1^*(u_2)v_3^*(u_3) + v_2^*(u_1)v_3^*(u_2)v_1^*(u_3) \\ &\quad + v_3^*(u_1)v_1^*(u_2)v_2^*(u_3) + v_3^*(u_1)v_2^*(u_2)v_1^*(u_3). \end{aligned}$$

Now the $(2,1)$- shuffles of $\{1,2,3\}$ are the following three permutations, namely

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

If $f \cong \mu_2(v_1^* \odot v_2^*)$ and $g \cong \mu_1(v_3^*)$, then $(*)$ implies that

$$\begin{aligned} (f \cdot g)(u_1, u_2, u_3) &= \sum_{\sigma \in \mathrm{shuffle}(2,1)} f(u_{\sigma(1)}, u_{\sigma_(2)}) g(u_{\sigma(3)}) \\ &= f(u_1, u_2)g(u_3) + f(u_1, u_3)g(u_2) + f(u_2, u_3)g(u_1) \\ &= \mu_2(v_1^* \odot v_2^*)(u_1, u_2)\mu_1(v_3^*)(u_3) + \mu_2(v_1^* \odot v_2^*)(u_1, u_3)\mu_1(v_3^*)(u_2) \\ &\quad + \mu_2(v_1^* \odot v_2^*)(u_2, u_3)\mu_1(v_3^*)(u_1) \\ &= (v_1^*(u_1)v_2^*(u_2) + v_2^*(u_1)v_1^*(u_2))v_3^*(u_3) \\ &\quad + (v_1^*(u_1)v_2^*(u_3) + v_2^*(u_1)v_1^*(u_3))v_3^*(u_2) \\ &\quad + (v_1^*(u_2)v_2^*(u_3) + v_2^*(u_2)v_1^*(u_3))v_3^*(u_1) \\ &= \mu_3(v_1^* \odot v_2^* \odot v_3^*)(u_1, u_2, u_3). \end{aligned}$$

We leave it as an exercise for the reader to verify Equation $(*)$ for arbitrary nonnegative integers m and n.

Another useful canonical isomorphism (of K-algebras) is given below.

Proposition 33.32. *For any two vector spaces E and F, there is a canonical isomorphism (of K-algebras)*

$$\mathrm{S}(E \oplus F) \cong \mathrm{S}(E) \otimes \mathrm{S}(F).$$

33.12 Problems

Problem 33.1. Prove Proposition 33.4.

Problem 33.2. Given two linear maps $f\colon E \to E'$ and $g\colon F \to F'$, we defined the unique linear map

$$f \otimes g\colon E \otimes F \to E' \otimes F'$$

by

$$(f \otimes g)(u \otimes v) = f(u) \otimes g(v),$$

for all $u \in E$ and all $v \in F$. See Proposition 33.9. Thus $f \otimes g \in \mathrm{Hom}(E \otimes F, E' \otimes F')$. If we denote the tensor product $E \otimes F$ by $T(E, F)$, and we assume that E, E' and F, F' are finite dimensional, pick bases and show that the map induced by $f \otimes g \mapsto T(f, g)$ is an isomorphism

$$\mathrm{Hom}(E, F) \otimes \mathrm{Hom}(E', F') \cong \mathrm{Hom}(E \otimes F, E' \otimes F').$$

Problem 33.3. Adjust the proof of Proposition 33.13 (2) to show that

$$E \otimes (F \otimes G) \cong E \otimes F \otimes G,$$

whenever E, F, and G are arbitrary vector spaces.

Problem 33.4. Given a fixed vector space G, for any two vector spaces M and N and every linear map $f\colon M \to N$, we defined $\tau_G(f) = f \otimes \mathrm{id}_G$ to be the unique linear map making the following diagram commute.

$$\begin{tikzcd}
M \times G \arrow[r, "\iota_{M\otimes}"] \arrow[d, "f\times \mathrm{id}_G"'] & M \otimes G \arrow[d, "f\otimes \mathrm{id}_G"] \\
N \times G \arrow[r, "\iota_{N\otimes}"'] & N \otimes G
\end{tikzcd}$$

See the proof of Proposition 33.13 (3). Show that

(1) $\tau_G(0) = 0$,

(2) $\tau_G(\mathrm{id}_M) = (\mathrm{id}_M \otimes \mathrm{id}_G) = \mathrm{id}_{M\otimes G}$,

(3) If $f'\colon M \to N$ is another linear map, then $\tau_G(f + f') = \tau_G(f) + \tau_G(f')$.

Problem 33.5. Induct on $m \geq 2$ to prove the canonical isomorphism

$$V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes (m+n)}.$$

Use this isomorphism to show that $\cdot\colon V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes (m+n)}$ defined as

$$(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.$$

induces a multiplication on $T(V)$.
Hint. See Jacobson [99], Section 3.9, or Bertin [15], Chapter 4, Section 2.).

Problem 33.6. Prove Proposition 33.19.
Hint. See Knapp [104] (Appendix A, Proposition A.14) or Bertin [15] (Chapter 4, Theorem 2.4).

Problem 33.7. Given linear maps $f\colon E \to E'$ and $f'\colon E' \to E''$, show that

$$(f' \circ f) \odot (f' \circ f) = (f' \odot f') \circ (f \odot f).$$

Problem 33.8. Complete the proof of Proposition 33.28 for the case of an infinite dimensional vector space E.

Problem 33.9. Let I be a finite index set of cardinality p. Let m be a nonnegative integer. Show that the number of multisets over I with cardinality m is $\binom{p+m-1}{m}$.

Problem 33.10. Prove Proposition 33.29.

Problem 33.11. Using bases, show that the bilinear map at $(*)$ in Section 33.10 produces a nondegenerate pairing.

Problem 33.12. Let $\mathfrak{I}$ be the two-sided ideal generated by all tensors of the form $u \otimes v - v \otimes u \in V^{\otimes 2}$. Prove that $\mathrm{S}^m(V) \cong V^{\otimes m}/(\mathfrak{I} \cap V^{\otimes m})$.

Problem 33.13. Verify Equation $(*)$ of Section 33.11 for arbitrary nonnegative integers m and n.

Chapter 34

Exterior Tensor Powers and Exterior Algebras

34.1 Exterior Tensor Powers

In this chapter we consider *alternating* (also called *skew-symmetric*) multilinear maps and *exterior tensor powers* (also called *alternating tensor powers*), denoted $\bigwedge^n(E)$. In many respects alternating multilinear maps and exterior tensor powers can be treated much like symmetric tensor powers, except that $\mathrm{sgn}(\sigma)$ needs to be inserted in front of the formulae valid for symmetric powers.

Roughly speaking, we are now in the world of determinants rather than in the world of permanents. However, there are also some fundamental differences, one of which being that the exterior tensor power $\bigwedge^n(E)$ is the trivial vector space (0) when E is finite-dimensional and when $n > \dim(E)$. This chapter provides the firm foundations for understanding differential forms.

As in the case of symmetric tensor powers, since we already have the tensor algebra $T(V)$, we can proceed rather quickly. But first let us review some basic definitions and facts.

Definition 34.1. Let $f \colon E^n \to F$ be a multilinear map. We say that f *alternating* iff for all $u_i \in E$, $f(u_1, \ldots, u_n) = 0$ whenever $u_i = u_{i+1}$, for some i with $1 \leq i \leq n-1$; that is, $f(u_1, \ldots, u_n) = 0$ whenever two adjacent arguments are identical. We say that f is *skew-symmetric* (or *anti-symmetric*) iff

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \mathrm{sgn}(\sigma) f(u_1, \ldots, u_n),$$

for every permutation $\sigma \in \mathfrak{S}_n$, and all $u_i \in E$.

For $n = 1$, we agree that every linear map $f \colon E \to F$ is alternating. The vector space of all multilinear alternating maps $f \colon E^n \to F$ is denoted $\mathrm{Alt}^n(E; F)$. Note that $\mathrm{Alt}^1(E; F) = \mathrm{Hom}(E, F)$. The following basic proposition shows the relationship between alternation and skew-symmetry.

Proposition 34.1. *Let $f\colon E^n \to F$ be a multilinear map. If f is alternating, then the following properties hold:*

(1) For all i, with $1 \leq i \leq n-1$,

$$f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots).$$

(2) For every permutation $\sigma \in \mathfrak{S}_n$,

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \mathrm{sgn}(\sigma) f(u_1, \ldots, u_n).$$

(3) For all i, j, with $1 \leq i < j \leq n$,

$$f(\ldots, u_i, \ldots u_j, \ldots) = 0 \quad \textit{whenever } u_i = u_j.$$

Moreover, if our field K has characteristic different from 2, then every skew-symmetric multilinear map is alternating.

Proof. (1) By multilinearity applied twice, we have

$$\begin{aligned} f(\ldots, u_i + u_{i+1}, u_i + u_{i+1}, \ldots) = f(\ldots, u_i, u_i, \ldots) &+ f(\ldots, u_i, u_{i+1}, \ldots) \\ &+ f(\ldots, u_{i+1}, u_i, \ldots) + f(\ldots, u_{i+1}, u_{i+1}, \ldots). \end{aligned}$$

Since f is alternating, we get

$$0 = f(\ldots, u_i, u_{i+1}, \ldots) + f(\ldots, u_{i+1}, u_i, \ldots);$$

that is, $f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots)$.

(2) Clearly, the symmetric group, $\mathfrak{S}_n$, acts on $\mathrm{Alt}^n(E; F)$ on the left, *via*

$$\sigma \cdot f(u_1, \ldots, u_n) \ = f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}).$$

Consequently, as $\mathfrak{S}_n$ is generated by the transpositions (permutations that swap exactly two elements), since for a transposition, (2) is simply (1), we deduce (2) by induction on the number of transpositions in σ.

(3) There is a permutation σ that sends u_i and u_j respectively to u_1 and u_2. By hypothesis $u_i = u_j$, so we have $u_{\sigma(1)} = u_{\sigma(2)}$, and as f is alternating we have

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.$$

However, by (2),

$$f(u_1, \ldots, u_n) = \mathrm{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.$$

Now when f is skew-symmetric, if σ is the transposition swapping u_i and $u_{i+1} = u_i$, as $\mathrm{sgn}(\sigma) = -1$, we get

$$f(\ldots, u_i, u_i, \ldots) = -f(\ldots, u_i, u_i, \ldots),$$

so that

$$2f(\ldots, u_i, u_i, \ldots) = 0,$$

and in every characteristic except 2, we conclude that $f(\ldots, u_i, u_i, \ldots) = 0$, namely f is alternating. ☐

Proposition 34.1 shows that in every characteristic except 2, alternating and skew-symmetric multilinear maps are identical. Using Proposition 34.1 we easily deduce the following crucial fact.

Proposition 34.2. *Let $f\colon E^n \to F$ be an alternating multilinear map. For any families of vectors, $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, with $u_i, v_i \in E$, if*

$$v_j = \sum_{i=1}^{n} a_{ij} u_i, \qquad 1 \leq j \leq n,$$

then

$$f(v_1, \ldots, v_n) = \left(\sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma)\, a_{\sigma(1),1} \cdots a_{\sigma(n),n}\right) f(u_1, \ldots, u_n) = \det(A) f(u_1, \ldots, u_n),$$

where A is the $n \times n$ matrix, $A = (a_{ij})$.

Proof. Use Property (ii) of Proposition 34.1. ☐

We are now ready to define and construct exterior tensor powers.

Definition 34.2. An *n-th exterior tensor power* of a vector space E, where $n \geq 1$, is a vector space A together with an alternating multilinear map $\varphi\colon E^n \to A$, such that for every vector space F and for every alternating multilinear map $f\colon E^n \to F$, there is a unique linear map $f_\wedge\colon A \to F$ with

$$f(u_1, \ldots, u_n) = f_\wedge(\varphi(u_1, \ldots, u_n)),$$

for all $u_1, \ldots, u_n \in E$, or for short

$$f = f_\wedge \circ \varphi.$$

Equivalently, there is a unique linear map $f_\wedge$ such that the following diagram commutes:

$$\begin{tikzcd} E^n \arrow[r, "\varphi"] \arrow[dr, "f"'] & A \arrow[d, "f_\wedge"] \\ & F. \end{tikzcd}$$

The above property is called the *universal mapping property* of the exterior tensor power (A, φ).

We now show that any two n-th exterior tensor powers (A_1, φ_1) and (A_2, φ_2) for E are isomorphic.

Proposition 34.3. *Given any two n-th exterior tensor powers (A_1, φ_1) and (A_2, φ_2) for E, there is an isomorphism $h\colon A_1 \to A_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Replace tensor product by n-th exterior tensor power in the proof of Proposition 33.5. ☐

We next give a construction that produces an n-th exterior tensor power of a vector space E.

Theorem 34.4. *Given a vector space E, an n-th exterior tensor power $(\bigwedge^n(E), \varphi)$ for E can be constructed ($n \geq 1$). Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \wedge \cdots \wedge u_n$, the exterior tensor power $\bigwedge^n(E)$ is generated by the vectors $u_1 \wedge \cdots \wedge u_n$, where $u_1, \ldots, u_n \in E$, and for every alternating multilinear map $f\colon E^n \to F$, the unique linear map $f_\wedge\colon \bigwedge^n(E) \to F$ such that $f = f_\wedge \circ \varphi$ is defined by*

$$f_\wedge(u_1 \wedge \cdots \wedge u_n) = f(u_1, \ldots, u_n)$$

on the generators $u_1 \wedge \cdots \wedge u_n$ of $\bigwedge^n(E)$.

Proof sketch. We can give a quick proof using the tensor algebra $T(E)$. Let $\mathfrak{I}_a$ be the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E^{\otimes 2}$. Then let

$$\bigwedge^n(E) = E^{\otimes n}/(\mathfrak{I}_a \cap E^{\otimes n})$$

and let π be the projection $\pi\colon E^{\otimes n} \to \bigwedge^n(E)$. If we let $u_1 \wedge \cdots \wedge u_n = \pi(u_1 \otimes \cdots \otimes u_n)$, it is easy to check that $(\bigwedge^n(E), \wedge)$ satisfies the conditions of Theorem 34.4. ☐

Remark: We can also define

$$\bigwedge(E) = T(E)/\mathfrak{I}_a = \bigoplus_{n \geq 0} \bigwedge^n(E),$$

the *exterior algebra* of E. This is the skew-symmetric counterpart of $\mathrm{S}(E)$, and we will study it a little later.

For simplicity of notation, we may write $\bigwedge^n E$ for $\bigwedge^n(E)$. We also abbreviate "exterior tensor power" as "exterior power." Clearly, $\bigwedge^1(E) \cong E$, and it is convenient to set $\bigwedge^0(E) = K$.

The fact that the map $\varphi\colon E^n \to \bigwedge^n(E)$ is alternating and multilinear can also be expressed as follows:

$$\begin{aligned}
u_1 \wedge \cdots \wedge (u_i + v_i) \wedge \cdots \wedge u_n &= (u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n) \\
&\quad + (u_1 \wedge \cdots \wedge v_i \wedge \cdots \wedge u_n), \\
u_1 \wedge \cdots \wedge (\lambda u_i) \wedge \cdots \wedge u_n &= \lambda(u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n), \\
u_{\sigma(1)} \wedge \cdots \wedge u_{\sigma(n)} &= \mathrm{sgn}(\sigma)\, u_1 \wedge \cdots \wedge u_n,
\end{aligned}$$

for all $\sigma \in \mathfrak{S}_n$.

The map φ from E^n to $\bigwedge^n(E)$ is often denoted $\iota_\wedge$, so that

$$\iota_\wedge(u_1, \ldots, u_n) = u_1 \wedge \cdots \wedge u_n.$$

Theorem 34.4 implies the following result.

Proposition 34.5. *There is a canonical isomorphism*

$$\mathrm{Hom}(\bigwedge^n(E), F) \cong \mathrm{Alt}^n(E; F)$$

between the vector space of linear maps $\mathrm{Hom}(\bigwedge^n(E), F)$ *and the vector space of alternating multilinear maps* $\mathrm{Alt}^n(E; F)$*, given by the linear map* $- \circ \varphi$ *defined by* $\mapsto h \circ \varphi$*, with* $h \in \mathrm{Hom}(\bigwedge^n(E), F)$*. In particular, when* $F = K$*, we get a canonical isomorphism*

$$\left(\bigwedge^n(E)\right)^* \cong \mathrm{Alt}^n(E; K).$$

Definition 34.3. Tensors $\alpha \in \bigwedge^n(E)$ are called *alternating n-tensors* or *alternating tensors of degree n* and we write $\deg(\alpha) = n$. Tensors of the form $u_1 \wedge \cdots \wedge u_n$, where $u_i \in E$, are called *simple (or decomposable) alternating n-tensors*. Those alternating n-tensors that are not simple are often called *compound alternating n-tensors*. Simple tensors $u_1 \wedge \cdots \wedge u_n \in \bigwedge^n(E)$ are also called *n-vectors* and tensors in $\bigwedge^n(E^*)$ are often called *(alternating) n-forms*.

Given a linear map $f\colon E \to E'$, since the map $\iota'_\wedge \circ (f \times f)$ is bilinear and alternating, there is a unique linear map $f \wedge f\colon \bigwedge^2(E) \to \bigwedge^2(E')$ making the following diagram commute:

$$\begin{tikzcd}
E^2 \arrow[r, "\iota_\wedge"] \arrow[d, "f \times f"'] & \bigwedge^2(E) \arrow[d, "f \wedge f"] \\
(E')^2 \arrow[r, "\iota'_\wedge"'] & \bigwedge^2(E').
\end{tikzcd}$$

The map $f \wedge f\colon \bigwedge^2(E) \to \bigwedge^2(E')$ is determined by

$$(f \wedge f)(u \wedge v) = f(u) \wedge f(v).$$

Proposition 34.6. *Given any two linear maps $f\colon E \to E'$ and $f'\colon E' \to E''$, we have*

$$(f' \circ f) \wedge (f' \circ f) = (f' \wedge f') \circ (f \wedge f).$$

The generalization to the alternating product $f \wedge \cdots \wedge f$ of $n \geq 3$ copies of the linear map $f\colon E \to E'$ is immediate, and left to the reader.

34.2 Bases of Exterior Powers

Definition 34.4. Let E be any vector space. For any basis $(u_i)_{i\in\Sigma}$ for E, we assume that some total ordering $\leq$ on the index set Σ has been chosen. Call the pair $((u_i)_{i\in\Sigma}, \leq)$ an *ordered basis*. Then for any nonempty finite subset $I \subseteq \Sigma$, let

$$u_I = u_{i_1} \wedge \cdots \wedge u_{i_m},$$

where $I = \{i_1, \ldots, i_m\}$, with $i_1 < \cdots < i_m$.

Since $\bigwedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$, in view of skew-symmetry, it is clear that the tensors u_I with $|I| = n$ generate $\bigwedge^n(E)$ (where $((u_i)_{i\in\Sigma}, \leq)$ is an ordered basis). Actually they form a basis. To gain an intuitive understanding of this statement, let $m = 2$ and E be a 3-dimensional vector space lexicographically ordered basis $\{e_1, e_2, e_3\}$. We claim that

$$e_1 \wedge e_2, \qquad e_1 \wedge e_3, \qquad e_2 \wedge e_3$$

form a basis for $\bigwedge^2(E)$ since they not only generate $\bigwedge^2(E)$ but are linearly independent. The linear independence is argued as follows: given any vector space F, if w_{12}, w_{13}, w_{23} are any vectors in F, there is an alternating bilinear map $h\colon E^2 \to F$ such that

$$h(e_1, e_2) = w_{12}, \qquad h(e_1, e_3) = w_{13}, \qquad h(e_2, e_3) = w_{23}.$$

Because h yields a unique linear map $h_\wedge\colon \bigwedge^2 E \to F$ such that

$$h_\wedge(e_i \wedge e_j) = w_{ij}, \quad 1 \leq i < j \leq 3,$$

by Proposition 33.4, the vectors

$$e_1 \wedge e_2, \qquad e_1 \wedge e_3, \qquad e_2 \wedge e_3$$

are linearly independent. This suggests understanding how an alternating bilinear function $f\colon E^2 \to F$ is expressed in terms of its values $f(e_i, e_j)$ on the basis vectors (e_1, e_2, e_3). Using bilinearity and alternation, we obtain

$$\begin{aligned} f(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) &= (u_1v_2 - u_2v_1)f(e_1, e_2) + (u_1v_3 - u_3v_1)f(e_1, e_3) \\ &\quad + (u_2v_3 - u_3v_2)f(e_2, e_3). \end{aligned}$$

Therefore, given $w_{12}, w_{13}, w_{23} \in F$, the function h given by

$$\begin{aligned} h(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) = (u_1v_2 - u_2v_1)w_{12} + (u_1v_3 - u_3v_1)w_{13} \\ + (u_2v_3 - u_3v_2)w_{23} \end{aligned}$$

is clearly bilinear and alternating, and by construction $h(e_i, e_j) = w_{ij}$, with $1 \leq i < j \leq 3$ does the job.

We now prove the assertion that tensors u_I with $|I| = n$ generate $\bigwedge^n(E)$ for arbitrary n.

Proposition 34.7. *Given any vector space E, if E has finite dimension $d = \dim(E)$, then for all $n > d$, the exterior power $\bigwedge^n(E)$ is trivial; that is $\bigwedge^n(E) = (0)$. If $n \leq d$ or if E is infinite dimensional, then for every ordered basis $((u_i)_{i\in\Sigma}, \leq)$, the family (u_I) is basis of $\bigwedge^n(E)$, where I ranges over finite nonempty subsets of Σ of size $|I| = n$.*

Proof. First assume that E has finite dimension $d = \dim(E)$ and that $n > d$. We know that $\bigwedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$. If $u_1, \ldots, u_d$ is a basis of E, as every v_i is a linear combination of the u_j, when we expand $v_1 \wedge \cdots \wedge v_n$ using multilinearity, we get a linear combination of the form

$$v_1 \wedge \cdots \wedge v_n = \sum_{(j_1,\ldots,j_n)} \lambda_{(j_1,\ldots,j_n)}\, u_{j_1} \wedge \cdots \wedge u_{j_n},$$

where each $(j_1, \ldots, j_n)$ is some sequence of integers $j_k \in \{1, \ldots, d\}$. As $n > d$, each sequence $(j_1, \ldots, j_n)$ must contain two identical elements. By alternation, $u_{j_1} \wedge \cdots \wedge u_{j_n} = 0$, and so $v_1 \wedge \cdots \wedge v_n = 0$. It follows that $\bigwedge^n(E) = (0)$.

Now assume that either $\dim(E) = d$ and $n \leq d$, or that E is infinite dimensional. The argument below shows that the u_I are nonzero and linearly independent. As usual, let $u_i^* \in E^*$ be the linear form given by

$$u_i^*(u_j) = \delta_{ij}.$$

For any nonempty subset $I = \{i_1, \ldots, i_n\} \subseteq \Sigma$ with $i_1 < \cdots < i_n$, for any n vectors $v_1, \ldots, v_n \in E$, let

$$l_I(v_1, \ldots, v_n) = \det(u_{i_j}^*(v_k)) = \begin{vmatrix} u_{i_1}^*(v_1) & \cdots & u_{i_1}^*(v_n) \\ \vdots & \ddots & \vdots \\ u_{i_n}^*(v_1) & \cdots & u_{i_n}^*(v_n) \end{vmatrix}.$$

If we let the n-tuple $(v_1, \ldots, v_n)$ vary we obtain a map l_I from E^n to K, and it is easy to check that this map is alternating multilinear. Thus l_I induces a unique linear map $L_I \colon \bigwedge^n(E) \to K$ making the following diagram commute.

$$\begin{tikzcd} E^n \arrow[r, "\iota_\wedge"] \arrow[dr, "l_I"'] & \bigwedge^n(E) \arrow[d, "L_I"] \\ & K \end{tikzcd}$$

Observe that for any nonempty finite subset $J \subseteq \Sigma$ with $|J| = n$, we have

$$L_I(u_J) = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{if } I \neq J. \end{cases}$$

Note that when $\dim(E) = d$ and $n \leq d$, or when E is infinite-dimensional, the forms $u^*_{i_1}, \ldots, u^*_{i_n}$ are all distinct, so the above does hold. Since $L_I(u_I) = 1$, we conclude that $u_I \neq 0$. If we have a linear combination

$$\sum_I \lambda_I u_I = 0,$$

where the above sum is finite and involves nonempty finite subset $I \subseteq \Sigma$ with $|I| = n$, for every such I, when we apply L_I we get $\lambda_I = 0$, proving linear independence. $\square$

As a corollary, if E is finite dimensional, say $\dim(E) = d$, and if $1 \leq n \leq d$, then we have

$$\dim(\bigwedge^n(E)) = \binom{n}{d},$$

and if $n > d$, then $\dim(\bigwedge^n(E)) = 0$.

Remark: When $n = 0$, if we set $u_\emptyset = 1$, then $(u_\emptyset) = (1)$ is a basis of $\bigwedge^0(V) = K$.

It follows from Proposition 34.7 that the family $(u_I)_I$ where $I \subseteq \Sigma$ ranges over finite subsets of Σ is a basis of $\bigwedge(V) = \bigoplus_{n \geq 0} \bigwedge^n(V)$.

As a corollary of Proposition 34.7 we obtain the following useful criterion for linear independence.

Proposition 34.8. *For any vector space E, the vectors $u_1, \ldots, u_n \in E$ are linearly independent iff $u_1 \wedge \cdots \wedge u_n \neq 0$.*

Proof. If $u_1 \wedge \cdots \wedge u_n \neq 0$, then $u_1, \ldots, u_n$ must be linearly independent. Otherwise, some u_i would be a linear combination of the other u_j's (with $j \neq i$), and then, as in the proof of Proposition 34.7, $u_1 \wedge \cdots \wedge u_n$ would be a linear combination of wedges in which two vectors are identical, and thus zero.

Conversely, assume that $u_1, \ldots, u_n$ are linearly independent. Then we have the linear forms $u^*_i \in E^*$ such that

$$u^*_i(u_j) = \delta_{i,j} \qquad 1 \leq i, j \leq n.$$

As in the proof of Proposition 34.7, we have a linear map $L_{u_1,\ldots,u_n} \colon \bigwedge^n(E) \to K$ given by

$$L_{u_1,\ldots,u_n}(v_1 \wedge \cdots \wedge v_n) = \det(u^*_j(v_i)) = \begin{vmatrix} u^*_1(v_1) & \cdots & u^*_1(v_n) \\ \vdots & \ddots & \vdots \\ u^*_n(v_1) & \cdots & u^*_n(v_n) \end{vmatrix},$$

for all $v_1 \wedge \cdots \wedge v_n \in \bigwedge^n(E)$. As $L_{u_1,\ldots,u_n}(u_1 \wedge \cdots \wedge u_n) = 1$, we conclude that $u_1 \wedge \cdots \wedge u_n \neq 0$. $\square$

Proposition 34.8 shows that *geometrically every nonzero wedge $u_1 \wedge \cdots \wedge u_n$ corresponds to some oriented version of an n-dimensional subspace of E.*

34.3 Some Useful Isomorphisms for Exterior Powers

We can show the following property of the exterior tensor product, using the proof technique of Proposition 33.13.

Proposition 34.9. *We have the following isomorphism:*

$$\bigwedge^n(E \oplus F) \cong \bigoplus_{k=0}^{n} \bigwedge^k(E) \otimes \bigwedge^{n-k}(F).$$

34.4 Duality for Exterior Powers

In this section *all vector spaces are assumed to have finite dimension*. We define a nondegenerate pairing $\bigwedge^n(E^*) \times \bigwedge^n(E) \longrightarrow K$ as follows: Consider the multilinear map

$$(E^*)^n \times E^n \longrightarrow K$$

given by

$$\begin{aligned}(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma)\, v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n) &= \det(v_j^*(u_i)) \\ &= \begin{vmatrix} v_1^*(u_1) & \cdots & v_1^*(u_n) \\ \vdots & \ddots & \vdots \\ v_n^*(u_1) & \cdots & v_n^*(u_n) \end{vmatrix}.\end{aligned}$$

It is easily checked that this expression is alternating w.r.t. the u_i's and also w.r.t. the v_j^*. For any fixed $(v_1^*, \ldots, v_n^*) \in (E^*)^n$, we get an alternating multilinear map

$$l_{v_1^*, \ldots, v_n^*} \colon (u_1, \ldots, u_n) \mapsto \det(v_j^*(u_i))$$

from E^n to K. The map $l_{v_1^*, \ldots, v_n^*}$ extends uniquely to a linear map $L_{v_1^*, \ldots, v_n^*} \colon \bigwedge^n(E) \to K$ making the following diagram commute:

$$\begin{tikzcd} E^n \arrow[r, "\iota_\wedge"] \arrow[dr, "l_{v_1^*, \ldots, v_n^*}"'] & \bigwedge^n(E) \arrow[d, "L_{v_1^*, \ldots, v_n^*}"] \\ & K. \end{tikzcd}$$

We also have the alternating multilinear map

$$(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*, \ldots, v_n^*}$$

from $(E^*)^n$ to $\mathrm{Hom}(\bigwedge^n(E), K)$, which extends to a linear map L from $\bigwedge^n(E^*)$ to $\mathrm{Hom}(\bigwedge^n(E), K)$ making the following diagram commute:

$$\begin{tikzcd}
(E^*)^n \arrow[r, "\iota_{\wedge^*}"] \arrow[dr] & \bigwedge^n(E^*) \arrow[d, "L"] \\
& \mathrm{Hom}(\bigwedge^n(E), K).
\end{tikzcd}$$

However, in view of the isomorphism

$$\mathrm{Hom}(U \otimes V, W) \cong \mathrm{Hom}(U, \mathrm{Hom}(V, W)),$$

with $U = \bigwedge^n(E^*)$, $V = \bigwedge^n(E)$ and $W = K$, we can view L as a linear map

$$L\colon \bigwedge^n(E^*) \otimes \bigwedge^n(E) \longrightarrow K,$$

which by Proposition 33.8 corresponds to a bilinear map

$$\langle -, - \rangle\colon \bigwedge^n(E^*) \times \bigwedge^n(E) \longrightarrow K. \qquad (*)$$

This pairing is given explicitly in terms of generators by

$$\langle v_1^* \wedge \cdots \wedge v_n^*, u_1, \ldots, u_n \rangle = \det(v_j^*(u_i)).$$

Now this pairing in nondegenerate. This can be shown using bases. Given any basis $(e_1, \ldots, e_m)$ of E, for every basis element $e_{i_1}^* \wedge \cdots \wedge e_{i_n}^*$ of $\bigwedge^n(E^*)$ (with $1 \leq i_1 < \cdots < i_n \leq m$), we have

$$\langle e_{i_1}^* \wedge \cdots \wedge e_{i_n}^*, e_{j_1}, \ldots, e_{j_n} \rangle = \begin{cases} 1 & \text{if } (j_1, \ldots, j_n) = (i_1, \ldots, i_n) \\ 0 & \text{otherwise.} \end{cases}$$

We leave the details as an exercise to the reader. As a consequence we get the following canonical isomorphisms.

Proposition 34.10. *There is a canonical isomorphism*

$$(\bigwedge^n(E))^* \cong \bigwedge^n(E^*).$$

There is also a canonical isomorphism

$$\mu\colon \bigwedge^n(E^*) \cong \mathrm{Alt}^n(E; K)$$

which allows us to interpret alternating tensors over E^* as alternating multilinear maps.

Proof. The second isomorphism follows from the canonical isomorphism $(\bigwedge^n(E))^* \cong \bigwedge^n(E^*)$ and the canonical isomorphism $(\bigwedge^n(E))^* \cong \mathrm{Alt}^n(E; K)$ given by Proposition 34.5. ☐

Remarks:

1. The isomorphism $\mu\colon \bigwedge^n(E^*) \cong \mathrm{Alt}^n(E; K)$ discussed above can be described explicitly as the linear extension of the map given by
$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_j^*(u_i)).$$

2. The canonical isomorphism of Proposition 34.10 holds under more general conditions. Namely, that K is a commutative ring with identity and that E is a finitely-generated projective K-module (see Definition 35.7). See Bourbaki, [25] (Chapter III, §11, Section 5, Proposition 7).

3. Variants of our isomorphism μ are found in the literature. For example, there is a version μ', where
$$\mu' = \frac{1}{n!}\mu,$$
with the factor $\frac{1}{n!}$ added in front of the determinant. Each version has its its own merits and inconveniences. Morita [129] uses μ' because it is more convenient than μ when dealing with characteristic classes. On the other hand, μ' may not be defined for a field with positive characteristic, and when using μ', some extra factor is needed in defining the wedge operation of alternating multilinear forms (see Section 34.5) and for exterior differentiation. The version μ is the one adopted by Warner [186], Knapp [104], Fulton and Harris [68], and Cartan [34, 35].

If $f\colon E \to F$ is any linear map, by transposition we get a linear map $f^\top\colon F^* \to E^*$ given by
$$f^\top(v^*) = v^* \circ f, \qquad v^* \in F^*.$$
Consequently, we have
$$f^\top(v^*)(u) = v^*(f(u)), \qquad \text{for all } u \in E \text{ and all } v^* \in F^*.$$
For any $p \geq 1$, the map
$$(u_1, \ldots, u_p) \mapsto f(u_1) \wedge \cdots \wedge f(u_p)$$
from E^p to $\bigwedge^p F$ is multilinear alternating, so it induces a unique linear map $\bigwedge^p f\colon \bigwedge^p E \to \bigwedge^p F$ making the following diagram commute
$$\begin{tikzcd} E^p \arrow[r, "\iota_\wedge"] \arrow[dr] & \bigwedge^p E \arrow[d, "\bigwedge^p f"] \\ & \bigwedge^p F, \end{tikzcd}$$

and defined on generators by

$$\Big(\bigwedge^p f\Big)(u_1 \wedge \cdots \wedge u_p) = f(u_1) \wedge \cdots \wedge f(u_p).$$

Combining $\bigwedge^p$ and duality, we get a linear map $\bigwedge^p f^\top \colon \bigwedge^p F^* \to \bigwedge^p E^*$ defined on generators by

$$\Big(\bigwedge^p f^\top\Big)(v_1^* \wedge \cdots \wedge v_p^*) = f^\top(v_1^*) \wedge \cdots \wedge f^\top(v_p^*).$$

Proposition 34.11. *If $f\colon E \to F$ is any linear map between two finite-dimensional vector spaces E and F, then*

$$\mu\Big(\Big(\bigwedge^p f^\top\Big)(\omega)\Big)(u_1, \ldots, u_p) = \mu(\omega)(f(u_1), \ldots, f(u_p)), \qquad \omega \in \bigwedge^p F^*, \;\; u_1, \ldots, u_p \in E.$$

Proof. It is enough to prove the formula on generators. By definition of μ, we have

$$\begin{aligned}
\mu\Big(\Big(\bigwedge^p f^\top\Big)(v_1^* \wedge \cdots \wedge v_p^*)\Big)(u_1, \ldots, u_p) &= \mu(f^\top(v_1^*) \wedge \cdots \wedge f^\top(v_p^*))(u_1, \ldots, u_p)\\
&= \det(f^\top(v_j^*)(u_i))\\
&= \det(v_j^*(f(u_i)))\\
&= \mu(v_1^* \wedge \cdots \wedge v_p^*)(f(u_1), \ldots, f(u_p)),
\end{aligned}$$

as claimed. □

Remark: The map $\bigwedge^p f^\top$ is often denoted f^*, although this is an ambiguous notation since p is dropped. Proposition 34.11 gives us the behavior of $\bigwedge^p f^\top$ under the identification of $\bigwedge^p E^*$ and $\mathrm{Alt}^p(E; K)$ *via* the isomorphism μ.

As in the case of symmetric powers, the map from E^n to $\bigwedge^n(E)$ given by $(u_1, \ldots, u_n) \mapsto u_1 \wedge \cdots \wedge u_n$ yields a surjection $\pi\colon E^{\otimes n} \to \bigwedge^n(E)$. Now this map has some section, so there is some injection $\eta\colon \bigwedge^n(E) \to E^{\otimes n}$ with $\pi \circ \eta = \mathrm{id}$. As we saw in Proposition 34.10 there is a canonical isomorphism

$$(\bigwedge^n(E))^* \cong \bigwedge^n(E^*)$$

for any field K, even of positive characteristic. However, if our field K has characteristic 0, then there is a special section having a natural definition involving an antisymmetrization process.

Recall, from Section 33.10 that we have a left action of the symmetric group $\mathfrak{S}_n$ on $E^{\otimes n}$. The tensors $z \in E^{\otimes n}$ such that

$$\sigma \cdot z = \mathrm{sgn}(\sigma)\, z, \quad \text{for all} \quad \sigma \in \mathfrak{S}_n$$

are called *antisymmetrized* tensors. We define the map $\eta\colon E^n \to E^{\otimes n}$ by

$$\eta(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma)\, u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$ [1]

[1] It is the division by $n!$ that requires the field to have characteristic zero.

As the right hand side is an alternating map, we get a unique linear map $\bigwedge^n \eta\colon \bigwedge^n(E) \to E^{\otimes n}$ making the following diagram commute.

$$\begin{tikzcd} E^n \arrow[r, "\iota_\wedge"] \arrow[dr, "\eta"'] & \bigwedge^n(E) \arrow[d, "\bigwedge^n \eta"] \\ & E^{\otimes n}. \end{tikzcd}$$

Clearly, $\bigwedge^n \eta(\bigwedge^n(E))$ is the set of antisymmetrized tensors in $E^{\otimes n}$. If we consider the map $A = (\bigwedge^n \eta) \circ \pi\colon E^{\otimes n} \longrightarrow E^{\otimes n}$, it is easy to check that $A \circ A = A$. Therefore, A is a projection, and by linear algebra, we know that

$$E^{\otimes n} = A(E^{\otimes n}) \oplus \operatorname{Ker} A = \bigwedge^n \eta(\bigwedge^n(E)) \oplus \operatorname{Ker} A.$$

It turns out that $\operatorname{Ker} A = E^{\otimes n} \cap \mathfrak{I}_a = \operatorname{Ker} \pi$, where $\mathfrak{I}_a$ is the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E^{\otimes 2}$ (for example, see Knapp [104], Appendix A). Therefore, $\bigwedge^n \eta$ is injective,

$$E^{\otimes n} = \bigwedge^n \eta(\bigwedge^n(E)) \oplus (E^{\otimes n} \cap \mathfrak{I}_a) = \bigwedge^n \eta(\bigwedge^n(E)) \oplus \operatorname{Ker} \pi,$$

and the exterior tensor power $\bigwedge^n(E)$ is naturally embedded into $E^{\otimes n}$.

34.5 Exterior Algebras

As in the case of symmetric tensors, we can pack together all the exterior powers $\bigwedge^n(V)$ into an algebra.

Definition 34.5. Gieven any vector space V, the vector space

$$\bigwedge(V) = \bigoplus_{m \geq 0} \bigwedge^m(V)$$

is called the *exterior algebra (or Grassmann algebra) of V*.

To make $\bigwedge(V)$ into an algebra, we mimic the procedure used for symmetric powers. If $\mathfrak{I}_a$ is the two-sided ideal generated by all tensors of the form $u \otimes u \in V^{\otimes 2}$, we set

$$\bigwedge^\bullet(V) = T(V)/\mathfrak{I}_a.$$

Then $\bigwedge^\bullet(V)$ automatically inherits a multiplication operation, called *wedge product*, and since $T(V)$ is graded, that is

$$T(V) = \bigoplus_{m \geq 0} V^{\otimes m},$$

we have

$$\overset{\bullet}{\bigwedge}(V) = \bigoplus_{m\geq 0} V^{\otimes m}/(\mathfrak{I}_a \cap V^{\otimes m}).$$

However, it is easy to check that

$$\bigwedge^m(V) \cong V^{\otimes m}/(\mathfrak{I}_a \cap V^{\otimes m}),$$

so

$$\overset{\bullet}{\bigwedge}(V) \cong \bigwedge(V).$$

When V has finite dimension d, we actually have a finite direct sum (coproduct)

$$\bigwedge(V) = \bigoplus_{m=0}^{d} \bigwedge^m(V),$$

and since each $\bigwedge^m(V)$ has dimension $\binom{d}{m}$, we deduce that

$$\dim(\bigwedge(V)) = 2^d = 2^{\dim(V)}.$$

The multiplication, $\wedge\colon \bigwedge^m(V) \times \bigwedge^n(V) \to \bigwedge^{m+n}(V)$, is skew-symmetric in the following precise sense:

Proposition 34.12. *For all $\alpha \in \bigwedge^m(V)$ and all $\beta \in \bigwedge^n(V)$, we have*

$$\beta \wedge \alpha = (-1)^{mn}\alpha \wedge \beta.$$

Proof. Since $v \wedge u = -u \wedge v$ for all $u, v \in V$, Proposition 34.12 follows by induction. $\square$

Since $\alpha \wedge \alpha = 0$ for every *simple* (also called *decomposable*) tensor $\alpha = u_1 \wedge \cdots \wedge u_n$, it seems natural to infer that $\alpha \wedge \alpha = 0$ for *every* tensor $\alpha \in \bigwedge(V)$. If we consider the case where $\dim(V) \leq 3$, we can indeed prove the above assertion. However, if $\dim(V) \geq 4$, the above fact is generally false! For example, when $\dim(V) = 4$, if (u_1, u_2, u_3, u_4) is a basis for V, for $\alpha = u_1 \wedge u_2 + u_3 \wedge u_4$, we check that

$$\alpha \wedge \alpha = 2u_1 \wedge u_2 \wedge u_3 \wedge u_4,$$

which is nonzero. However, if $\alpha \in \bigwedge^m E$ with m odd, since m^2 is also odd, we have

$$\alpha \wedge \alpha = (-1)^{m^2} \alpha \wedge \alpha = -\alpha \wedge \alpha,$$

so indeed $\alpha \wedge \alpha = 0$ (if K is not a field of characteristic 2).

The above discussion suggests that it might be useful to know when an alternating tensor is simple (decomposable). We will show in Section 34.7 that for tensors $\alpha \in \bigwedge^2(V)$, $\alpha \wedge \alpha = 0$ iff α is simple.

A general criterion for decomposability can be given in terms of some operations known as *left hook* and *right hook* (also called *interior products*); see Section 34.7.

It is easy to see that $\bigwedge(V)$ satisfies the following universal mapping property.

Proposition 34.13. *Given any K-algebra A, for any linear map $f\colon V \to A$, if $(f(v))^2 = 0$ for all $v \in V$, then there is a unique K-algebra homomorphism $\overline{f}\colon \bigwedge(V) \to A$ so that*

$$f = \overline{f} \circ i,$$

as in the diagram below.

$$\begin{tikzcd} V \arrow[r, "i"] \arrow[dr, "f"'] & \bigwedge(V) \arrow[d, "\overline{f}"] \\ & A \end{tikzcd}$$

When E is finite-dimensional, recall the isomorphism $\mu\colon \bigwedge^n(E^*) \longrightarrow \mathrm{Alt}^n(E;K)$, defined as the linear extension of the map given by

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_j^*(u_i)).$$

Now, we have also a multiplication operation $\bigwedge^m(E^*) \times \bigwedge^n(E^*) \longrightarrow \bigwedge^{m+n}(E^*)$. The following question then arises:

Can we define a multiplication $\mathrm{Alt}^m(E;K) \times \mathrm{Alt}^n(E;K) \longrightarrow \mathrm{Alt}^{m+n}(E;K)$ directly on alternating multilinear forms, so that the following diagram commutes?

$$\begin{tikzcd} \bigwedge^m(E^*) \times \bigwedge^n(E^*) \arrow[r, "\wedge"] \arrow[d, "\mu_m \times \mu_n"] & \bigwedge^{m+n}(E^*) \arrow[d, "\mu_{m+n}"] \\ \mathrm{Alt}^m(E;K) \times \mathrm{Alt}^n(E;K) \arrow[r, "\wedge"] & \mathrm{Alt}^{m+n}(E;K) \end{tikzcd}$$

As in the symmetric case, the answer is *yes*! The solution is to define this multiplication such that, for $f \in \mathrm{Alt}^m(E;K)$ and $g \in \mathrm{Alt}^n(E;K)$,

$$(f \wedge g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}), \quad (**)$$

where shuffle(m,n) consists of all (m,n)-"shuffles;" that is, permutations σ of $\{1, \ldots m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$. For example, when $m = n = 1$, we have

$$(f \wedge g)(u, v) = f(u)g(v) - g(u)f(v).$$

When $m = 1$ and $n \geq 2$, check that

$$(f \wedge g)(u_1, \ldots, u_{m+1}) = \sum_{i=1}^{m+1} (-1)^{i-1} f(u_i) g(u_1, \ldots, \widehat{u_i}, \ldots, u_{m+1}),$$

where the hat over the argument u_i means that it should be omitted.

Here is another explicit example. Suppose $m = 2$ and $n = 1$. Given $v_1^*, v_2^*, v_3^* \in E^*$, the multiplication structure on $\bigwedge(E^*)$ implies that $(v_1^* \wedge v_2^*) \cdot v_3^* = v_1^* \wedge v_2^* \wedge v_3^* \in \bigwedge^3(E^*)$. Furthermore, for $u_1, u_2, u_3, \in E$,

$$\begin{aligned}
\mu_3(v_1^* \wedge v_2^* \wedge v_3^*)(u_1, u_2, u_3) &= \sum_{\sigma \in \mathfrak{S}_3} \mathrm{sgn}(\sigma) v_{\sigma(1)}^*(u_1) v_{\sigma(2)}^*(u_2) v_{\sigma(3)}^*(u_3) \\
&= v_1^*(u_1) v_2^*(u_2) v_3^*(u_3) - v_1^*(u_1) v_3^*(u_2) v_2^*(u_3) \\
&\quad - v_2^*(u_1) v_1^*(u_2) v_3^*(u_3) + v_2^*(u_1) v_3^*(u_2) v_1^*(u_3) \\
&\quad + v_3^*(u_1) v_1^*(u_2) v_2^*(u_3) - v_3^*(u_1) v_2^*(u_2) v_1^*(u_3).
\end{aligned}$$

Now the $(2, 1)$- shuffles of $\{1, 2, 3\}$ are the following three permutations, namely

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

If $f \cong \mu_2(v_1^* \wedge v_2^*)$ and $g \cong \mu_1(v_3^*)$, then $(**)$ implies that

$$\begin{aligned}
(f \cdot g)(u_1, u_2, u_3) &= \sum_{\sigma \in \text{shuffle}(2,1)} \mathrm{sgn}(\sigma) f(u_{\sigma(1)}, u_{\sigma(2)}) g(u_{\sigma(3)}) \\
&= f(u_1, u_2) g(u_3) - f(u_1, u_3) g(u_2) + f(u_2, u_3) g(u_1) \\
&= \mu_2(v_1^* \wedge v_2^*)(u_1, u_2) \mu_1(v_3^*)(u_3) - \mu_2(v_1^* \wedge v_2^*)(u_1, u_3) \mu_1(v_3^*)(u_2) \\
&\quad + \mu_2(v_1^* \wedge v_2^*)(u_2, u_3) \mu_1(v_3^*)(u_1) \\
&= (v_1^*(u_1) v_2^*(u_2) - v_2^*(u_1) v_1^*(u_2)) v_3^*(u_3) \\
&\quad - (v_1^*(u_1) v_2^*(u_3) - v_2^*(u_1) v_1^*(u_3)) v_3^*(u_2) \\
&\quad + (v_1^*(u_2) v_2^*(u_3) - v_2^*(u_2) v_1^*(u_3)) v_3^*(u_1) \\
&= \mu_3(v_1^* \wedge v_2^* \wedge v_3^*)(u_1, u_2, u_3).
\end{aligned}$$

As a result of all this, the direct sum

$$\mathrm{Alt}(E) = \bigoplus_{n \geq 0} \mathrm{Alt}^n(E; K)$$

is an algebra under the above multiplication, and this algebra is isomorphic to $\bigwedge(E^*)$. For the record we state

Proposition 34.14. *When E is finite dimensional, the maps $\mu\colon \bigwedge^n(E^*) \longrightarrow \mathrm{Alt}^n(E;K)$ induced by the linear extensions of the maps given by*

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_j^*(u_i))$$

yield a canonical isomorphism of algebras $\mu\colon \bigwedge(E^*) \longrightarrow \mathrm{Alt}(E)$, where the multiplication in $\mathrm{Alt}(E)$ is defined by the maps $\wedge\colon \mathrm{Alt}^m(E;K) \times \mathrm{Alt}^n(E;K) \longrightarrow \mathrm{Alt}^{m+n}(E;K)$, with

$$(f \wedge g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}),$$

where $\mathrm{shuffle}(m,n)$ *consists of all (m,n)-"shuffles," that is, permutations σ of $\{1, \ldots m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.*

Remark: The algebra $\bigwedge(E)$ is a graded algebra. Given two graded algebras E and F, we can make a new tensor product $E \mathbin{\widehat{\otimes}} F$, where $E \mathbin{\widehat{\otimes}} F$ is equal to $E \otimes F$ as a vector space, but with a skew-commutative multiplication given by

$$(a \otimes b) \wedge (c \otimes d) = (-1)^{\deg(b)\deg(c)}(ac) \otimes (bd),$$

where $a \in E^m, b \in F^p$, $c \in E^n, d \in F^q$. Then, it can be shown that

$$\bigwedge(E \oplus F) \cong \bigwedge(E) \mathbin{\widehat{\otimes}} \bigwedge(F).$$

34.6 The Hodge ∗-Operator

In order to define a generalization of the Laplacian that applies to differential forms on a Riemannian manifold, we need to define isomorphisms

$$\bigwedge^k V \longrightarrow \bigwedge^{n-k} V,$$

for any Euclidean vector space V of dimension n and any k, with $0 \leq k \leq n$. If $\langle -, - \rangle$ denotes the inner product on V, we define an inner product on $\bigwedge^k V$, denoted $\langle -, - \rangle_\wedge$, by setting

$$\langle u_1 \wedge \cdots \wedge u_k, v_1 \wedge \cdots \wedge v_k \rangle_\wedge = \det(\langle u_i, v_j \rangle),$$

for all $u_i, v_i \in V$, and extending $\langle -, - \rangle_\wedge$ by bilinearity.

It is easy to show that if $(e_1, \ldots, e_n)$ is an orthonormal basis of V, then the basis of $\bigwedge^k V$ consisting of the e_I (where $I = \{i_1, \ldots, i_k\}$, with $1 \leq i_1 < \cdots < i_k \leq n$) is an orthonormal basis of $\bigwedge^k V$. Since the inner product on V induces an inner product on V^* (recall that $\langle \omega_1, \omega_2 \rangle = \langle \omega_1^\sharp, \omega_2^\sharp \rangle$, for all $\omega_1, \omega_2 \in V^*$), we also get an inner product on $\bigwedge^k V^*$.

Definition 34.6. An *orientation* of a vector space V of dimension n is given by the choice of some basis $(e_1, \ldots, e_n)$. We say that a basis $(u_1, \ldots, u_n)$ of V is *positively oriented* iff $\det(u_1, \ldots, u_n) > 0$ (where $\det(u_1, \ldots, u_n)$ denotes the determinant of the matrix whose jth column consists of the coordinates of u_j over the basis $(e_1, \ldots, e_n)$), otherwise it is *negatively oriented*. An *oriented vector space* is a vector space V together with an orientation of V.

If V is oriented by the basis $(e_1, \ldots, e_n)$, then V^* is oriented by the dual basis $(e_1^*, \ldots, e_n^*)$. If σ is any permutation of $\{1, \ldots, n\}$, then the basis $(e_{\sigma(1)}, \ldots, e_{\sigma(n)})$ has positive orientation iff the signature $\mathrm{sgn}(\sigma)$ of the permutation σ is even.

If V is an oriented vector space of dimension n, then we can define a linear isomorphism

$$*\colon \bigwedge^k V \to \bigwedge^{n-k} V,$$

called the *Hodge $*$-operator*. The existence of this operator is guaranteed by the following proposition.

Proposition 34.15. *Let V be any oriented Euclidean vector space whose orientation is given by some chosen orthonormal basis $(e_1, \ldots, e_n)$. For any alternating tensor $\alpha \in \bigwedge^k V$, there is a unique alternating tensor $*\alpha \in \bigwedge^{n-k} V$ such that*

$$\alpha \wedge \beta = \langle *\alpha, \beta \rangle_\wedge \, e_1 \wedge \cdots \wedge e_n$$

for all $\beta \in \bigwedge^{n-k} V$. The alternating tensor $*\alpha$ is independent of the choice of the positive orthonormal basis $(e_1, \ldots, e_n)$.

Proof. Since $\bigwedge^n V$ has dimension 1, the alternating tensor $e_1 \wedge \cdots \wedge e_n$ is a basis of $\bigwedge^n V$. It follows that for any fixed $\alpha \in \bigwedge^k V$, the linear map λ_α from $\bigwedge^{n-k} V$ to $\bigwedge^n V$ given by

$$\lambda_\alpha(\beta) = \alpha \wedge \beta$$

is of the form

$$\lambda_\alpha(\beta) = f_\alpha(\beta) \, e_1 \wedge \cdots \wedge e_n$$

for some linear form $f_\alpha \in \left(\bigwedge^{n-k} V\right)^*$. But then, by the duality induced by the inner product $\langle -, - \rangle$ on $\bigwedge^{n-k} V$, there is a unique vector $*\alpha \in \bigwedge^{n-k} V$ such that

$$f_\lambda(\beta) = \langle *\alpha, \beta \rangle_\wedge \quad \text{for all } \beta \in \bigwedge^{n-k} V,$$

which implies that

$$\alpha \wedge \beta = \lambda_\alpha(\beta) = f_\alpha(\beta) \, e_1 \wedge \cdots \wedge e_n = \langle *\alpha, \beta \rangle_\wedge \, e_1 \wedge \cdots \wedge e_n,$$

as claimed. If $(e'_1, \ldots, e'_n)$ is any other positively oriented orthonormal basis, by Proposition 34.2, $e'_1 \wedge \cdots \wedge e'_n = \det(P) \, e_1 \wedge \cdots \wedge e_n = e_1 \wedge \cdots \wedge e_n$, since $\det(P) = 1$ where P is the change of basis from $(e_1, \ldots, e_n)$ to $(e'_1, \ldots, e'_n)$ and both bases are positively oriented. $\square$

Definition 34.7. The operator $*$ from $\bigwedge^k V$ to $\bigwedge^{n-k} V$ defined by Proposition 34.15 is called the *Hodge $*$-operator*.

Obseve that the Hodge $*$-operator is linear.

The Hodge $*$-operator is defined in terms of the orthonormal basis elements of $\bigwedge V$ as follows: For any increasing sequence $(i_1, \ldots, i_k)$ of elements $i_p \in \{1, \ldots, n\}$, if $(j_1, \ldots, j_{n-k})$ is the increasing sequence of elements $j_q \in \{1, \ldots, n\}$ such that

$$\{i_1, \ldots, i_k\} \cup \{j_1, \ldots, j_{n-k}\} = \{1, \ldots, n\},$$

then

$$*(e_{i_1} \wedge \cdots \wedge e_{i_k}) = \mathrm{sign}(i_1, \ldots i_k, j_1, \ldots, j_{n-k})\, e_{j_1} \wedge \cdots \wedge e_{j_{n-k}}.$$

In particular, for $k = 0$ and $k = n$, we have

$$\begin{aligned} *(1) &= e_1 \wedge \cdots \wedge e_n \\ *(e_1 \wedge \cdots \wedge e_n) &= 1. \end{aligned}$$

For example, if $n = 3$, we have

$$\begin{aligned} *e_1 &= e_2 \wedge e_3 \\ *e_2 &= -e_1 \wedge e_3 \\ *e_3 &= e_1 \wedge e_2 \\ *(e_1 \wedge e_2) &= e_3 \\ *(e_1 \wedge e_3) &= -e_2 \\ *(e_2 \wedge e_3) &= e_1. \end{aligned}$$

The Hodge $*$-operators $*\colon \bigwedge^k V \to \bigwedge^{n-k} V$ induce a linear map $*\colon \bigwedge(V) \to \bigwedge(V)$. We also have Hodge $*$-operators $*\colon \bigwedge^k V^* \to \bigwedge^{n-k} V^*$.

The following proposition shows that the linear map $*\colon \bigwedge(V) \to \bigwedge(V)$ is an isomorphism.

Proposition 34.16. *If V is any oriented vector space of dimension n, for every k with $0 \leq k \leq n$, we have*

(i) $** = (-\mathrm{id})^{k(n-k)}$.

(ii) $\langle x, y \rangle_\wedge = *(x \wedge *y) = *(y \wedge *x)$, *for all* $x, y \in \bigwedge^k V$.

Proof. (1) Let $(e_i)_{i=1}^n$ is an orthonormal basis of V. It is enough to check the identity on basis elements. We have

$$*(e_{i_1} \wedge \cdots \wedge e_{i_k}) = \mathrm{sign}(i_1, \ldots i_k, j_1, \ldots, j_{n-k})\, e_{j_1} \wedge \cdots \wedge e_{j_{n-k}}$$

and

$$\begin{aligned}
**(e_{i_1} \wedge \cdots \wedge e_{i_k}) &= \mathrm{sign}(i_1, \ldots i_k, j_1, \ldots, j_{n-k}) \; *(e_{j_1} \wedge \cdots \wedge e_{j_{n-k}}) \\
&= \mathrm{sign}(i_1, \ldots i_k, j_1, \ldots, j_{n-k}) \, \mathrm{sign}(j_1, \ldots, j_{n-k}, i_1, \ldots i_k) \, e_{i_1} \wedge \cdots \wedge e_{i_k}.
\end{aligned}$$

It is easy to see that

$$\mathrm{sign}(i_1, \ldots i_k, j_1, \ldots, j_{n-k}) \, \mathrm{sign}(j_1, \ldots, j_{n-k}, i_1, \ldots i_k) \; = (-1)^{k(n-k)},$$

which yields

$$**(e_{i_1} \wedge \cdots \wedge e_{i_k}) = (-1)^{k(n-k)} \, e_{i_1} \wedge \cdots \wedge e_{i_k},$$

as claimed.

(ii) These identities are easily checked on basis elements; see Jost [101], Chapter 2, Lemma 2.1.1. In particular let

$$x = e_{i_1} \wedge \cdots \wedge e_{i_k}, \qquad y = e_{i_j} \wedge \cdots \wedge e_{i_j}, \qquad x, y \in \bigwedge^k V,$$

where $(e_i)_{i=1}^n$ is an orthonormal basis of V. If $x \neq y$, $\langle x, y \rangle_\wedge = 0$ since there is some e_{i_p} of x not equal to any e_{j_q} of y by the orthonormality of the basis, this means the p^{th} row of $(\langle e_{i_l}, e_{j_s} \rangle)$ consists entirely of zeroes. Also $x \neq y$ implies that $y \wedge *x = 0$ since

$$*x = \mathrm{sign}(i_1, \ldots i_k, l_1, \ldots, l_{n-k}) e_{l_1} \wedge \cdots \wedge e_{l_{n-k}},$$

where e_{l_s} is the same as some e_p in y. A similar argument shows that if $x \neq y$, $x \wedge *y = 0$. So now assume $x = y$. Then

$$\begin{aligned}
*(e_{i_1} \wedge \cdots \wedge e_{i_k} \wedge *(e_{i_1} \wedge \cdots \wedge e_{i_k})) &= *(e_1 \wedge e_2 \cdots \wedge e_n) \\
&= 1 = \langle x, x \rangle_\wedge.
\end{aligned}$$

□

It is possible to express $*(1)$ in terms of any basis (not necessarily orthonormal) of V.

Proposition 34.17. *If V is any finite-dimensional oriented vector space, for any basis $(v_!, \ldots, v_n)$ of V, we have*

$$*(1) = \frac{1}{\sqrt{\det(\langle v_i, v_j \rangle)}} \, v_1 \wedge \cdots \wedge v_n.$$

Proof. If $(e_1, \ldots, e_n)$ is an orthonormal basis of V and $(v_1, \ldots, v_n)$ is any other basis of V, then

$$\langle v_1 \wedge \cdots \wedge v_n, v_1 \wedge \cdots \wedge v_n \rangle_\wedge = \det(\langle v_i, v_j \rangle),$$

and since

$$v_1 \wedge \cdots \wedge v_n = \det(A) \, e_1 \wedge \cdots \wedge e_n$$

where A is the matrix expressing the v_j in terms of the e_i, we have

$$\langle v_1 \wedge \cdots \wedge v_n, v_1 \wedge \cdots \wedge v_n \rangle_\wedge = \det(A)^2 \langle e_1 \wedge \cdots \wedge e_n, e_1 \wedge \cdots \wedge e_n \rangle = \det(A)^2.$$

As a consequence, $\det(A) = \sqrt{\det(\langle v_i, v_j \rangle)}$, and

$$v_1 \wedge \cdots \wedge v_n = \sqrt{\det(\langle v_i, v_j \rangle)}\, e_1 \wedge \cdots \wedge e_n,$$

from which it follows that

$$*(1) = \frac{1}{\sqrt{\det(\langle v_i, v_j \rangle)}}\, v_1 \wedge \cdots \wedge v_n$$

(see Jost [101], Chapter 2, Lemma 2.1.3). □

34.7 Left and Right Hooks ⊛

In this section *all vector spaces are assumed to have finite dimension*. Say $\dim(E) = n$. Using our nonsingular pairing

$$\langle -, - \rangle \colon \bigwedge^p E^* \times \bigwedge^p E \longrightarrow K \qquad (1 \leq p \leq n)$$

defined on generators by

$$\langle u_1^* \wedge \cdots \wedge u_p^*, v_1 \wedge \cdots \wedge u_p \rangle = \det(u_i^*(v_j)),$$

we define various contraction operations (partial evaluation operators)

$$\lrcorner \colon \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^* \qquad \text{(left hook)}$$

and

$$\llcorner \colon \bigwedge^{p+q} E^* \times \bigwedge^p E \longrightarrow \bigwedge^q E^* \qquad \text{(right hook)},$$

as well as the versions obtained by replacing E by E^* and E^{**} by E. We begin with the *left interior product or left hook*, $\lrcorner$.

Let $u \in \bigwedge^p E$. For any q such that $p + q \leq n$, multiplication on the right by u is a linear map

$$\wedge_R(u) \colon \bigwedge^q E \longrightarrow \bigwedge^{p+q} E$$

given by

$$v \mapsto v \wedge u$$

where $v \in \bigwedge^q E$. The transpose of $\wedge_R(u)$ yields a linear map

$$(\wedge_R(u))^\top \colon \left(\bigwedge^{p+q} E\right)^* \longrightarrow \left(\bigwedge^{q} E\right)^*,$$

which, using the isomorphisms $\left(\bigwedge^{p+q} E\right)^* \cong \bigwedge^{p+q} E^*$ and $\left(\bigwedge^{q} E\right)^* \cong \bigwedge^{q} E^*$, can be viewed as a map

$$(\wedge_R(u))^\top \colon \bigwedge^{p+q} E^* \longrightarrow \bigwedge^{q} E^*$$

given by

$$z^* \mapsto z^* \circ \wedge_R(u),$$

where $z^* \in \bigwedge^{p+q} E^*$. We denote $z^* \circ \wedge_R(u)$ by $u \lrcorner\, z^*$. In terms of our pairing, the adjoint $u \lrcorner$ of $\wedge_R(u)$ defined by

$$\langle u \lrcorner\, z^*, v\rangle = \langle z^*, \wedge_R(u)(v)\rangle;$$

this in turn leads to the following definition.

Definition 34.8. Let $u \in \bigwedge^p E$ and $z^* \in \bigwedge^{p+q} E^*$. We define $u \lrcorner\, z^* \in \bigwedge^q E^*$ to be q-vector uniquely determined by

$$\langle u \lrcorner\, z^*, v\rangle = \langle z^*, v \wedge u\rangle, \quad \text{for all } v \in \bigwedge^q E.$$

Remark: Note that to be precise the operator

$$\lrcorner \colon \bigwedge^{p} E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^{q} E^*$$

depends of p, q, so we really defined a family of operators $\lrcorner_{p,q}$. This family of operators $\lrcorner_{p,q}$ induces a map

$$\lrcorner \colon \bigwedge E \times \bigwedge E^* \longrightarrow \bigwedge E^*,$$

with

$$\lrcorner_{p,q} \colon \bigwedge^{p} E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^{q} E^*$$

as defined before. The common practice is to omit the subscripts of $\lrcorner$.

It is immediately verified that

$$(u \wedge v) \lrcorner\, z^* = u \lrcorner\, (v \lrcorner\, z^*),$$

for all $u \in \bigwedge^k E, v \in \bigwedge^{p-k} E$, $z^* \in \bigwedge^{p+q} E^*$ since

$$\langle (u \wedge v) \lrcorner\, z^*, w\rangle = \langle z^*, w \wedge u \wedge v\rangle = \langle v \lrcorner\, z^*, w \wedge u\rangle = \langle u \lrcorner\, (v \lrcorner\, z^*), w\rangle,$$

whenever $w \in \bigwedge^q E$. This means that

$$\lrcorner : \bigwedge E \times \bigwedge E^* \longrightarrow \bigwedge E^*$$

is a left action of the (noncommutative) ring $\bigwedge E$ with multiplication $\wedge$ on $\bigwedge E^*$, which makes $\bigwedge E^*$ into a left $\bigwedge E$-module.

By interchanging E and E^* and using the isomorphism

$$\left(\bigwedge^k F\right)^* \cong \bigwedge^k F^*,$$

we can also define some maps

$$\lrcorner : \bigwedge^p E^* \times \bigwedge^{p+q} E \longrightarrow \bigwedge^q E,$$

and make the following definition.

Definition 34.9. Let $u^* \in \bigwedge^p E^*$, and $z \in \bigwedge^{p+q} E$. We define $u^* \lrcorner z \in \bigwedge^q$ as the q-vector uniquely defined by

$$\langle v^* \wedge u^*, z\rangle = \langle v^*, u^* \lrcorner z\rangle, \quad \text{for all } v^* \in \bigwedge^q E^*.$$

As for the previous version, we have a family of operators $\lrcorner_{p,q}$ which define an operator

$$\lrcorner : \bigwedge E^* \times \bigwedge E \longrightarrow \bigwedge E.$$

We easily verify that

$$(u^* \wedge v^*) \lrcorner z = u^* \lrcorner (v^* \lrcorner z),$$

whenever $u^* \in \bigwedge^k E^*$, $v^* \in \bigwedge^{p-k} E^*$, and $z \in \bigwedge^{p+q} E$; so this version of $\lrcorner$ is a left action of the ring $\bigwedge E^*$ on $\bigwedge E$ which makes $\bigwedge E$ into a left $\bigwedge E^*$-module.

In order to proceed any further we need some combinatorial properties of the basis of $\bigwedge^p E$ constructed from a basis $(e_1, \ldots, e_n)$ of E. Recall that for any (nonempty) subset $I \subseteq \{1, \ldots, n\}$, we let

$$e_I = e_{i_1} \wedge \cdots \wedge e_{i_p},$$

where $I = \{i_1, \ldots, i_p\}$ with $i_1 < \cdots < i_p$. We also let $e_\emptyset = 1$.

Given any two nonempty subsets $H, L \subseteq \{1, \ldots, n\}$ both listed in increasing order, say $H = \{h_1 < \ldots < h_p\}$ and $L = \{\ell_1 < \ldots < \ell_q\}$, if H and L are disjoint, let $H \cup L$ be union of H and L considered as the ordered sequence

$$(h_1, \ldots, h_p, \ell_1, \ldots, \ell_q).$$

Then let

$$\rho_{H,L} = \begin{cases} 0 & \text{if } H \cap L \neq \emptyset, \\ (-1)^\nu & \text{if } H \cap L = \emptyset, \end{cases}$$

where
$$\nu = |\{(h, l) \mid (h, l) \in H \times L,\, h > l\}|.$$
Observe that when $H \cap L = \emptyset$, $|H| = p$ and $|L| = q$, the number ν is the number of inversions of the sequence
$$(h_1, \cdots, h_p, \ell_1, \cdots, \ell_q),$$
where an inversion is a pair (h_i, ℓ_j) such that $h_i > \ell_j$.

Unless $p + q = n$, the function whose graph is given by
$$\begin{pmatrix} 1 & \cdots & p & p+1 & \cdots & p+q \\ h_1 & \cdots & h_p & \ell_1 & \cdots & \ell_q \end{pmatrix}$$
is **not** a permutation of $\{1, \ldots, n\}$. We can view ν as a slight generalization of the notion of the number of inversions of a permutation.

Proposition 34.18. *For any basis $(e_1, \ldots, e_n)$ of E the following properties hold:*

(1) If $H \cap L = \emptyset$, $|H| = p$, and $|L| = q$, then
$$\rho_{H,L}\rho_{L,H} = (-1)^{\nu}(-1)^{pq-\nu} = (-1)^{pq}.$$

(2) For $H, L \subseteq \{1, \ldots, m\}$ listed in increasing order, we have
$$e_H \wedge e_L = \rho_{H,L} e_{H \cup L}.$$
Similarly,
$$e_H^* \wedge e_L^* = \rho_{H,L} e_{H \cup L}^*.$$

(3) For the left hook
$$\lrcorner \colon \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*,$$
we have
$$\begin{aligned} e_H \lrcorner\, e_L^* &= 0 \quad \textit{if } H \not\subseteq L \\ e_H \lrcorner\, e_L^* &= \rho_{L-H,H} e_{L-H}^* \quad \textit{if } H \subseteq L. \end{aligned}$$

(4) For the left hook
$$\lrcorner \colon \bigwedge^p E^* \times \bigwedge^{p+q} E \longrightarrow \bigwedge^q E,$$
we have
$$\begin{aligned} e_H^* \lrcorner\, e_L &= 0 \quad \textit{if } H \not\subseteq L \\ e_H^* \lrcorner\, e_L &= \rho_{L-H,H} e_{L-H} \quad \textit{if } H \subseteq L. \end{aligned}$$

Proof. These are proved in Bourbaki [25] (Chapter III, §11, Section 11), but the proofs of (3) and (4) are very concise. We elaborate on the proofs of (2) and (4), the proof of (3) being similar.

In (2) if $H \cap L \neq \emptyset$, then $e_H \wedge e_L$ contains some vector twice and so $e_H \wedge e_L = 0$. Otherwise, $e_H \wedge e_L$ consists of

$$e_{h_1} \wedge \cdots \wedge e_{h_p} \wedge e_{\ell_1} \wedge \cdots \wedge e_{\ell_q},$$

and to order the sequence of indices in increasing order we need to transpose any two indices (h_i, ℓ_j) corresponding to an inversion, which yields $\rho_{H,L} e_{H \cup L}$.

Let us now consider (4). We have $|L| = p + q$ and $|H| = p$, and the q-vector $e_H^* \lrcorner e_L$ is characterized by

$$\langle v^*, e_H^* \lrcorner e_L \rangle = \langle v^* \wedge e_H^*, e_L \rangle$$

for all $v^* \in \bigwedge^q E^*$. There are two cases.

Case 1: $H \not\subseteq L$. If so, no matter what $v^* \in \bigwedge^q E^*$ is, since H contains some index h not in L, the hth row $(e_h^*(e_{\ell_1}), \ldots, e_h^*(e_{\ell_{p+q}}))$ of the determinant $\langle v^* \wedge e_H^*, e_L \rangle$ must be zero, so $\langle v^* \wedge e_H^*, e_L \rangle = 0$ for all $v^* \in \bigwedge^q E^*$, and since the pairing is nongenerate, we must have $e_H^* \lrcorner e_L = 0$.

Case 2: $H \subseteq L$. In this case, for $v^* = e_{L-H}^*$, by (2) we have

$$\langle e_{L-H}^*, e_H^* \lrcorner e_L \rangle = \langle e_{L-H}^* \wedge e_H^*, e_L \rangle = \langle \rho_{L-H,H} e_L^*, e_L \rangle = \rho_{L-H,H},$$

which yields

$$\langle e_{L-H}^*, e_H^* \lrcorner e_L \rangle = \rho_{L-H,H}.$$

The q-vector $e_H^* \lrcorner e_L$ can be written as a linear combination $e_H^* \lrcorner e_L = \sum_J \lambda_J e_J$ with $|J| = q$ so

$$\langle e_{L-H}^*, e_H^* \lrcorner e_L \rangle = \sum_J \lambda_J \langle e_{L-H}^*, e_J \rangle.$$

By definition of the pairing, $\langle e_{L-H}^*, e_J \rangle = 0$ unless $J = L - H$, which means that

$$\langle e_{L-H}^*, e_H^* \lrcorner e_L \rangle = \lambda_{L-H} \langle e_{L-H}^*, e_{L-H} \rangle = \lambda_{L-H},$$

so $\lambda_{L-H} = \rho_{L-H,H}$, as claimed. ☐

Using Proposition 34.18, we have the

Proposition 34.19. *For the left hook*

$$\lrcorner : E \times \bigwedge^{q+1} E^* \longrightarrow \bigwedge^q E^*,$$

for every $u \in E$, $x^* \in \bigwedge^{q+1-s} E^*$, and $y^* \in \bigwedge^s E^*$, we have

$$u \lrcorner (x^* \wedge y^*) = (-1)^s (u \lrcorner x^*) \wedge y^* + x^* \wedge (u \lrcorner y^*).$$

Proof. We can prove the above identity assuming that x^* and y^* are of the form e_I^* and e_J^* using Proposition 34.18 and leave the details as an exercise for the reader. $\square$

Thus, $\lrcorner : E \times \bigwedge^{q+1} E^* \longrightarrow \bigwedge^q E^*$ is almost an anti-derivation, except that the sign $(-1)^s$ is applied to the wrong factor.

We have a similar identity for the other version of the left hook

$$\lrcorner : E^* \times \bigwedge^{q+1} E \longrightarrow \bigwedge^{q} E,$$

namely

$$u^* \lrcorner (x \wedge y) = (-1)^s (u^* \lrcorner x) \wedge y + x \wedge (u^* \lrcorner y)$$

for every $u^* \in E^*$, $x \in \bigwedge^{q+1-s} E$, and $y \in \bigwedge^s E$.

An application of this formula when $q = 3$ and $s = 2$ yields an interesting equation. In this case, $u^* \in E^*$ and $x, y \in \bigwedge^2 E$, so we get

$$u^* \lrcorner (x \wedge y) = (u^* \lrcorner x) \wedge y + x \wedge (u^* \lrcorner y).$$

In particular, for $x = y$, since $x \in \bigwedge^2 E$ and $u^* \lrcorner x \in E$, Proposition 34.12 implies that $(u^* \lrcorner x) \wedge x = x \wedge (u^* \lrcorner x)$, and we obtain

$$u^* \lrcorner (x \wedge x) = 2((u^* \lrcorner x) \wedge x). \tag{$\dagger$}$$

As a consequence, $(u^* \lrcorner x) \wedge x = 0$ iff $u^* \lrcorner (x \wedge x) = 0$. We will use this identity together with Proposition 34.25 to prove that a 2-vector $x \in \bigwedge^2 E$ is decomposable iff $x \wedge x = 0$.

It is also possible to define a *right interior product or right hook* $\llcorner$, using multiplication on the left rather than multiplication on the right. Then we use the maps

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*$$

to make the following definition.

Definition 34.10. Let $u \in \bigwedge^p E$ and $z^* \in \bigwedge^{p+q} E^*$. We define $z^* \llcorner u \in \bigwedge^q E^*$ to be the q-vector uniquely defined as

$$\langle z^* \llcorner u, v \rangle = \langle z^*, u \wedge v \rangle, \qquad \text{for all } v \in \bigwedge^q E.$$

This time we can prove that

$$z^* \llcorner (u \wedge v) = (z^* \llcorner u) \llcorner v,$$

so the family of operators $\llcorner_{p,q}$ defines a right action

$$\llcorner : \bigwedge E^* \times \bigwedge E \longrightarrow \bigwedge E^*$$

of the ring $\bigwedge E$ on $\bigwedge E^*$ which makes $\bigwedge E^*$ into a right $\bigwedge E$-module.

Similarly, we have maps

$$\llcorner : \bigwedge^{p+q} E \times \bigwedge^{p} E^* \longrightarrow \bigwedge^{q} E$$

which in turn leads to the following dual formation of the right hook.

Definition 34.11. Let $u^* \in \bigwedge^p E^*$ and $z \in \bigwedge^{p+q} E$. We define $z \llcorner u^* \in \bigwedge^q$ to be the q-vector uniquely defined by

$$\langle u^* \wedge v^*, z\rangle = \langle v^*, z \llcorner u^*\rangle, \qquad \text{for all } v^* \in \textstyle\bigwedge^q E^*.$$

We can prove that

$$z \llcorner (u^* \wedge v^*) = (z \llcorner u^*) \llcorner v^*,$$

so the family of operators $\llcorner_{p,q}$ defines a right action

$$\llcorner : \bigwedge E \times \bigwedge E^* \longrightarrow \bigwedge E$$

of the ring $\bigwedge E^*$ on $\bigwedge E$ which makes $\bigwedge E$ into a right $\bigwedge E^*$-module.

Since the left hook $\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*$ is defined by

$$\langle u \lrcorner z^*, v\rangle = \langle z^*, v \wedge u\rangle, \quad \text{for all } u \in \textstyle\bigwedge^p E,\ v \in \bigwedge^q E \text{ and } z^* \in \bigwedge^{p+q} E^*,$$

the right hook

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*$$

by

$$\langle z^* \llcorner u, v\rangle = \langle z^*, u \wedge v\rangle, \quad \text{for all } u \in \textstyle\bigwedge^p E,\ v \in \bigwedge^q E, \text{ and } z^* \in \bigwedge^{p+q} E^*,$$

and $v \wedge u = (-1)^{pq} u \wedge v$, we conclude that

$$z^* \llcorner u = (-1)^{pq}\, u \lrcorner z^*.$$

Similarly, since

$$\begin{aligned}
\langle v^* \wedge u^*, z\rangle &= \langle v^*, u^* \lrcorner z\rangle, \quad \text{for all } u^* \in \textstyle\bigwedge^p E^*,\ v^* \in \bigwedge^q E^* \text{ and } z \in \bigwedge^{p+q} E\\
\langle u^* \wedge v^*, z\rangle &= \langle v^*, z \llcorner u^*\rangle, \quad \text{for all } u^* \in \textstyle\bigwedge^p E^*,\ v^* \in \bigwedge^q E^*, \text{ and } z \in \bigwedge^{p+q} E,
\end{aligned}$$

and $v^* \wedge u^* = (-1)^{pq} u^* \wedge v^*$, we have

$$z \llcorner u^* = (-1)^{pq}\, u^* \lrcorner z.$$

We summarize the above facts in the following proposition.

Proposition 34.20. *The following identities hold:*

$$
\begin{aligned}
z^* \llcorner u &= (-1)^{pq}\, u \lrcorner z^* \quad \text{for all } u \in \textstyle\bigwedge^p E \text{ and all } z^* \in \bigwedge^{p+q} E^* \\
z \llcorner u^* &= (-1)^{pq}\, u^* \lrcorner z \quad \text{for all } u^* \in \textstyle\bigwedge^p E^* \text{ and all } z \in \bigwedge^{p+q} E.
\end{aligned}
$$

Therefore the left and right hooks are not independent, and in fact each one determines the other. As a consequence, we can restrict our attention to only one of the hooks, for example the left hook, but there are a few situations where it is nice to use both, for example in Proposition 34.23.

A version of Proposition 34.18 holds for right hooks, but beware that the indices in $\rho_{L-H,H}$ are permuted. This permutation has to do with the fact that the left hook and the right hook are related *via* a sign factor.

Proposition 34.21. *For any basis $(e_1, \ldots, e_n)$ of E the following properties hold:*

(1) For the right hook

$$\llcorner \colon \bigwedge^{p+q} E \times \bigwedge^{p} E^* \longrightarrow \bigwedge^{q} E$$

we have

$$
\begin{aligned}
e_L \llcorner e_H^* &= 0 \quad \text{if } H \not\subseteq L \\
e_L \llcorner e_H^* &= \rho_{H,L-H} e_{L-H} \quad \text{if } H \subseteq L.
\end{aligned}
$$

(2) For the right hook

$$\llcorner \colon \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*$$

we have

$$
\begin{aligned}
e_L^* \llcorner e_H &= 0 \quad \text{if } H \not\subseteq L \\
e_L^* \llcorner e_H &= \rho_{H,L-H} e_{L-H}^* \quad \text{if } H \subseteq L.
\end{aligned}
$$

Remark: Our definition of left hooks as left actions $\lrcorner \colon \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*$ and $\lrcorner \colon \bigwedge^p E^* \times \bigwedge^{p+q} E \longrightarrow \bigwedge^q E$ and right hooks as right actions $\llcorner \colon \bigwedge^{p+q} E^* \times \bigwedge^p E \longrightarrow \bigwedge^q E^*$ and $\llcorner \colon \bigwedge^{p+q} E \times \bigwedge^p E^* \longrightarrow \bigwedge^q E$ is identical to the definition found in Fulton and Harris [68] (Appendix B). However, the reader should be aware that this is not a universally accepted notation. In fact, the left hook $u^* \lrcorner z$ defined in Bourbaki [25] is our right hook $z \llcorner u^*$, up to the sign $(-1)^{p(p-1)/2}$. This has to do with the fact that Bourbaki uses a different pairing which also involves an extra sign, namely

$$\langle v^*, u^* \lrcorner z \rangle = (-1)^{p(p-1)/2} \langle u^* \wedge v^*, z \rangle.$$

One of the side-effects of this choice is that Bourbaki's version of Formula (4) of Proposition 34.18 (Bourbaki [25], Chapter III, page 168) is

$$\begin{aligned} &e_H^* \lrcorner e_L = 0 \quad \text{if } H \not\subseteq L \\ &e_H^* \lrcorner e_L = (-1)^{p(p-1)/2} \rho_{H,L-H} e_{L-H} \quad \text{if } H \subseteq L, \end{aligned}$$

where $|H| = p$ and $|L| = p + q$. This correspond to Formula (1) of Proposition 34.21 up to the sign factor $(-1)^{p(p-1)/2}$, which we find horribly confusing. Curiously, an older edition of Bourbaki (1958) uses the same pairing as Fulton and Harris [68]. The reason (and the advantage) for this change of sign convention is not clear to us.

We also have the following version of Proposition 34.19 for the right hook.

Proposition 34.22. *For the right hook*

$$\llcorner : \bigwedge^{q+1} E^* \times E \longrightarrow \bigwedge^{q} E^*,$$

for every $u \in E$, $x^* \in \bigwedge^r E^*$, and $y^* \in \bigwedge^{q+1-r} E^*$, we have

$$(x^* \wedge y^*) \llcorner u = (x^* \llcorner u) \wedge y^* + (-1)^r x^* \wedge (y^* \llcorner u).$$

Proof. A proof involving determinants can be found in Warner [186], Chapter 2. $\square$

Thus, $\llcorner : \bigwedge^{q+1} E^* \times E \longrightarrow \bigwedge^q E^*$ is an anti-derivation. A similar formula holds for the the right hook $\llcorner : \bigwedge^{q+1} E \times E^* \longrightarrow \bigwedge^q E$, namely

$$(x \wedge y) \llcorner u^* = (x \llcorner u^*) \wedge y + (-1)^r x \wedge (y \llcorner u^*),$$

for every $u^* \in E$, $\in \bigwedge^r E$, and $y \in \bigwedge^{q+1-r} E$. This formula is used by Shafarevitch [158] to define a hook, but beware that Shafarevitch use the left hook notation $u^* \lrcorner x$ rather than the right hook notation. Shafarevitch uses the terminology *convolution*, which seems very unfortunate.

For $u \in E$, the right hook $z^* \llcorner u$ is also denoted $i(u)z^*$, and called *insertion operator* or *interior product*. This operator plays an important role in differential geometry.

Definition 34.12. Let $u \in E$ and $z^* \in \bigwedge^{n+1}(E^*)$. If we view z^* as an alternating multilinear map in $\mathrm{Alt}^{n+1}(E; K)$, then we define $i(u)z^* \in \mathrm{Alt}^n(E; K)$ as given by

$$(i(u)z^*)(v_1, \ldots, v_n) = z^*(u, v_1, \ldots, v_n).$$

Using the left hook $\lrcorner$ and the right hook $\llcorner$ we can define two linear maps $\gamma \colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta \colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ as follows:

Definition 34.13. For any basis $(e_1, \ldots, e_n)$ of E, if we let $M = \{1, \ldots, n\}$, $e = e_1 \wedge \cdots \wedge e_n$, and $e^* = e_1^* \wedge \cdots \wedge e_n^*$, define $\gamma\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ as

$$\gamma(u) = u \lrcorner\, e^* \quad \text{and} \quad \delta(v^*) = e \llcorner\, v^*,$$

for all $u \in \bigwedge^p E$ and all $v^* \in \bigwedge^p E^*$.

Proposition 34.23. *The linear maps $\gamma\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ are isomorphims, and $\gamma^{-1} = \delta$. The isomorphisms γ and δ map decomposable vectors to decomposable vectors. Furthermore, if $z \in \bigwedge^p E$ is decomposable, say $z = u_1 \wedge \cdots \wedge u_p$ for some $u_i \in E$, then $\gamma(z) = v_1^* \wedge \cdots \wedge v_{n-p}^*$ for some $v_j^* \in E^*$, and $v_j^*(u_i) = 0$ for all i, j. A similar property holds for $v^* \in \bigwedge^p E^*$ and $\delta(v^*)$. If $(e_1', \ldots, e_n')$ is any other basis of E and $\gamma'\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta'\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ are the corresponding isomorphisms, then $\gamma' = \lambda\gamma$ and $\delta' = \lambda^{-1}\delta$ for some nonzero $\lambda \in K$.*

Proof. Using Propositions 34.18 and 34.21, for any subset $J \subseteq \{1, \ldots, n\} = M$ such that $|J| = p$, we have

$$\gamma(e_J) = e_J \lrcorner\, e^* = \rho_{M-J,J} e_{M-J}^* \quad \text{and} \quad \delta(e_{M-J}^*) = e \llcorner\, e_{M-J}^* = \rho_{M-J,J} e_J.$$

Thus,

$$\delta \circ \gamma(e_J) = \rho_{M-J,J}\rho_{M-J,J} e_J = e_J,$$

since $\rho_{M-J,J} = \pm 1$. A similar result holds for $\gamma \circ \delta$. This implies that

$$\delta \circ \gamma = \mathrm{id} \quad \text{and} \quad \gamma \circ \delta = \mathrm{id}.$$

Thus, γ and δ are inverse isomorphisms.

If $z \in \bigwedge^p E$ is decomposable, then $z = u_1 \wedge \cdots \wedge u_p$ where $u_1, \ldots, u_p$ are linearly independent since $z \neq 0$, and we can pick a basis of E of the form $(u_1, \ldots, u_n)$. Then the above formulae show that

$$\gamma(z) = \pm u_{p+1}^* \wedge \cdots \wedge u_n^*.$$

Since $(u_1^*, \ldots, u_n^*)$ is the dual basis of $(u_1, \ldots, u_n)$, we have $u_i^*(u_j) = \delta_{ij}$, If $(e_1', \ldots, e_n')$ is any other basis of E, because $\bigwedge^n E$ has dimension 1, we have

$$e_1' \wedge \cdots \wedge e_n' = \lambda e_1 \wedge \cdots \wedge e_n$$

for some nonzero $\lambda \in K$, and the rest is trivial. $\square$

Applying Proposition 34.23 to the case where $p = n-1$, the isomorphism $\gamma\colon \bigwedge^{n-1} E \to \bigwedge^1 E^*$ maps indecomposable vectors in $\bigwedge^{n-1} E$ to indecomposable vectors in $\bigwedge^1 E^* = E^*$. But every vector in E^* is decomposable, so every vector in $\bigwedge^{n-1} E$ is decomposable.

Corollary 34.24. *If E is a finite-dimensional vector space, then every vector in $\bigwedge^{n-1} E$ is decomposable.*

34.8 Testing Decomposability ⊛

We are now ready to tackle the problem of finding criteria for decomposability. Such criteria will use the left hook. Once again, in this section *all vector spaces are assumed to have finite dimension*. But before stating our criteria, we need a few preliminary results.

Proposition 34.25. *Given $z \in \bigwedge^p E$ with $z \neq 0$, the smallest vector space $W \subseteq E$ such that $z \in \bigwedge^p W$ is generated by the vectors of the form*

$$u^* \lrcorner\, z, \qquad \text{with } u^* \in \bigwedge^{p-1} E^*.$$

Proof. First let W be any subspace such that $z \in \bigwedge^p(W)$ and let $(e_1, \ldots, e_r, e_{r+1}, \ldots, e_n)$ be a basis of E such that $(e_1, \ldots, e_r)$ is a basis of W. Then, $u^* = \sum_I \lambda_I e_I^*$, where $I \subseteq \{1, \ldots, n\}$ and $|I| = p - 1$, and $z = \sum_J \mu_J e_J$, where $J \subseteq \{1, \ldots, r\}$ and $|J| = p \leq r$. It follows immediately from the formula of Proposition 34.18 (4), namely

$$e_I^* \lrcorner\, e_J = \rho_{J-I,J} e_{J-I},$$

that $u^* \lrcorner\, z \in W$, since $J - I \subseteq \{1, \ldots, r\}$.

Next we prove that if W is the smallest subspace of E such that $z \in \bigwedge^p(W)$, then W is generated by the vectors of the form $u^* \lrcorner\, z$, where $u^* \in \bigwedge^{p-1} E^*$. Suppose not. Then the vectors $u^* \lrcorner\, z$ with $u^* \in \bigwedge^{p-1} E^*$ span a proper subspace U of W. We prove that for every subspace W' of W with $\dim(W') = \dim(W) - 1 = r - 1$, it is not possible that $u^* \lrcorner\, z \in W'$ for all $u^* \in \bigwedge^{p-1} E^*$. But then, as U is a proper subspace of W, it is contained in some subspace W' with $\dim(W') = r - 1$, and we have a contradiction.

Let $w \in W - W'$ and pick a basis of W formed by a basis $(e_1, \ldots, e_{r-1})$ of W' and w. Any $z \in \bigwedge^p(W)$ can be written as $z = z' + w \wedge z''$, where $z' \in \bigwedge^p W'$ and $z'' \in \bigwedge^{p-1} W'$, and since W is the smallest subspace containing z, we have $z'' \neq 0$. Consequently, if we write $z'' = \sum_I \lambda_I e_I$ in terms of the basis $(e_1, \ldots, e_{r-1})$ of W', there is some e_I, with $I \subseteq \{1, \ldots, r - 1\}$ and $|I| = p - 1$, so that the coefficient λ_I is nonzero. Now, using any basis of E containing $(e_1, \ldots, e_{r-1}, w)$, by Proposition 34.18 (4), we see that

$$e_I^* \lrcorner\, (w \wedge e_I) = \lambda w, \qquad \lambda = \pm 1.$$

It follows that

$$e_I^* \lrcorner\, z = e_I^* \lrcorner\, (z' + w \wedge z'') = e_I^* \lrcorner\, z' + e_I^* \lrcorner\, (w \wedge z'') = e_I^* \lrcorner\, z' + \lambda \lambda_I w,$$

with $e_I^* \lrcorner\, z' \in W'$, which shows that $e_I^* \lrcorner\, z \notin W'$. Therefore, W is indeed generated by the vectors of the form $u^* \lrcorner\, z$, where $u^* \in \bigwedge^{p-1} E^*$. □

To help understand Proposition 34.25, let E be the vector space with basis $\{e_1, e_2, e_3, e_4\}$ and $z = e_1 \wedge e_2 + e_2 \wedge e_3$. Note that $z \in \bigwedge^2 E$. To find the smallest vector space $W \subseteq E$

such that $z \in \bigwedge^2 W$, we calculate $u^* \lrcorner z$, where $u^* \in \bigwedge^1 E^*$. The multilinearity of $\lrcorner$ implies it is enough to calculate $u^* \lrcorner z$ for $u^* \in \{e_1^*, e_2^*, e_3^*, e_4^*\}$. Proposition 34.18 (4) implies that

$$\begin{aligned} e_1^* \lrcorner z &= e_1^* \lrcorner (e_1 \wedge e_2 + e_2 \wedge e_3) = e_1^* \lrcorner e_1 \wedge e_2 = -e_2 \\ e_2^* \lrcorner z &= e_2^* \lrcorner (e_1 \wedge e_2 + e_2 \wedge e_3) = e_1 - e_3 \\ e_3^* \lrcorner z &= e_3^* \lrcorner (e_1 \wedge e_2 + e_2 \wedge e_3) = e_3^* \lrcorner e_2 \wedge e_3 = e_2 \\ e_4^* \lrcorner z &= e_4^* \lrcorner (e_1 \wedge e_2 + e_2 \wedge e_3) = 0. \end{aligned}$$

Thus W is the two-dimensional vector space generated by the basis $\{e_2, e_1 - e_3\}$. This is not surprising since $z = -e_2 \wedge (e_1 - e_3)$ and is in fact decomposable. As this example demonstrates, the action of the left hook provides a way of extracting a basis of W from z.

Proposition 34.25 implies the following corollary.

Corollary 34.26. *Any nonzero $z \in \bigwedge^p E$ is decomposable iff the smallest subspace W of E such that $z \in \bigwedge^p W$ has dimension p. Furthermore, if $z = u_1 \wedge \cdots \wedge u_p$ is decomposable, then $(u_1, \ldots, u_p)$ is a basis of the smallest subspace W of E such that $z \in \bigwedge^p W$*

Proof. If $\dim(W) = p$, then for any basis $(e_1, \ldots, e_p)$ of W we know that $\bigwedge^p W$ has $e_1 \wedge \cdots \wedge e_p$ has a basis, and thus has dimension 1. Since $z \in \bigwedge^p W$, we have $z = \lambda e_1 \wedge \cdots \wedge e_p$ for some nonzero λ, so z is decomposable.

Conversely assume that $z \in \bigwedge^p W$ is nonzero and decomposable. Then, $z = u_1 \wedge \cdots \wedge u_p$, and since $z \neq 0$, by Proposition 34.8 $(u_1, \ldots, u_p)$ are linearly independent. Then for any $v_i^* = u_1^* \wedge \cdots u_{i-1}^* \wedge u_{i+1}^* \wedge \cdots \wedge u_p^*$ (where u_i^* is omitted), we have

$$v_i^* \lrcorner z = (u_1^* \wedge \cdots u_{i-1}^* \wedge u_{i+1}^* \wedge \cdots \wedge u_p^*) \lrcorner (u_1 \wedge \cdots \wedge u_p) = \pm u_i,$$

so by Proposition 34.25 we have $u_i \in W$ for $i = 1, \ldots, p$. This shows that $\dim(W) \geq p$, but since $z = u_1 \wedge \cdots \wedge u_p$, we have $\dim(W) = p$, which means that $(u_1, \ldots, u_p)$ is a basis of W. ☐

Finally we are ready to state and prove the criterion for decomposability with respect to left hooks.

Proposition 34.27. *Any nonzero $z \in \bigwedge^p E$ is decomposable iff*

$$(u^* \lrcorner z) \wedge z = 0, \qquad \text{for all } u^* \in \textstyle\bigwedge^{p-1} E^*.$$

Proof. First assume that $z \in \bigwedge^p E$ is decomposable. If so, by Corollary 34.26, the smallest subspace W of E such that $z \in \bigwedge^p W$ has dimension p, so we have $z = e_1 \wedge \cdots \wedge e_p$ where $e_1, \ldots, e_p$ form a basis of W. By Proposition 34.25, for every $u^* \in \bigwedge^{p-1} E^*$, we have $u^* \lrcorner z \in W$, so each $u^* \lrcorner z$ is a linear combination of the e_i's, say

$$u^* \lrcorner z = \alpha_1 e_1 + \cdots + \alpha_p e_p,$$

and

$$(u^* \lrcorner z) \wedge z = \sum_{i=1}^{p} \alpha_i e_i \wedge e_1 \wedge \cdots \wedge e_i \wedge \cdots \wedge e_p = 0.$$

Now assume that $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$, and that $\dim(W) = m > p$, where W is the smallest subspace of E such that $z \in \bigwedge^p W$ If $e_1, \ldots, e_m$ is a basis of W, then we have $z = \sum_I \lambda_I e_I$, where $I \subseteq \{1, \ldots, m\}$ and $|I| = p$. Recall that $z \neq 0$, and so, some λ_I is nonzero. By Proposition 34.25, each e_i can be written as $u^* \lrcorner z$ for some $u^* \in \bigwedge^{p-1} E^*$, and since $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$, we get

$$e_j \wedge z = 0 \quad \text{for} \quad j = 1, \ldots, m.$$

By wedging $z = \sum_I \lambda_I e_I$ with each e_j, as $m > p$, we deduce $\lambda_I = 0$ for all I, so $z = 0$, a contradiction. Therefore, $m = p$ and Corollary 34.26 implies that z is decomposable. ☐

As a corollary of Proposition 34.27 we obtain the following fact that we stated earlier without proof.

Proposition 34.28. *Given any vector space E of dimension n, a vector $x \in \bigwedge^2 E$ is decomposable iff $x \wedge x = 0$.*

Proof. Recall that as an application of Proposition 34.19 we proved the formula (†), namely

$$u^* \lrcorner (x \wedge x) = 2((u^* \lrcorner x) \wedge x)$$

for all $x \in \bigwedge^2 E$ and all $u^* \in E^*$. As a consequence, $(u^* \lrcorner x) \wedge x = 0$ iff $u^* \lrcorner (x \wedge x) = 0$. By Proposition 34.27, the 2-vector x is decomposable iff $u^* \lrcorner (x \wedge x) = 0$ for all $u^* \in E^*$ iff $x \wedge x = 0$. Therefore, a 2-vector x is decomposable iff $x \wedge x = 0$. ☐

As an application of Proposition 34.28, assume that $\dim(E) = 3$ and that (e_1, e_2, e_3) is a basis of E. Then any 2-vector $x \in \bigwedge^2 E$ is of the form

$$x = \alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3.$$

We have

$$x \wedge x = (\alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3) \wedge (\alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3) = 0,$$

because all the terms involved are of the form $c\, e_{i_1} \wedge e_{i_2} \wedge e_{i_3} \wedge e_{i_4}$ with $i_1, i_2, i_3, i_4 \in \{1, 2, 3\}$, and so at least two of these indices are identical. Therefore, every 2-vector $x = \alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3$ is decomposable, although this not obvious at first glance. For example,

$$e_1 \wedge e_2 + e_1 \wedge e_3 + e_2 \wedge e_3 = (e_1 + e_2) \wedge (e_2 + e_3).$$

We now show that Proposition 34.27 yields an equational criterion for the decomposability of an alternating tensor $z \in \bigwedge^p E$.

34.9 The Grassmann-Plücker's Equations and Grassmannian Manifolds ⊛

We follow an argument adapted from Bourbaki [25] (Chapter III, §11, Section 13).

Let E be a vector space of dimensions n, let $(e_1, \ldots, e_n)$ be a basis of E, and let $(e_1^*, \ldots, e_n^*)$ be its dual basis. Our objective is to determine whether a nonzero vector $z \in \bigwedge^p E$ is decomposable. By Proposition 34.27, the vector z is decomposable iff $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$. We can let u^* range over a basis of $\bigwedge^{p-1} E^*$, and then the conditions are

$$(e_H^* \lrcorner z) \wedge z = 0$$

for all $H \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$. Since $(e_H^* \lrcorner z) \wedge z \in \bigwedge^{p+1} E$, this is equivalent to

$$\langle e_J^*, (e_H^* \lrcorner z) \wedge z \rangle = 0$$

for all $H, J \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$ and $|J| = p + 1$. Then, for all $I, I' \subseteq \{1, \ldots, n\}$ with $|I| = |I'| = p$, Formulae (2) and (4) of Proposition 34.18 show that

$$\langle e_J^*, (e_H^* \lrcorner e_I) \wedge e_{I'} \rangle = 0,$$

unless there is some $i \in \{1, \ldots, n\}$ such that

$$I - H = \{i\}, \quad J - I' = \{i\}.$$

In this case, $I = H \cup \{i\}$ and $I' = J - \{i\}$, and using Formulae (2) and (4) of Proposition 34.18, we have

$$\langle e_J^*, (e_H^* \lrcorner e_{H\cup\{i\}}) \wedge e_{J-\{i\}} \rangle = \langle e_J^*, \rho_{\{i\},H} e_i \wedge e_{J-\{i\}} \rangle = \langle e_J^*, \rho_{\{i\},H} \rho_{\{i\},J-\{i\}} e_J \rangle = \rho_{\{i\},H} \rho_{\{i\},J-\{i\}}.$$

If we let

$$\epsilon_{i,J,H} = \rho_{\{i\},H} \rho_{\{i\},J-\{i\}},$$

we have $\epsilon_{i,J,H} = +1$ if the parity of the number of $j \in J$ such that $j < i$ is the same as the parity of the number of $h \in H$ such that $h < i$, and $\epsilon_{i,J,H} = -1$ otherwise.

Finally we obtain the following criterion in terms of quadratic equations (*Plücker's equations*) for the decomposability of an alternating tensor.

Proposition 34.29. *(Grassmann-Plücker's Equations) For $z = \sum_I \lambda_I e_I \in \bigwedge^p E$, the conditions for $z \neq 0$ to be decomposable are*

$$\sum_{i \in J-H} \epsilon_{i,J,H} \lambda_{H\cup\{i\}} \lambda_{J-\{i\}} = 0,$$

with $\epsilon_{i,J,H} = \rho_{\{i\},H} \rho_{\{i\},J-\{i\}}$, for all $H, J \subseteq \{1, \ldots, n\}$ such that $|H| = p-1$, $|J| = p+1$, and all $i \in J - H$.

Using the above criterion, it is a good exercise to reprove that if $\dim(E) = n$, then every tensor in $\bigwedge^{n-1}(E)$ is decomposable. We already proved this fact as a corollary of Proposition 34.23.

Given any $z = \sum_I \lambda_I e_I \in \bigwedge^p E$ where $\dim(E) = n$, the family of scalars (λ_I) (with $I = \{i_1 < \cdots < i_p\} \subseteq \{1, \ldots, n\}$ listed in increasing order) is called the *Plücker coordinates* of z.The Grassmann-Plücker's equations give necessary and sufficient conditions for any nonzero z to be decomposable.

For example, when $\dim(E) = n = 4$ and $p = 2$, these equations reduce to the single equation

$$\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0.$$

However, it should be noted that the equations given by Proposition 34.29 are not independent in general.

We are now in the position to prove that the Grassmannian $G(p, n)$ can be embedded in the projective space $\mathbb{RP}^{\binom{n}{p}-1}$,

For any $n \geq 1$ and any k with $1 \leq p \leq n$, recall that the Grassmannian $G(p, n)$ is the set of all linear p-dimensional subspaces of $\mathbb{R}^n$ (also called *p-planes*). Any p-dimensional subspace U of $\mathbb{R}^n$ is spanned by p linearly independent vectors $u_1, \ldots, u_p$ in $\mathbb{R}^n$; write $U = \operatorname{span}(u_1, \ldots, u_k)$. By Proposition 34.8, $(u_1, \ldots, u_p)$ are linearly independent iff $u_1 \wedge \cdots \wedge u_p \neq 0$. If $(v_1, \ldots, v_p)$ are any other linearly independent vectors spanning U, then we have

$$v_j = \sum_{i=1}^{p} a_{ij} u_i, \quad 1 \leq j \leq p,$$

for some $a_{ij} \in \mathbb{R}$, and by Proposition 34.2

$$v_1 \wedge \cdots \wedge v_p = \det(A)\, u_1 \wedge \cdots \wedge u_p,$$

where $A = (a_{ij})$. As a consequence, we can define a map $i_G \colon G(p, n) \to \mathbb{RP}^{\binom{n}{p}-1}$ such that for any k-plane U, for any basis $(u_1, \ldots, u_p)$ of U,

$$i_G(U) = [u_1 \wedge \cdots \wedge u_p],$$

the point of $\mathbb{RP}^{\binom{n}{p}-1}$ given by the one-dimensional subspace of $\mathbb{R}^{\binom{n}{p}}$ spanned by $u_1 \wedge \cdots \wedge u_p$.

Proposition 34.30. *The map $i_G \colon G(p, n) \to \mathbb{RP}^{\binom{n}{p}-1}$ is injective.*

Proof. Let U and V be any two p-planes and assume that $i_G(U) = i_G(V)$. This means that there is a basis $(u_1, \ldots, u_p)$ of U and a basis $(v_1, \ldots, v_p)$ of V such that

$$v_1 \wedge \cdots \wedge v_p = c\, u_1 \wedge \cdots \wedge u_p$$

for some nonzero $c \in \mathbb{R}$. The above implies that the smallest subspaces W and W' of $\mathbb{R}^n$ such that $u_1 \wedge \cdots \wedge u_p \in \bigwedge^p W$ and $v_1 \wedge \cdots \wedge v_p \in \bigwedge^p W'$ are identical, so $W = W'$. By Corollary 34.26, this smallest subspace W has both $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_p)$ as bases, so the v_j are linear combinations of the u_i (and vice-versa), and $U = V$. $\square$

Since any nonzero $z \in \bigwedge^p \mathbb{R}^n$ can be uniquely written as

$$z = \sum_I \lambda_I e_I$$

in terms of its Plücker coordinates (λ_I), every point of $\mathbb{RP}^{\binom{n}{p}-1}$ is defined by the Plücker coordinates (λ_I) viewed as homogeneous coordinates. The points of $\mathbb{RP}^{\binom{n}{p}-1}$ corresponding to one-dimensional spaces associated with decomposable alternating p-tensors are the points whose coordinates satisfy the Grassmann-Plücker's equations of Proposition 34.29. Therefore, the map i_G embeds the Grassmannian $G(p, n)$ as an algebraic variety in $\mathbb{RP}^{\binom{n}{p}-1}$ defined by equations of degree 2.

We can replace the field $\mathbb{R}$ by $\mathbb{C}$ in the above reasoning and we obtain an embedding of the complex Grassmannian $G_{\mathbb{C}}(p, n)$ as an algebraic variety in $\mathbb{CP}^{\binom{n}{p}-1}$ defined by equations of degree 2.

In particular, if $n = 4$ and $p = 2$, the equation

$$\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0$$

is the homogeneous equation of a quadric in $\mathbb{CP}^5$ known as the *Klein quadric*. The points on this quadric are in one-to-one correspondence with the lines in $\mathbb{CP}^3$.

There is also a simple algebraic criterion to decide whether the smallest subspaces U and V associated with two nonzero decomposable vectors $u_1 \wedge \cdots \wedge u_p$ and $v_1 \wedge \cdots \wedge v_q$ have a nontrivial intersection.

Proposition 34.31. *Let E be any n-dimensional vector space over a field K, and let U and V be the smallest subspaces of E associated with two nonzero decomposable vectors $u = u_1 \wedge \cdots \wedge u_p \in \bigwedge^p U$ and $v = v_1 \wedge \cdots \wedge v_q \in \bigwedge^q V$. The following properties hold:*

(1) We have $U \cap V = (0)$ iff $u \wedge v \neq 0$.

(2) If $U \cap V = (0)$, then $U + V$ is the least subspace associated with $u \wedge v$.

Proof. Assume $U \cap V = (0)$. We know by Corollary 34.26 that $(u_1, \ldots, u_p)$ is a basis of U and $(v_1, \ldots, v_q)$ is a basis of V. Since $U \cap V = (0)$, $(u_1, \ldots, u_p, v_1, \ldots, v_q)$ is a basis of $U + V$, and by Proposition 34.8, we have

$$u \wedge v = u_1 \wedge \cdots \wedge u_p \wedge v_1 \wedge \cdots \wedge v_q \neq 0.$$

This also proves (2).

Conversely, assume that $\dim(U \cap V) \geq 1$. Pick a basis $(w_1, \ldots, w_r)$ of $W = U \cap V$, and extend this basis to a basis $(w_1, \ldots, w_r, w_{r+1}, \ldots, w_p)$ of U and to a basis $(w_1, \ldots, w_r, w_{p+1}, \ldots, w_{p+q-r})$ of V. By Corollary 34.26, $(u_1, \ldots, u_p)$ is also basis of U, so

$$u_1 \wedge \cdots \wedge u_p = a\, w_1 \wedge \cdots \wedge w_r \wedge w_{r+1} \wedge \cdots \wedge w_p$$

for some $a \in K$, and $(v_1, \ldots, v_q)$ is also basis of V, so

$$v_1 \wedge \cdots \wedge v_q = b\, w_1 \cdots \wedge w_r \wedge w_{p+1} \wedge \cdots \wedge w_{p+q-r}$$

for some $b \in K$, and thus

$$u \wedge v = u_1 \wedge \cdots \wedge u_p \wedge v_1 \wedge \cdots \wedge v_q = 0$$

since it contains some repeated w_i, with $1 \leq i \leq r$. $\square$

As an application of Proposition 34.31, consider two projective lines D_1 and D_2 in $\mathbb{RP}^3$, which means that D_1 and D_2 correspond to two 2-planes in $\mathbb{R}^4$, and thus by Proposition 34.30, to two points in $\mathbb{RP}^{\binom{4}{2}-1} = \mathbb{RP}^5$. These two points correspond to the 2-vectors

$$z = a_{1,2}e_1 \wedge e_2 + a_{1,3}e_1 \wedge e_3 + a_{1,4}e_1 \wedge e_4 + a_{2,3}e_2 \wedge e_3 + a_{2,4}e_2 \wedge e_4 + a_{3,4}e_3 \wedge e_4$$

and

$$z' = a'_{1,2}e_1 \wedge e_2 + a'_{1,3}e_1 \wedge e_3 + a'_{1,4}e_1 \wedge e_4 + a'_{2,3}e_2 \wedge e_3 + a'_{2,4}e_2 \wedge e_4 + a'_{3,4}e_3 \wedge e_4$$

whose Plücker coordinates, (where $a_{i,j} = \lambda_{ij}$), satisfy the equation

$$\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0$$

of the Klein quadric, and D_1 and D_2 intersect iff $z \wedge z' = 0$ iff

$$a_{1,2}a'_{3,4} - a_{1,3}a'_{3,4} + a_{1,4}a'_{2,3} + a_{2,3}a'_{1,4} - a_{2,4}a'_{1,3} + a_{3,4}a'_{1,2} = 0.$$

Observe that for D_1 fixed, this is a linear condition. This fact is very helpful for solving problems involving intersections of lines. A famous problem is to find how many lines in $\mathbb{RP}^3$ meet four given lines in general position. The answer is at most 2.

34.10 Vector-Valued Alternating Forms

The purpose of this section is to present the technical background needed to understand vector-valued differential forms, in particular in the case of Lie groups where differential forms taking their values in a Lie algebra arise naturally.

In this section the vector space E is assumed to have *finite dimension*. We know that there is a canonical isomorphism $\bigwedge^n(E^*) \cong \mathrm{Alt}^n(E; K)$ between alternating n-forms and alternating multilinear maps. As in the case of general tensors, the isomorphisms provided by Propositions 34.5, 33.17, and 34.10, namely

$$\begin{aligned}
\mathrm{Alt}^n(E;F) &\cong \mathrm{Hom}\Big(\bigwedge^n(E), F\Big) \\
\mathrm{Hom}\Big(\bigwedge^n(E), F\Big) &\cong \Big(\bigwedge^n(E)\Big)^* \otimes F \\
\Big(\bigwedge^n(E)\Big)^* &\cong \bigwedge^n(E^*)
\end{aligned}$$

yield a canonical isomorphism

$$\mathrm{Alt}^n(E;F) \cong \Big(\bigwedge^n(E^*)\Big) \otimes F$$

which we record as a corollary.

Corollary 34.32. *For any finite-dimensional vecgtor space E and any vector space F, we have a canonical isomorphism*

$$\mathrm{Alt}^n(E;F) \cong \Big(\bigwedge^n(E^*)\Big) \otimes F.$$

Note that F may have infinite dimension. This isomorphism allows us to view the tensors in $\bigwedge^n(E^*) \otimes F$ as *vector-valued alternating forms*, a point of view that is useful in differential geometry. If $(f_1, \ldots, f_r)$ is a basis of F, every tensor $\omega \in \bigwedge^n(E^*) \otimes F$ can be written as some linear combination

$$\omega = \sum_{i=1}^{r} \alpha_i \otimes f_i,$$

with $\alpha_i \in \bigwedge^n(E^*)$.We also let

$$\bigwedge(E;F) = \bigoplus_{n=0} \Big(\bigwedge^n(E^*)\Big) \otimes F = \Big(\bigwedge(E)\Big) \otimes F.$$

Given three vector spaces, F, G, H, if we have some bilinear map $\Phi \colon F \times G \to H$, then we can define a multiplication operation

$$\wedge_\Phi \colon \bigwedge(E;F) \times \bigwedge(E;G) \to \bigwedge(E;H)$$

as follows: For every pair (m,n), we define the multiplication

$$\wedge_\Phi \colon \Big(\Big(\bigwedge^m(E^*)\Big) \otimes F\Big) \times \Big(\Big(\bigwedge^n(E^*)\Big) \otimes G\Big) \longrightarrow \Big(\bigwedge^{m+n}(E^*)\Big) \otimes H$$

by

$$\omega \wedge_\Phi \eta = (\alpha \otimes f) \wedge_\Phi (\beta \otimes g) = (\alpha \wedge \beta) \otimes \Phi(f, g).$$

As in Section 34.5 (following H. Cartan [35]), we can also define a multiplication

$$\wedge_\Phi \colon \mathrm{Alt}^m(E; F) \times \mathrm{Alt}^n(E; G) \longrightarrow \mathrm{Alt}^{m+n}(E; H)$$

directly on alternating multilinear maps as follows: For $f \in \mathrm{Alt}^m(E; F)$ and $g \in \mathrm{Alt}^n(E; G)$,

$$(f \wedge_\Phi g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, \Phi\Big(f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}), g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)})\Big),$$

where shuffle(m, n) consists of all (m, n)-"shuffles;" that is, permutations σ of $\{1, \ldots m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

A special case of interest is the case where $F = G = H$ is a Lie algebra and $\Phi(a, b) = [a, b]$ is the Lie bracket of F. In this case, using a basis $(f_1, \ldots, f_r)$ of F, if we write $\omega = \sum_i \alpha_i \otimes f_i$ and $\eta = \sum_j \beta_j \otimes f_j$, we have

$$\omega \wedge_\Phi \eta = [\omega, \eta] = \sum_{i,j} \alpha_i \wedge \beta_j \otimes [f_i, f_j].$$

It is customary to denote $\omega \wedge_\Phi \eta$ by $[\omega, \eta]$ (unfortunately, the bracket notation is overloaded). Consequently,

$$[\eta, \omega] = (-1)^{mn+1}[\omega, \eta].$$

In general not much can be said about $\wedge_\Phi$, unless Φ has some additional properties. In particular, $\wedge_\Phi$ is generally not associative.

We now use vector-valued alternating forms to generalize both the μ map of Proposition 34.14 and generalize Proposition 33.17 by defining the map

$$\mu_F \colon \left(\bigwedge^n (E^*)\right) \otimes F \longrightarrow \mathrm{Alt}^n(E; F)$$

on generators by

$$\mu_F((v_1^* \wedge \cdots \wedge v_n^*) \otimes f)(u_1, \ldots, u_n) = (\det(v_j^*(u_i))f,$$

with $v_1^*, \ldots, v_n^* \in E^*$, $u_1, \ldots, u_n \in E$, and $f \in F$.

Proposition 34.33. *The map*

$$\mu_F \colon \left(\bigwedge^n (E^*)\right) \otimes F \longrightarrow \mathrm{Alt}^n(E; F)$$

defined as above is a canonical isomorphism for every $n \geq 0$. Furthermore, given any three vector spaces, F, G, H, and any bilinear map $\Phi \colon F \times G \to H$, for all $\omega \in (\bigwedge^n (E^*)) \otimes F$ and all $\eta \in (\bigwedge^n (E^*)) \otimes G$,

$$\mu_H(\omega \wedge_\Phi \eta) = \mu_F(\omega) \wedge_\Phi \mu_G(\eta).$$

Proof. Since we already know that $(\bigwedge^n(E^*))\otimes F$ and $\mathrm{Alt}^n(E;F)$ are isomorphic, it is enough to show that μ_F maps some basis of $(\bigwedge^n(E^*))\otimes F$ to linearly independent elements. Pick some bases $(e_1,\ldots,e_p)$ in E and $(f_j)_{j\in J}$ in F. Then we know that the vectors $e_I^*\otimes f_j$, where $I\subseteq\{1,\ldots,p\}$ and $|I|=n$, form a basis of $(\bigwedge^n(E^*))\otimes F$. If we have a linear dependence

$$\sum_{I,j}\lambda_{I,j}\mu_F(e_I^*\otimes f_j)=0,$$

applying the above combination to each $(e_{i_1},\ldots,e_{i_n})$ $(I=\{i_1,\ldots,i_n\},\ i_1<\cdots<i_n)$, we get the linear combination

$$\sum_j \lambda_{I,j}f_j=0,$$

and by linear independence of the f_j's, we get $\lambda_{I,j}=0$ for all I and all j. Therefore, the $\mu_F(e_I^*\otimes f_j)$ are linearly independent, and we are done. The second part of the proposition is checked using a simple computation. ☐

The following proposition will be useful in dealing with vector-valued differential forms.

Proposition 34.34. *If $(e_1,\ldots,e_p)$ is any basis of E, then every element $\omega\in(\bigwedge^n(E^*))\otimes F$ can be written in a unique way as*

$$\omega=\sum_I e_I^*\otimes f_I,\qquad f_I\in F,$$

where the e_I^* are defined as in Section 34.2.

Proof. Since, by Proposition 34.7, the e_I^* form a basis of $\bigwedge^n(E^*)$, elements of the form $e_I^*\otimes f$ span $(\bigwedge^n(E^*))\otimes F$. Now if we apply $\mu_F(\omega)$ to $(e_{i_1},\ldots,e_{i_n})$, where $I=\{i_1,\ldots,i_n\}\subseteq\{1,\ldots,p\}$, we get

$$\mu_F(\omega)(e_{i_1},\ldots,e_{i_n})=\mu_F(e_I^*\otimes f_I)(e_{i_1},\ldots,e_{i_n})=f_I.$$

Therefore, the f_I are uniquely determined by f. ☐

Proposition 34.34 can also be formulated in terms of alternating multilinear maps, a fact that will be useful to deal with differential forms.

Corollary 34.35. *Define the product* $\cdot\colon \mathrm{Alt}^n(E;\mathbb{R})\times F\to\mathrm{Alt}^n(E;F)$ *as follows: For all* $\omega\in\mathrm{Alt}^n(E;\mathbb{R})$ *and all* $f\in F$,

$$(\omega\cdot f)(u_1,\ldots,u_n)=\omega(u_1,\ldots,u_n)f,$$

for all $u_1,\ldots,u_n\in E$. Then for every $\omega\in(\bigwedge^n(E^*))\otimes F$ of the form

$$\omega=u_1^*\wedge\cdots\wedge u_n^*\otimes f,$$

we have

$$\mu_F(u_1^*\wedge\cdots\wedge u_n^*\otimes f)=\mu_F(u_1^*\wedge\cdots\wedge u_n^*)\cdot f.$$

Then Proposition 34.34 yields the following result.

Proposition 34.36. *If $(e_1, \ldots, e_p)$ is any basis of E, then every element $\omega \in \mathrm{Alt}^n(E; F)$ can be written in a unique way as*

$$\omega = \sum_I e_I^* \cdot f_I, \qquad f_I \in F,$$

where the e_I^* are defined as in Section 34.2.

34.11 Problems

Problem 34.1. Complete the induction argument used in the proof of Proposition 34.1 (2).

Problem 34.2. Prove Proposition 34.2.

Problem 34.3. Prove Proposition 34.9.

Problem 34.4. Show that the pairing given by $(*)$ in Section 34.4 is nondegenerate.

Problem 34.5. Let $\mathfrak{I}_a$ be the two-sided ideal generated by all tensors of the form $u \otimes u \in V^{\otimes 2}$. Prove that

$$\bigwedge^m(V) \cong V^{\otimes m}/(\mathfrak{I}_a \cap V^{\otimes m}).$$

Problem 34.6. Complete the induction proof of Proposition 34.12.

Problem 34.7. Prove the following lemma: If V is a vector space with $\dim(V) \leq 3$, then $\alpha \wedge \alpha = 0$ whenever $\alpha \in \bigwedge(V)$.

Problem 34.8. Prove Proposition 34.13.

Problem 34.9. Given two graded algebras E and F, define $E \mathbin{\widehat{\otimes}} F$ to be the vector space $E \otimes F$, but with a skew-commutative multiplication given by

$$(a \otimes b) \wedge (c \otimes d) = (-1)^{\deg(b)\deg(c)}(ac) \otimes (bd),$$

where $a \in E^m, b \in F^p$, $c \in E^n, d \in F^q$. Show that

$$\bigwedge(E \oplus F) \cong \bigwedge(E) \mathbin{\widehat{\otimes}} \bigwedge(F).$$

Problem 34.10. If $\langle -, - \rangle$ denotes the inner product on V, recall that we defined an inner product on $\bigwedge^k V$, also denoted $\langle -, - \rangle$, by setting

$$\langle u_1 \wedge \cdots \wedge u_k, v_1 \wedge \cdots \wedge v_k \rangle = \det(\langle u_i, v_j \rangle),$$

for all $u_i, v_i \in V$, and extending $\langle -, - \rangle$ by bilinearity.

Show that if $(e_1, \ldots, e_n)$ is an orthonormal basis of V, then the basis of $\bigwedge^k V$ consisting of the e_I (where $I = \{i_1, \ldots, i_k\}$, with $1 \leq i_1 < \cdots < i_k \leq n$) is also an orthonormal basis of $\bigwedge^k V$.

Problem 34.11. Show that

$$(u^* \wedge v^*) \lrcorner z = u^* \lrcorner (v^* \lrcorner z),$$

whenever $u^* \in \bigwedge^k E^*$, $v^* \in \bigwedge^{p-k} E^*$, and $z \in \bigwedge^{p+q} E$.

Problem 34.12. Prove Statement (3) of Proposition 34.18.

Problem 34.13. Prove Proposition 34.19.

Also prove the identity

$$u^* \lrcorner (x \wedge y) = (-1)^s (u^* \lrcorner x) \wedge y + x \wedge (u^* \lrcorner y),$$

where $u^* \in E^*$, $x \in \bigwedge^{q+1-s} E$, and $y \in \bigwedge^s E$.

Problem 34.14. Use the Grassmann-Plücker's equations prove that if $\dim(E) = n$, then every tensor in $\bigwedge^{n-1}(E)$ is decomposable.

Problem 34.15. Recall that the map

$$\mu_F \colon \left(\bigwedge^n (E^*)\right) \otimes F \longrightarrow \mathrm{Alt}^n(E; F)$$

is defined on generators by

$$\mu_F((v_1^* \wedge \cdots \wedge v_n^*) \otimes f)(u_1, \ldots, u_n) = (\det(v_j^*(u_i))f,$$

with $v_1^*, \ldots, v_n^* \in E^*$, $u_1, \ldots, u_n \in E$, and $f \in F$.

Given any three vector spaces, F, G, H, and any bilinear map $\Phi \colon F \times G \to H$, for all $\omega \in (\bigwedge^n (E^*)) \otimes F$ and all $\eta \in (\bigwedge^n (E^*)) \otimes G$ prove that

$$\mu_H(\omega \wedge_\Phi \eta) = \mu_F(\omega) \wedge_\Phi \mu_G(\eta).$$

Chapter 35

Introduction to Modules; Modules over a PID

35.1 Modules over a Commutative Ring

In this chapter we introduce modules over a commutative ring (with unity). After a quick overview of fundamental concepts such as free modules, torsion modules, and some basic results about them, we focus on finitely generated modules over a PID and we prove the structure theorems for this class of modules (invariant factors and elementary divisors). Our main goal is not to give a comprehensive exposition of modules, but instead to apply the structure theorem to the $K[X]$-module E_f defined by a linear map f acting on a finite-dimensional vector space E, and to obtain several normal forms for f, including the rational canonical form.

A module is the generalization of a vector space E over a field K obtained replacing the field K by a commutative ring A (with unity 1). Although formally the definition is the same, the fact that some nonzero elements of A are not invertible has some serious consequences. For example, it is possible that $\lambda \cdot u = 0$ for some nonzero $\lambda \in A$ and some nonzero $u \in E$, and a module may no longer have a basis.

For the sake of completeness, we give the definition of a module, although it is the same as Definition 3.1 with the field K replaced by a ring A. In this chapter, *all rings under consideration are assumed to be commutative and to have an identity element* 1.

Definition 35.1. Given a ring A, a *(left) module over* A (or A-*module*) is a set M (of vectors) together with two operations $+\colon M \times M \to M$ (called *vector addition*),[1] and $\cdot\colon A \times M \to M$ (called *scalar multiplication*) satisfying the following conditions for all $\alpha, \beta \in A$ and all $u, v \in M$;

(M0) M is an abelian group w.r.t. $+$, with identity element 0;

[1]The symbol $+$ is overloaded, since it denotes both addition in the ring A and addition of vectors in M. It is usually clear from the context which $+$ is intended.

(M1) $\alpha \cdot (u + v) = (\alpha \cdot u) + (\alpha \cdot v)$;

(M2) $(\alpha + \beta) \cdot u = (\alpha \cdot u) + (\beta \cdot u)$;

(M3) $(\alpha * \beta) \cdot u = \alpha \cdot (\beta \cdot u)$;

(M4) $1 \cdot u = u$.

Given $\alpha \in A$ and $v \in M$, the element $\alpha \cdot v$ is also denoted by αv. The ring A is often called the ring of scalars.

Unless specified otherwise or unless we are dealing with several different rings, in the rest of this chapter, we assume that all A-modules are defined with respect to a fixed ring A. Thus, we will refer to a A-module simply as a module.

From (M0), a module always contains the null vector 0, and thus is nonempty. From (M1), we get $\alpha \cdot 0 = 0$, and $\alpha \cdot (-v) = -(\alpha \cdot v)$. From (M2), we get $0 \cdot v = 0$, and $(-\alpha) \cdot v = -(\alpha \cdot v)$. The ring A itself can be viewed as a module over itself, addition of vectors being addition in the ring, and multiplication by a scalar being multiplication in the ring.

When the ring A is a field, an A-module is a vector space. When $A = \mathbb{Z}$, a $\mathbb{Z}$-module is just an abelian group, with the action given by

$$\begin{aligned} 0 \cdot u &= 0, & \\ n \cdot u &= \underbrace{u + \cdots + u}_{n}, & n > 0 \\ n \cdot u &= -(-n) \cdot u, & n < 0. \end{aligned}$$

All definitions from Section 3.4, linear combinations, linear independence and linear dependence, subspaces renamed as *submodules*, apply unchanged to modules. Proposition 3.5 also holds for the module spanned by a set of vectors. The definition of a basis (Definition 3.6) also applies to modules, but the only result from Section 3.5 that holds for modules is Proposition 3.12. Unfortunately, it is longer true that every module has a basis. For example, for any nonzero integer $n \in \mathbb{Z}$, the $\mathbb{Z}$-module $\mathbb{Z}/n\mathbb{Z}$ has no basis since $n \cdot \overline{x} = 0$ for all $\overline{x} \in \mathbb{Z}/n\mathbb{Z}$. Similarly, $\mathbb{Q}$, as a $\mathbb{Z}$-module, has no basis. Any two distinct nonzero elements p_1/q_1 and p_2/q_2 are linearly dependent, since

$$(p_2 q_1) \left(\frac{p_1}{q_1} \right) - (p_1 q_2) \left(\frac{p_2}{q_2} \right) = 0.$$

Furthermore, the $\mathbb{Z}$-module $\mathbb{Q}$ is not finitely generated. For if $\{p_1/q_1, \cdots p_n/q_n\} \subset \mathbb{Q}$ generated $\mathbb{Q}$, then for any $x = r/s \in \mathbb{Q}$, we have

$$c_1 \frac{p_1}{q_1} + \cdots + c_n \frac{p_n}{q_n} = \frac{r}{s},$$

where $c_i \in \mathbb{Z}$ for $i = 1, \ldots, n$. The left hand side of the preceding line is equivalent to

$$\frac{c_1 p_1 q_2 \cdots q_n + \cdots + c_n p_n q_1 \cdots q_{n-1}}{q_1 q_2 \cdots q_n},$$

where the numerator is an element of the ideal in $\mathbb{Z}$ spanned by $(c_1, c_2, \cdots, c_n)$. Since $\mathbb{Z}$ is a PID, there exists $a \in \mathbb{Z}$ such that (a) is the ideal spanned by $(c_1, c_2, \cdots, c_n)$. Thus

$$c_1 \frac{p_1}{q_1} + \cdots + c_n \frac{p_n}{q_n} = \frac{ma}{q_1 q_2 \cdots q_n} = \frac{r}{s},$$

where $m \in \mathbb{Z}$. Set

$$\frac{a}{q_1 q_2 \cdots q_n} = \frac{a_1}{b}, \qquad (a_1, b) = 1.$$

Then if $\mathbb{Q}$ was a finitely generated $\mathbb{Z}$-module, we deduce that for all $x \in \mathbb{Q}$

$$x = \frac{r}{s} = m \frac{a_1}{b},$$

whenever a_1/b is a fixed rational number, clearly a contradiction. (In particular let $x = 1/p$ where p is a fixed prime $p > b$. If $ma_1/b = 1/p$, then $ma_1 \in \mathbb{Z}$ with $ma_1 = b_1/p$, an impossiblity since $(b_1, p) = 1$ and $p > b_1$.)

Definition 3.11 can be generalized to rings and yields free modules.

Definition 35.2. Given a commutative ring A and any (nonempty) set I, let $A^{(I)}$ be the subset of the cartesian product A^I consisting of all families $(\lambda_i)_{i \in I}$ with finite support of scalars in A.[2] We define addition and multiplication by a scalar as follows:

$$(\lambda_i)_{i \in I} + (\mu_i)_{i \in I} = (\lambda_i + \mu_i)_{i \in I},$$

and

$$\lambda \cdot (\mu_i)_{i \in I} = (\lambda \mu_i)_{i \in I}.$$

It is immediately verified that addition and multiplication by a scalar are well defined. Thus, $A^{(I)}$ is a module. Furthermore, because families with finite support are considered, the family $(e_i)_{i \in I}$ of vectors e_i, defined such that $(e_i)_j = 0$ if $j \neq i$ and $(e_i)_i = 1$, is clearly a basis of the module $A^{(I)}$. When $I = \{1, \ldots, n\}$, we denote $A^{(I)}$ by A^n. The function $\iota\colon I \to A^{(I)}$, such that $\iota(i) = e_i$ for every $i \in I$, is clearly an injection.

Definition 35.3. An A-module M is *free* iff it has a basis.

The module $A^{(I)}$ is a free module.

All definitions from Section 3.7 apply to modules, linear maps, kernel, image, except the definition of rank, which has to be defined differently. Propositions 3.17, 3.18, 3.19, and

[2]Where A^I denotes the set of all functions from I to A.

3.20 hold for modules. However, the other propositions do not generalize to modules. The definition of an isomorphism generalizes to modules. As a consequence, a module is free iff it is isomorphic to a module of the form $A^{(I)}$.

Section 3.8 generalizes to modules. Given a submodule N of a module M, we can define the quotient module M/N.

If $\mathfrak{a}$ is an ideal in A and if M is an A-module, we define $\mathfrak{a}M$ as the set of finite sums of the form

$$a_1 m_1 + \cdots + a_k m_k, \quad a_i \in \mathfrak{a},\ m_i \in M.$$

It is immediately verified that $\mathfrak{a}M$ is a submodule of M.

Interestingly, the part of Theorem 3.11 that asserts that any two bases of a vector space have the same cardinality holds for modules. One way to prove this fact is to "pass" to a vector space by a quotient process.

Theorem 35.1. *For any free module M, any two bases of M have the same cardinality.*

Proof sketch. We give the argument for finite bases, but it also holds for infinite bases. The trick is to pick any maximal ideal $\mathfrak{m}$ in A (whose existence is guaranteed by Theorem C.3). Then, $A/\mathfrak{m}$ is a field, and $M/\mathfrak{m}M$ can be made into a vector space over $A/\mathfrak{m}$; we leave the details as an exercise. If $(u_1, \ldots, u_n)$ is a basis of M, then it is easy to see that the image of this basis is a basis of the vector space $M/\mathfrak{m}M$. By Theorem 3.11, the number n of elements in any basis of $M/\mathfrak{m}M$ is an invariant, so any two bases of M must have the same number of elements. ☐

Definition 35.4. The common number of elements in any basis of a free module is called the *dimension* (or *rank*) of the free module.

One should realize that the notion of linear independence in a module is a little tricky. According to the definition, the one-element sequence (u) consisting of a single nonzero vector is linearly independent if for all $\lambda \in A$, if $\lambda u = 0$ then $\lambda = 0$. However, there are free modules that contain nonzero vectors that are not linearly independent! For example, the ring $A = \mathbb{Z}/6\mathbb{Z}$ viewed as a module over itself has the basis (1), but the zero-divisors, such as 2 or 4, are not linearly independent. Using language introduced in Definition 35.5, a free module may have torsion elements. There are also nonfree modules such that every nonzero vector is linearly independent, such as $\mathbb{Q}$ over $\mathbb{Z}$.

All definitions from Section 4.1 about matrices apply to free modules, and so do all the propositions. Similarly, all definitions from Section 6.1 about direct sums and direct products apply to modules. All propositions that do not involve extending bases still hold. The important Proposition 6.15 survives in the following form.

Proposition 35.2. *Let $f\colon E \to F$ be a surjective linear map between two A-modules with F a free module. Given any basis $(v_1, \ldots, v_r)$ of F, for any r vectors $u_1, \ldots, u_r \in E$ such that $f(u_i) = v_i$ for $i = 1, \ldots, r$, the vectors $(u_1, \ldots, u_r)$ are linearly independent and the module E is the direct sum*

$$E = \mathrm{Ker}\,(f) \oplus U,$$

where U is the free submodule of E spanned by the basis $(u_1, \ldots, u_r)$.

Proof. Pick any $w \in E$, write $f(w)$ over the basis $(v_1, \ldots, v_r)$ as $f(w) = a_1 v_1 + \cdots + a_r v_r$, and let $u = a_1 u_1 + \cdots + a_r u_r$. Observe that

$$\begin{aligned} f(w-u) &= f(w) - f(u) \\ &= a_1 v_1 + \cdots + a_r v_r - (a_1 f(u_1) + \cdots + a_r f(u_r)) \\ &= a_1 v_1 + \cdots + a_r v_r - (a_1 v_1 + \cdots + a_r v_r) \\ &= 0. \end{aligned}$$

Therefore, $h = w - u \in \mathrm{Ker}\,(f)$, and since $w = h + u$ with $h \in \mathrm{Ker}\,(f)$ and $u \in U$, we have $E = \mathrm{Ker}\,(f) + U$.

If $u = a_1 u_1 + \cdots + a_r u_r \in U$ also belongs to $\mathrm{Ker}\,(f)$, then

$$0 = f(u) = f(a_1 u_1 + \cdots + a_r u_r) = a_1 v_1 + \cdots + a_r v_r,$$

and since $(v_1, \ldots, v_r)$ is a basis, $a_i = 0$ for $i = 1, \ldots, r$, which shows that $\mathrm{Ker}\,(f) \cap U = (0)$. Therefore, we have a direct sum

$$E = \mathrm{Ker}\,(f) \oplus U.$$

Finally, if

$$a_1 u_1 + \cdots + a_r u_r = 0,$$

the above reasoning shows that $a_i = 0$ for $i = 1, \ldots, r$, so $(u_1, \ldots, u_r)$ are linearly independent. Therefore, the module U is a free module. $\square$

One should be aware that if we have a direct sum of modules

$$U = U_1 \oplus \cdots \oplus U_m,$$

every vector $u \in U$ can be written is a unique way as

$$u = u_1 + \cdots + u_m,$$

with $u_i \in U_i$ but, unlike the case of vector spaces, this does not imply that any m nonzero vectors $(u_1, \ldots, u_m)$ are linearly independent. For example, we have the direct sum

$$\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$$

where $\mathbb{Z}/2\mathbb{Z}$ is viewed as a $\mathbb{Z}$-modules, but $(1,0)$ and $(0,1)$ are not linearly independent, since

$$2(1,0) + 2(0,1) = (0,0).$$

A useful fact is that every module is a quotient of some free module. Indeed, if M is an A-module, pick any spanning set I for M (such a set exists, for example, $I = M$), and consider the unique homomorphism $\varphi\colon A^{(I)} \to M$ extending the identity function from I to itself. Then we have an isomorphism $A^{(I)}/\mathrm{Ker}\,(\varphi) \approx M$.

In particular, if M is finitely generated, we can pick I to be a finite set of generators, in which case we get an isomorphism $A^n/\mathrm{Ker}\,(\varphi) \approx M$, for some natural number n. A finitely generated module is sometimes called a module of *finite type*.

The case $n = 1$ is of particular interest. A module M is said to be *cyclic* if it is generated by a single element. In this case $M = Ax$, for some $x \in M$. We have the linear map $m_x\colon A \to M$ given by $a \mapsto ax$ for every $a \in A$, and it is obviously surjective since $M = Ax$. Since the kernel $\mathfrak{a} = \mathrm{Ker}\,(m_x)$ of m_x is an ideal in A, we get an isomorphism $A/\mathfrak{a} \approx Ax$. Conversely, for any ideal $\mathfrak{a}$ of A, if $M = A/\mathfrak{a}$, we see that M is generated by the image x of 1 in M, so M is a cyclic module.

The ideal $\mathfrak{a} = \mathrm{Ker}\,(m_x)$ is the set of all $a \in A$ such that $ax = 0$. This is called the *annihilator* of x, and it is the special case of the following more general situation.

Definition 35.5. If M is any A-module, for any subset S of M, the set of all $a \in A$ such that $ax = 0$ for all $x \in S$ is called the *annihilator* of S, and it is denoted by $\mathrm{Ann}(S)$. If $S = \{x\}$, we write $\mathrm{Ann}(x)$ instead of $\mathrm{Ann}(\{x\})$. A nonzero element $x \in M$ is called a *torsion element* iff $\mathrm{Ann}(x) \neq (0)$. The set consisting of all torsion elements in M and 0 is denoted by M_{tor}.

It is immediately verified that $\mathrm{Ann}(S)$ is an ideal of A, and by definition,

$$M_{\mathrm{tor}} = \{x \in M \mid (\exists a \in A,\ a \neq 0)(ax = 0)\}.$$

If a ring has zero divisors, then the set of all torsion elements in an A-module M may not be a submodule of M. For example, if $M = A = \mathbb{Z}/6\mathbb{Z}$, then $M_{\mathrm{tor}} = \{2, 3, 4\}$, but $3 + 4 = 1$ is not a torsion element. Also, a free module may not be torsion-free because there may be torsion elements, as the example of $\mathbb{Z}/6\mathbb{Z}$ as a free module over itself shows.

However, if A is an integral domain, then a free module is torsion-free and M_{tor} is a submodule of M. (Recall that an integral domain is commutative).

Proposition 35.3. *If A is an integral domain, then for any A-module M, the set M_{tor} of torsion elements in M is a submodule of M.*

Proof. If $x, y \in M$ are torsion elements $(x, y \neq 0)$, then there exist some nonzero elements $a, b \in A$ such that $ax = 0$ and $by = 0$. Since A is an integral domain, $ab \neq 0$, and then for all $\lambda, \mu \in A$, we have

$$ab(\lambda x + \mu y) = b\lambda ax + a\mu by = 0.$$

Therefore, M_{tor} is a submodule of M. ☐

The module M_{tor} is called the *torsion submodule* of M. If $M_{\text{tor}} = (0)$, then we say that M is *torsion-free*, and if $M = M_{\text{tor}}$, then we say that M is a *torsion module*.

If M is not finitely generated, then it is possible that $M_{\text{tor}} \neq 0$, yet the annihilator of M_{tor} is reduced to 0. For example, let take the $\mathbb{Z}$-module

$$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/5\mathbb{Z} \times \cdots \times \mathbb{Z}/p\mathbb{Z} \times \cdots ,$$

where p ranges over the set of primes. Call this module M and the set of primes P. Observe that M is generated by $\{\alpha_p\}_{p\in P}$, where α_p is the tuple whose only nonzero entry is $\overline{1}_p$, the generator of $\mathbb{Z}/p\mathbb{Z}$, i.e.,

$$\alpha_p = (\overline{0}, \overline{0}, \overline{0}, \cdots , \overline{1}_p, \overline{0}, \cdots), \qquad \mathbb{Z}/p\mathbb{Z} = \{n \cdot \overline{1}_p\}_{n=0}^{p-1}.$$

In other words, M is not finitely generated. Furthermore, since $p \cdot \overline{1}_p = \overline{0}$, we have $\{\alpha_p\}_{p\in P} \subset M_{\text{tor}}$. However, because p ranges over all primes, the only possible nonzero annihilator of $\{\alpha_p\}_{p\in P}$ would be the product of all the primes. Hence $\text{Ann}(\{\alpha_p\}_{p\in P}) = (0)$. Because of the subset containment, we conclude that $\text{Ann}(M_{\text{tor}}) = (0)$.

However, if M is finitely generated, it is *not* possible that $M_{\text{tor}} \neq 0$, yet the annihilator of M_{tor} is reduced to 0, since if $x_1, \ldots , x_n$ generate M and if $a_1, \ldots , a_n$ annihilate $x_1, \ldots , x_n$, then $a_1 \cdots a_n$ annihilates every element of M.

Proposition 35.4. *If A is an integral domain, then for any A-module M, the quotient module M/M_{tor} is torsion free.*

Proof. Let $\overline{x}$ be an element of M/M_{tor} and assume that $a\overline{x} = 0$ for some $a \neq 0$ in A. This means that $ax \in M_{\text{tor}}$, so there is some $b \neq 0$ in A such that $bax = 0$. Since $a, b \neq 0$ and A is an integral domain, $ba \neq 0$, so $x \in M_{\text{tor}}$, which means that $\overline{x} = 0$. ☐

If A is an integral domain and if F is a free A-module with basis $(u_1, \ldots , u_n)$, then F can be embedded in a K-vector space F_K isomorphic to K^n, where $K = \text{Frac}(A)$ is the fraction field of A. Similarly, any submodule M of F is embedded into a subspace M_K of F_K. Note that any linearly independent vectors $(u_1, \ldots , u_m)$ in the A-module M remain linearly independent in the vector space M_K, because any linear dependence over K is of the form

$$\frac{a_1}{b_1} u_1 + \cdots + \frac{a_m}{b_m} u_m = 0$$

for some $a_i, b_i \in A$, with $b_1 \cdots b_m \neq 0$, so if we multiply by $b_1 \cdots b_m \neq 0$, we get a linear dependence in the A-module M. Then we see that the maximum number of linearly independent vectors in the A-module M is at most n. The maximum number of linearly independent vectors in a finitely generated submodule of a free module (over an integral domain) is called the *rank* of the module M. If $(u_1, \ldots , u_m)$ are linearly independent where

m is the rank of m, then for every nonzero $v \in M$, there are some $a, a_1, \ldots, a_m \in A$, not all zero, such that

$$av = a_1u_1 + \cdots + a_mu_m.$$

We must have $a \neq 0$, since otherwise, linear independence of the u_i would imply that $a_1 = \cdots = a_m = 0$, contradicting the fact that $a, a_1, \ldots, a_m \in A$ are not all zero.

Unfortunately, in general, a torsion-free module is not free. For example, $\mathbb{Q}$ as a $\mathbb{Z}$-module is torsion-free but not free. If we restrict ourselves to finitely generated modules over PID's, then such modules split as the direct sum of their torsion module with a free module, and a torsion module has a nice decomposition in terms of cyclic modules.

The following proposition shows that over a PID, submodules of a free module are free. There are various ways of proving this result. We give a proof due to Lang [109] (see Chapter III, Section 7).

Proposition 35.5. *If A is a PID and if F is a free A-module of dimension n, then every submodule M of F is a free module of dimension at most n.*

Proof. Let $(u_1, \ldots, u_n)$ be a basis of F, and let $M_r = M \cap (Au_1 \oplus \cdots \oplus Au_r)$, the intersection of M with the free module generated by $(u_1, \ldots, u_r)$, for $r = 1, \ldots, n$. We prove by induction on r that each M_r is free and of dimension at most r. Since $M = M_r$ for some r, this will prove our result.

Consider $M_1 = M \cap Au_1$. If $M_1 = (0)$, we are done. Otherwise let

$$\mathfrak{a} = \{a \in A \mid au_1 \in M\}.$$

It is immediately verified that $\mathfrak{a}$ is an ideal, and since A is a PID, $\mathfrak{a} = a_1A$, for some $a_1 \in A$. Since we are assuming that $M_1 \neq (0)$, we have $a_1 \neq 0$, and $a_1u_1 \in M$. If $x \in M_1$, then $x = au_1$ for some $a \in A$, so $a \in a_1A$, and thus $a = ba_1$ for some $b \in A$. It follows that $M_1 = Aa_1u_1$, which is free.

Assume inductively that M_r is free of dimension at most $r < n$, and let

$$\mathfrak{a} = \{a \in A \mid (\exists b_1 \in A) \cdots (\exists b_r \in A)(b_1u_1 + \cdots + b_ru_r + au_{r+1} \in M)\}.$$

It is immediately verified that $\mathfrak{a}$ is an ideal, and since A is a PID, $\mathfrak{a} = a_{r+1}A$, for some $a_{r+1} \in A$. If $a_{r+1} = 0$, then $M_{r+1} = M_r$, and we are done.

If $a_{r+1} \neq 0$, then there is some $v_1 \in Au_1 \oplus \cdots \oplus Au_r$ such that

$$w = v_1 + a_{r+1}u_{r+1} \in M.$$

For any $x \in M_{r+1}$, there is some $v \in Au_1 \oplus \cdots \oplus Au_r$ and some $a \in A$ such that $x = v + au_{r+1}$. Then, $a \in a_{r+1}A$, so there is some $b \in A$ such that $a = ba_{r+1}$. As a consequence

$$x - bw = v - bv_1 \in M_r,$$

and so $x = x - bw + bw$ with $x - bw \in M_r$, which shows that

$$M_{r+1} = M_r + Aw.$$

On the other hand, if $u \in M_r \cap Aw$, then since $w = v_1 + a_{r+1}u_{r+1}$ we have

$$u = bv_1 + ba_{r+1}u_{r+1},$$

for some $b \in A$, with $u, v_1 \in Au_1 \oplus \cdots \oplus Au_r$, and if $b \neq 0$, this yields the nontrivial linear combination

$$bv_1 - u + ba_{r+1}u_{r+1} = 0,$$

contradicting the fact that $(u_1, \ldots, u_{r+1})$ are linearly independent. Therefore,

$$M_{r+1} = M_r \oplus Aw,$$

which shows that M_{r+1} is free of dimension at most $r + 1$. □

The following two examples show why the hypothesis of Proposition 35.5 requires A to be PID. First consider $6\mathbb{Z} = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}\}$ as a free $6\mathbb{Z}$-module with generator $\overline{1}$. The $6\mathbb{Z}$-submodule $\{\overline{0}, \overline{2}, \overline{4}\}$ is not free, even though it is generated by $\overline{2}$ since $\overline{3} \cdot \overline{2} = \overline{0}$. Proposition 35.5 fails since $6\mathbb{Z}$ is not even an integral domain. Next consider $\mathbb{Z}[X]$ as a free $\mathbb{Z}[X]$-module with generator 1. We claim the ideal

$$(2, X) = \{2p(X) + Xq(X) \mid p(X), q(X) \in \mathbb{Z}[X]\},$$

is not a free $\mathbb{Z}[X]$-module. Indeed any two nonzero elements of $(2, X)$, say $s(X)$ and $t(X)$, are linearly dependent since $t(X)s(X) - s(X)t(X) = 0$. Once again Proposition 35.5 fails since $\mathbb{Z}[X]$ is not a PID. See Example 32.1.

Proposition 35.5 implies that if M is a finitely generated module over a PID, then any submodule N of M is also finitely generated.

Indeed, if $(u_1, \ldots, u_n)$ generate M, then we have a surjection $\varphi\colon A^n \to M$ from the free module A^n onto M. The inverse image $\varphi^{-1}(N)$ of N is a submodule of the free module A^n, therefore by Proposition 35.5, $\varphi^{-1}(N)$ is free and finitely generated. This implies that N is finitely generated (and that it has a number of generators $\leq n$).

We can also prove that a finitely generated torsion-free module over a PID is actually free. We will give another proof of this fact later, but the following proof is instructive.

Proposition 35.6. *If A is a PID and if M is a finitely generated module which is torsion-free, then M is free.*

Proof. Let $(y_1, \ldots, y_n)$ be some generators for M, and let $(u_1, \ldots, u_m)$ be a maximal subsequence of $(y_1, \ldots, y_n)$ which is linearly independent. If $m = n$, we are done. Otherwise, due to the maximality of m, for $i = 1, \ldots, n$, there is some $a_i \neq 0$ such that such that

$a_i y_i$ can be expressed as a linear combination of $(u_1, \ldots, u_m)$. If we let $a = a_1 \ldots a_n$, then $a_1 \ldots a_n y_i \in Au_1 \oplus \cdots \oplus Au_m$ for $i = 1, \ldots, n$, which shows that

$$aM \subseteq Au_1 \oplus \cdots \oplus Au_m.$$

Now, A is an integral domain, and since $a_i \neq 0$ for $i = 1, \ldots, n$, we have $a = a_1 \ldots a_n \neq 0$, and because M is torsion-free, the map $x \mapsto ax$ is injective. It follows that M is isomorphic to a submodule of the free module $Au_1 \oplus \cdots \oplus Au_m$. By Proposition 35.5, this submodule if free, and thus, M is free. $\square$

Although we will obtain this result as a corollary of the structure theorem for finitely generated modules over a PID, we are in the position to give a quick proof of the following theorem.

Theorem 35.7. *Let M be a finitely generated module over a PID. Then M/M_{tor} is free, and there exit a free submodule F of M such that M is the direct sum*

$$M = M_{\mathrm{tor}} \oplus F.$$

The dimension of F is uniquely determined.

Proof. By Proposition 35.4 M/M_{tor} is torsion-free, and since M is finitely generated, it is also finitely generated. By Proposition 35.6, M/M_{tor} is free. We have the quotient linear map $\pi \colon M \to M/M_{\mathrm{tor}}$, which is surjective, and M/M_{tor} is free, so by Proposition 35.2, there is a free module F isomorphic to M/M_{tor} such that

$$M = \mathrm{Ker}\,(\pi) \oplus F = M_{\mathrm{tor}} \oplus F.$$

Since F is isomorphic to M/M_{tor}, the dimension of F is uniquely determined. $\square$

Theorem 35.7 reduces the study of finitely generated module over a PID to the study of finitely generated torsion modules. This is the path followed by Lang [109] (Chapter III, section 7).

35.2 Finite Presentations of Modules

Since modules are generally not free, it is natural to look for techniques for dealing with nonfree modules. The hint is that if M is an A-module and if $(u_i)_{i\in I}$ is any set of generators for M, then we know that there is a surjective homomorphism $\varphi \colon A^{(I)} \to M$ from the free module $A^{(I)}$ generated by I onto M. Furthermore M is isomorphic to $A^{(I)}/\mathrm{Ker}\,(\varphi)$. Then, we can pick a set of generators $(v_j)_{j\in J}$ for $\mathrm{Ker}\,(\varphi)$, and again there is a surjective map $\psi \colon A^{(J)} \to \mathrm{Ker}\,(\varphi)$ from the free module $A^{(J)}$ generated by J onto $\mathrm{Ker}\,(\varphi)$. The map ψ can be viewed a linear map from $A^{(J)}$ to $A^{(I)}$, we have

$$\mathrm{Im}(\psi) = \mathrm{Ker}\,(\varphi),$$

and φ is surjective. Note that M is isomorphic to $A^{(I)}/\mathrm{Im}(\psi)$. In such a situation we say that we have an *exact sequence* and this is denoted by the diagram

$$A^{(J)} \xrightarrow{\psi} A^{(I)} \xrightarrow{\varphi} M \longrightarrow 0.$$

Definition 35.6. Given an A-module M, a *presentation* of M is an exact sequence

$$A^{(J)} \xrightarrow{\psi} A^{(I)} \xrightarrow{\varphi} M \longrightarrow 0$$

which means that

1. $\mathrm{Im}(\psi) = \mathrm{Ker}\,(\varphi)$.
2. φ is surjective.

Consequently, M is isomorphic to $A^{(I)}/\mathrm{Im}(\psi)$. If I and J are both finite, we say that this is a *finite presentation* of M.

Observe that in the case of a finite presentation, I and J are finite, and if $|J| = n$ and $|I| = m$, then ψ is a linear map $\psi\colon A^n \to A^m$, so it is given by some $m \times n$ matrix R with coefficients in A called the *presentation matrix* of M. Every column R^j of R may thought of as a relation

$$a_{j1}e_1 + \cdots + a_{jm}e_m = 0$$

among the generators $e_1, \ldots, e_m$ of A^m, so we have n relations among these generators. Also the images of $e_1, \ldots, e_m$ in M are generators of M, so we can think of the above relations as relations among the generators of M.

The submodule of A^m spanned by the columns of R is *the set of relations* of M, and the columns of R are called a *complete set of relations* for M. The vectors $e_1, \ldots, e_m$ are called a set of *generators* for M. We may also say that the generators $e_1, \ldots, e_m$ and the relations $R^1, \ldots, R^n$ (the columns of R) are a (finite) presentation of the module M. The *module M presented by R is isomorphic to A^m/RA^n*, where we denote by RA^n the image of A^n by the linear map defined by R.

For example, the $\mathbb{Z}$-module presented by the 1×1 matrix $R = (5)$ is the quotient, $\mathbb{Z}/5\mathbb{Z}$, of $\mathbb{Z}$ by the submodule $5\mathbb{Z}$ corresponding to the single relation

$$5e_1 = 0.$$

But $\mathbb{Z}/5\mathbb{Z}$ has other presentations. For example, if we consider the matrix of relations

$$R = \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix},$$

presenting the module M, then we have the relations

$$\begin{aligned} 2e_1 + e_2 &= 0 \\ -e_1 + 2e_2 &= 0. \end{aligned}$$

From the first equation, we get $e_2 = -2e_1$, and substituting into the second equation we get

$$-5e_1 = 0.$$

It follows that the generator e_2 can be eliminated and M is generated by the single generator e_1 satisfying the relation

$$5e_1 = 0,$$

which shows that $M \approx \mathbb{Z}/5\mathbb{Z}$.

The above example shows that many different matrices can present the same module. Here are some useful rules for manipulating a relation matrix without changing the isomorphism class of the module M it presents.

Proposition 35.8. *If R is an $m \times n$ matrix presenting an A-module M, then the matrices S of the form listed below present the same module (a module isomorphic to M):*

(1) $S = QRP^{-1}$, where Q is a $m \times m$ invertible matrix and P a $n \times n$ invertible matrix (both over A).

(2) S is obtained from R by deleting a column of zeros.

(3) The jth column of R is e_i, and S is obtained from R by deleting the ith row and the jth column.

Proof. (1) By definition, we have an isomorphism $M \approx A^m/RA^n$, where we denote by RA^n the image of A^n by the linear map defined by R. Going from R to QRP^{-1} corresponds to making a change of basis in A^m and a change of basis in A^n, and this yields a quotient module isomorphic to M.

(2) A zero column does not contribute to the span of the columns of R, so it can be eliminated.

(3) If the jth column of R is e_i, then when taking the quotient A^m/RA^n, the generator e_i goes to zero. This means that the generator e_i is redundant, and when we delete it, we get a matrix of relations in which the ith row of R and the jth column of R are deleted. $\square$

The matrices P and Q are often products of elementary operations. One should be careful that rows of zeros cannnot be eliminated. For example, the 2×1 matrix

$$R_1 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

gives the single relation

$$4e_1 = 0,$$

but the second generator e_2 cannot be eliminated. This matrix presents the module $\mathbb{Z}/4\mathbb{Z}\times\mathbb{Z}$. On the other hand, the 1×2 matrix

$$R_2 = \begin{pmatrix} 4 & 0 \end{pmatrix}$$

gives two relations

$$\begin{aligned} 4e_1 &= 0, \\ 0 &= 0, \end{aligned}$$

so the second generator can be eliminated and R_2 presents the module $\mathbb{Z}/4\mathbb{Z}$.

The rules of Proposition 35.8 make it possible to simplify a presentation matrix quite a lot in some cases. For example, consider the relation matrix

$$R = \begin{pmatrix} 3 & 8 & 7 & 9 \\ 2 & 4 & 6 & 6 \\ 1 & 2 & 2 & 1 \end{pmatrix}.$$

By subtracting 2 times row 3 from row 2 and subtracting 3 times row 3 from row 1, we get

$$\begin{pmatrix} 0 & 2 & 1 & 6 \\ 0 & 0 & 2 & 4 \\ 1 & 2 & 2 & 1 \end{pmatrix}.$$

After deleting column 1 and row 3, we get

$$\begin{pmatrix} 2 & 1 & 6 \\ 0 & 2 & 4 \end{pmatrix}.$$

By subtracting 2 times row 1 from row 2, we get

$$\begin{pmatrix} 2 & 1 & 6 \\ -4 & 0 & -8 \end{pmatrix}.$$

After deleting column 2 and row 1, we get

$$\begin{pmatrix} -4 & -8 \end{pmatrix}.$$

By subtracting 2 times column 1 from column 2, we get

$$\begin{pmatrix} -4 & 0 \end{pmatrix}.$$

Finally, we can drop the second column and we get

$$(4),$$

which shows that R presents the module $\mathbb{Z}/4\mathbb{Z}$.

Unfortunately a submodule of a free module of finite dimension is not necessarily finitely generated but, by Proposition 35.5, if A is a PID, then any submodule of a finitely generated module is finitely generated. This property actually characterizes Noetherian rings. To prove it, we need a slightly different version of Proposition 35.2.

Proposition 35.9. *Let $f\colon E \to F$ be a linear map between two A-modules E and F.*

(1) Given any set of generators $(v_1, \ldots, v_r)$ of $\mathrm{Im}(f)$, for any r vectors $u_1, \ldots, u_r \in E$ such that $f(u_i) = v_i$ for $i = 1, \ldots, r$, if U is the finitely generated submodule of E generated by $(u_1, \ldots, u_r)$, then the module E is the sum

$$E = \mathrm{Ker}\,(f) + U.$$

Consequently, if both $\mathrm{Ker}\,(f)$ and $\mathrm{Im}(f)$ are finitely generated, then E is finitely generated.

(2) If E is finitely generated, then so is $\mathrm{Im}(f)$.

Proof. (1) Pick any $w \in E$, write $f(w)$ over the generators $(v_1, \ldots, v_r)$ of $\mathrm{Im}(f)$ as $f(w) = a_1 v_1 + \cdots + a_r v_r$, and let $u = a_1 u_1 + \cdots + a_r u_r$. Observe that

$$\begin{aligned} f(w-u) &= f(w) - f(u) \\ &= a_1 v_1 + \cdots + a_r v_r - (a_1 f(u_1) + \cdots + a_r f(u_r)) \\ &= a_1 v_1 + \cdots + a_r v_r - (a_1 v_1 + \cdots + a_r v_r) \\ &= 0. \end{aligned}$$

Therefore, $h = w - u \in \mathrm{Ker}\,(f)$, and since $w = h + u$ with $h \in \mathrm{Ker}\,(f)$ and $u \in U$, we have $E = \mathrm{Ker}\,(f) + U$, as claimed. If $\mathrm{Ker}\,(f)$ is also finitely generated, by taking the union of a finite set of generators for $\mathrm{Ker}\,(f)$ and $(v_1, \ldots, v_r)$, we obtain a finite set of generators for E.

(2) If $(u_1, \ldots, u_n)$ generate E, it is obvious that $(f(u_1), \ldots, f(u_n))$ generate $\mathrm{Im}(f)$. ☐

Theorem 35.10. *A ring A is Noetherian iff every submodule N of a finitely generated A-module M is itself finitely generated.*

Proof. First, assume that every submodule N of a finitely generated A-module M is itself finitely generated. The ring A is a module over itself and it is generated by the single element 1. Furthermore, every submodule of A is an ideal, so the hypothesis implies that every ideal in A is finitely generated, which shows that A is Noetherian.

Now, assume A is Noetherian. First, observe that it is enough to prove the theorem for the finitely generated free modules A^n (with $n \geq 1$). Indeed, assume that we proved for every $n \geq 1$ that every submodule of A^n is finitely generated. If M is any finitely generated A-module, then there is a surjection $\varphi\colon A^n \to M$ for some n (where n is the number of elements of a finite generating set for M). Given any submodule N of M, $L = \varphi^{-1}(N)$ is a submodule of A^n. Since A^n is finitely generated, the submodule N of A^n is finitely generated, and then $N = \varphi(L)$ is finitely generated.

It remains to prove the theorem for $M = A^n$. We proceed by induction on n. For $n = 1$, a submodule N of A is an ideal, and since A is Noetherian, N is finitely generated. For the induction step where $n > 1$, consider the projection $\pi\colon A^n \to A^{n-1}$ given by

$$\pi(a_1, \ldots, a_n) = (a_1, \ldots, a_{n-1}).$$

The kernel of π is the module

$$\mathrm{Ker}\,(\pi) = \{(0, \ldots, 0, a_n) \in A^n \mid a_n \in A\} \approx A.$$

For any submodule N of A^n, let $\varphi\colon N \to A^{n-1}$ be the restriction of π to N. Since $\varphi(N)$ is a submodule of A^{n-1}, by the induction hypothesis, $\mathrm{Im}(\varphi) = \varphi(N)$ is finitely generated. Also, $\mathrm{Ker}\,(\varphi) = N \cap \mathrm{Ker}\,(\pi)$ is a submodule of $\mathrm{Ker}\,(\pi) \approx A$, and thus $\mathrm{Ker}\,(\varphi)$ is isomorphic to an ideal of A, and thus is finitely generated (since A is Noetherian). Since both $\mathrm{Im}(\varphi)$ and $\mathrm{Ker}\,(\varphi)$ are finitely generated, by Proposition 35.9, the submodule N is also finitely generated. $\square$

As a consequence of Theorem 35.10, every finitely generated A-module over a Noetherian ring A is finitely presented, because if $\varphi\colon A^n \to M$ is a surjection onto the finitely generated module M, then $\mathrm{Ker}\,(\varphi)$ is finitely generated. In particular, if A is a PID, then every finitely generated module is finitely presented.

If the ring A is not Noetherian, then there exist finitely generated A-modules that are not finitely presented. This is not so easy to prove.

We will prove in Proposition 35.35 that if A is a PID then a matrix R can "diagonalized" as

$$R = QDP^{-1}$$

where D is a diagonal matrix (more computational versions of this proposition are given in Theorem 36.18 and Theorem 36.21). It follows from Proposition 35.8 that every finitely generated module M over a PID has a presentation with m generators and r relations of the form

$$\alpha_i e_i = 0,$$

where $\alpha_i \neq 0$ and $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$, which shows that M is isomorphic to the direct sum

$$M \approx A^{m-r} \oplus A/(\alpha_1 A) \oplus \cdots \oplus A/(\alpha_r A).$$

This is a version of Theorem 35.25 that will be proved in Section 35.5.

35.3 Tensor Products of Modules over a Commutative Ring

It is possible to define tensor products of modules over a ring, just as in Section 33.2, and the results of this section continue to hold. The results of Section 33.4 also continue to hold since they are based on the universal mapping property. However, the results of Section 33.3 on bases generally fail, except for free modules. Similarly, the results of Section 33.5 on duality generally fail. Tensor algebras can be defined for modules, as in Section 33.6. Symmetric tensor and alternating tensors can be defined for modules but again, results involving bases generally fail.

Tensor products of modules have some unexpected properties. For example, if p and q are relatively prime integers, then

$$\mathbb{Z}/p\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/q\mathbb{Z} = (0).$$

This is because, by Bezout's identity, there are $a, b \in \mathbb{Z}$ such that

$$ap + bq = 1,$$

so, for all $x \in \mathbb{Z}/p\mathbb{Z}$ and all $y \in \mathbb{Z}/q\mathbb{Z}$, we have

$$\begin{aligned} x \otimes y &= ap(x \otimes y) + bq(x \otimes y) \\ &= a(px \otimes y) + b(x \otimes qy) \\ &= a(0 \otimes y) + b(x \otimes 0) \\ &= 0. \end{aligned}$$

It is possible to salvage certain properties of tensor products holding for vector spaces by restricting the class of modules under consideration. For example, *projective modules* have a pretty good behavior w.r.t. tensor products.

A free A-module F, is a module that has a basis (*i.e.*, there is a family, $(e_i)_{i \in I}$, of linearly independent vectors in F that span F). Projective modules have many equivalent characterizations. Here is one that is best suited for our needs:

Definition 35.7. An A-module, P, is *projective* if it is a summand of a free module, that is, if there is a free A-module, F, and some A-module, Q, so that

$$F = P \oplus Q.$$

Given any A-module, M, we let $M^* = \mathrm{Hom}_A(M, A)$ be its *dual*. We have the following proposition:

Proposition 35.11. *For any finitely-generated projective A-modules, P, and any A-module, Q, we have the isomorphisms:*

$$\begin{aligned} P^{**} &\cong P \\ \mathrm{Hom}_A(P, Q) &\cong P^* \otimes_A Q. \end{aligned}$$

Proof sketch. We only consider the second isomorphism. Since P is projective, we have some A-modules, P_1, F, with

$$P \oplus P_1 = F,$$

where F is some free module. Now, we know that for any A-modules, U, V, W, we have

$$\mathrm{Hom}_A(U \oplus V, W) \cong \mathrm{Hom}_A(U, W) \prod \mathrm{Hom}_A(V, W) \cong \mathrm{Hom}_A(U, W) \oplus \mathrm{Hom}_A(V, W),$$

so

$$P^* \oplus P_1^* \cong F^*, \qquad \mathrm{Hom}_A(P, Q) \oplus \mathrm{Hom}_A(P_1, Q) \cong \mathrm{Hom}_A(F, Q).$$

By tensoring with Q and using the fact that tensor distributes w.r.t. coproducts, we get

$$(P^* \otimes_A Q) \oplus (P_1^* \otimes_A Q) \cong (P^* \oplus P_1^*) \otimes_A Q \cong F^* \otimes_A Q.$$

Now, the proof of Proposition 33.17 goes through because F is free and finitely generated, so

$$\alpha_\otimes \colon (P^* \otimes_A Q) \oplus (P_1^* \otimes_A Q) \cong F^* \otimes_A Q \longrightarrow \mathrm{Hom}_A(F, Q) \cong \mathrm{Hom}_A(P, Q) \oplus \mathrm{Hom}_A(P_1, Q)$$

is an isomorphism and as $\alpha_\otimes$ maps $P^* \otimes_A Q$ to $\mathrm{Hom}_A(P, Q)$, it yields an isomorphism between these two spaces. $\square$

The isomorphism $\alpha_\otimes \colon P^* \otimes_A Q \cong \mathrm{Hom}_A(P, Q)$ of Proposition 35.11 is still given by

$$\alpha_\otimes(u^* \otimes f)(x) = u^*(x) f, \qquad u^* \in P^*,\ f \in Q,\ x \in P.$$

It is convenient to introduce the *evaluation map*, $\mathrm{Ev}_x \colon P^* \otimes_A Q \to Q$, defined for every $x \in P$ by

$$\mathrm{Ev}_x(u^* \otimes f) = u^*(x) f, \qquad u^* \in P^*,\ f \in Q.$$

We will need the following generalization of part (4) of Proposition 33.13.

Proposition 35.12. *Given any two families of A-modules $(M_i)_{i \in I}$ and $(N_j)_{j \in J}$ (where I and J are finite index sets), we have an isomorphism*

$$\Big(\bigoplus_{i \in I} M_i\Big) \otimes \Big(\bigoplus_{j \in I} M_j\Big) \approx \bigoplus_{(i,j) \in I \times J} (M_i \otimes N_j).$$

Proposition 35.12 also holds for infinite index sets.

Proposition 35.13. *Let M and N be two A-module with N a free module, and pick any basis $(v_1, \ldots, v_n)$ for N. Then, every element of $M \otimes N$ can expressed in a unique way as a sum of the form*

$$u_1 \otimes v_1 + \cdots + u_n \otimes v_n, \quad u_i \in M,$$

so that $M \otimes N$ is isomorphic to M^n (as an A-module).

Proof. Since N is free with basis $(v_1, \ldots, v_n)$, we have an isomorphism

$$N \approx Av_1 \oplus \cdots \oplus Av_n.$$

By Proposition 35.12, we obtain an isomorphism

$$M \otimes N \approx M \otimes (Av_1 \oplus \cdots \oplus Av_n) \approx (M \otimes Av_1) \oplus \cdots \oplus (M \otimes Av_n).$$

Because $(v_1, \ldots, v_n)$ is a basis of N, each v_j is torsion-free so the map $a \mapsto av_j$ is an isomorphism of A onto Av_j, and because $M \otimes A \approx M$, we have the isomorphism

$$M \otimes N \approx (M \otimes A) \oplus \cdots \oplus (M \otimes A) \approx M \oplus \cdots \oplus M = M^n,$$

as claimed. $\square$

Proposition 35.13 also holds for an infinite basis $(v_j)_{j \in J}$ of N. Obviously, a version of Proposition 35.13 also holds if M is free and N is arbitrary.

The next proposition will be also be needed.

Proposition 35.14. *Given any A-module M and any ideal $\mathfrak{a}$ in A, there is an isomorphism*

$$(A/\mathfrak{a}) \otimes_A M \approx M/\mathfrak{a}M$$

given by the map $(\overline{a} \otimes u) \mapsto au \pmod{\mathfrak{a}M}$, for all $\overline{a} \in A/\mathfrak{a}$ and all $u \in M$.

Sketch of proof. Consider the map $\varphi\colon (A/\mathfrak{a}) \times M \to M/\mathfrak{a}M$ given by

$$\varphi(\overline{a}, u) = au \pmod{\mathfrak{a}M}$$

for all $\overline{a} \in A/\mathfrak{a}$ and all $u \in M$. It is immediately checked that φ is well-defined because $au \pmod{\mathfrak{a}M}$ does not depend on the representative $a \in A$ chosen in the equivalence class $\overline{a}$, and φ is bilinear. Therefore, φ induces a linear map $\varphi\colon (A/\mathfrak{a}) \otimes M \to M/\mathfrak{a}M$, such that $\varphi(\overline{a} \otimes u) = au \pmod{\mathfrak{a}M}$. We also define the map $\psi\colon M \to (A/\mathfrak{a}) \otimes M$ by

$$\psi(u) = \overline{1} \otimes u.$$

Since $\mathfrak{a}M$ is generated by vectors of the form au with $a \in \mathfrak{a}$ and $u \in M$, and since

$$\psi(au) = \overline{1} \otimes au = \overline{a} \otimes u = 0 \otimes u = 0,$$

we see that $\mathfrak{a}M \subseteq \mathrm{Ker}\,(\psi)$, so ψ induces a linear map $\psi\colon M/\mathfrak{a}M \to (A/\mathfrak{a}) \otimes M$. We have

$$\begin{aligned}\psi(\varphi(\overline{a} \otimes u)) &= \psi(au) \\ &= \overline{1} \otimes au \\ &= \overline{a} \otimes u\end{aligned}$$

and

$$\begin{aligned}\varphi(\psi(u)) &= \varphi(\overline{1} \otimes u) \\ &= 1u \\ &= u,\end{aligned}$$

which shows that φ and ψ are mutual inverses. $\square$

We now develop the theory necessary to understand the structure of finitely generated modules over a PID.

35.4 Torsion Modules over a PID; The Primary Decomposition

We begin by considering modules over a product ring obtained from a direct decomposition, as in Definition 32.3. In this section and the next, we closely follow Bourbaki [26] (Chapter VII). Let A be a commutative ring and let $(\mathfrak{b}_1, \ldots, \mathfrak{b}_n)$ be ideals in A such that there is an isomorphism $A \approx A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$. From Theorem 32.16 part (b), there exist some elements $e_1, \ldots, e_n$ of A such that

$$\begin{aligned}e_i^2 &= e_i \\ e_i e_j &= 0, \quad i \neq j \\ e_1 + \cdots + e_n &= 1_A,\end{aligned}$$

and $\mathfrak{b}_i = (1_A - e_i)A$, for $i, j = 1, \ldots, n$.

Given an A-module M with $A \approx A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$, let M_i be the subset of M annihilated by $\mathfrak{b}_i$; that is,

$$M_i = \{x \in M \mid bx = 0, \text{ for all } b \in \mathfrak{b}_i\}.$$

Because $\mathfrak{b}_i$ is an ideal, each M_i is a submodule of M. Observe that if $\lambda, \mu \in A$, $b \in \mathfrak{b}_i$, and if $\lambda - \mu = b$, then for any $x \in M_i$, since $bx = 0$,

$$\lambda x = (\mu + b)x = \mu x + bx = \mu x,$$

so M_i can be viewed as a $A/\mathfrak{b}_i$- module.

Proposition 35.15. *Given a ring $A \approx A/\mathfrak{b}_1 \times \cdots \times A/\mathfrak{b}_n$ as above, the A-module M is the direct sum*

$$M = M_1 \oplus \cdots \oplus M_n,$$

where M_i is the submodule of M annihilated by $\mathfrak{b}_i$.

Proof. For $i = 1, \ldots, n$, let $p_i\colon M \to M$ be the map given by

$$p_i(x) = e_i x, \quad x \in M.$$

The map p_i is clearly linear, and because of the properties satisfied by the e_is, we have

$$\begin{aligned} p_i^2 &= p_i \\ p_i p_j &= 0, \quad i \neq j \\ p_1 + \cdots + p_n &= \mathrm{id}. \end{aligned}$$

This shows that the p_i are projections, and by Proposition 6.8 (which also holds for modules), we have a direct sum

$$M = p_1(M) \oplus \cdots \oplus p_n(M) = e_1 M \oplus \cdots \oplus e_n M.$$

It remains to show that $M_i = e_i M$. Since $(1 - e_i)e_i = e_i - e_i^2 = e_i - e_i = 0$, we see that $e_i M$ is annihilated by $\mathfrak{b}_i = (1 - e_i)A$. Furthermore, for $i \neq j$, for any $x \in M$, we have $(1 - e_i)e_j x = (e_j - e_i e_j)x = e_j x$, so no nonzero element of $e_j M$ is annihilated by $1 - e_i$, and thus not annihilated by $\mathfrak{b}_i$. It follows that $e_i M = M_i$, as claimed. $\square$

Definition 35.8. Given an A-module M, for any nonzero $\alpha \in A$, let

$$M(\alpha) = \{x \in M \mid \alpha x = 0\},$$

the submodule of M annihilated by α. If α divides β, then $M(\alpha) \subseteq M(\beta)$, so we can define

$$M_\alpha = \bigcup_{n \geq 1} M(\alpha^n) = \{x \in M \mid (\exists n \geq 1)(\alpha^n x = 0)\},$$

the submodule of M consisting of all elements of M annihilated by some power of α.

If N is any submodule of M, it is clear that

$$N_\alpha = M \cap M_\alpha.$$

Recall that in a PID, an irreducible element is also called a *prime element*.

Definition 35.9. If A is a PID and p is a prime element in A, we say that a module M is *p-primary* if $M = M_p$.

Proposition 35.16. *Let M be module over a PID A. For every nonzero $\alpha \in A$, if*

$$\alpha = u p_1^{n_1} \cdots p_r^{n_r}$$

is a factorization of α into prime factors (where u is a unit), then the module $M(\alpha)$ annihilated by α is the direct sum

$$M(\alpha) = M(p_1^{n_1}) \oplus \cdots \oplus M(p_r^{n_r}).$$

Furthermore, the projection from $M(\alpha)$ onto $M(p_i^{n_i})$ is of the form $x \mapsto \gamma_i x$, for some $\gamma_i \in A$, and

$$M(p_i^{n_i}) = M(\alpha) \cap M_{p_i}.$$

Proof. First observe that since $M(\alpha)$ is annihilated by α, we can view $M(\alpha)$ as a $A/(\alpha)$-module. By the Chinese remainder theorem (Theorem 32.15) applied to the ideals $(up_1^{n_1}) = (p_1^{n_1}), (p_2^{n_2}), \ldots, (p_r^{n_r})$, we have an isomorphism

$$A/(\alpha) \approx A/(p_1^{n_1}) \times \cdots \times A/(p_r^{n_r}).$$

Since we also have isomorphisms

$$A/(p_i^{n_i}) \approx (A/(\alpha))/((p_i^{n_i})/(\alpha)),$$

we can apply Proposition 35.15, and we get a direct sum

$$M(\alpha) = N_1 \oplus \cdots \oplus N_r,$$

where N_i is the $A/(\alpha)$-submodule of $M(\alpha)$ annihilated by $(p_i^{n_i})/(\alpha)$, and the projections onto the N_i are of the form stated in the proposition. However, N_i is just the A-module $M(p_i^{n_i})$ annihilated by $p_i^{n_i}$, because every nonzero element of $(p_i^{n_i})/(\alpha)$ is an equivalence class modulo (α) of the form $\overline{ap_i^{n_i}}$ for some nonzero $a \in A$, and by definition, $x \in N_i$ iff

$$0 = \overline{ap_i^{n_i}}\, x = ap_i^{n_i} x, \quad \text{for all } a \in A - \{0\},$$

in particular for $a = 1$, which implies that $x \in M(p_i^{n_i})$.

The inclusion $M(p_i^{n_i}) \subseteq M(\alpha) \cap M_{p_i}$ is clear. Conversely, pick $x \in M(\alpha) \cap M_{p_i}$, which means that $\alpha x = 0$ and $p_i^s x = 0$ for some $s \geq 1$. If $s < n_i$, we are done, so assume $s \geq n_i$. Since $p_i^{n_i}$ is a gcd of α and p_i^s, by Bezout, we can write

$$p_i^{n_i} = \lambda p_i^s + \mu\alpha$$

for some $\lambda, \mu \in A$, and then $p_i^{n_i} x = \lambda p_i^s x + \mu\alpha x = 0$, which shows that $x \in M(p_i^{n_i})$, as desired. $\square$

Here is an example of Proposition 35.16. Let $M = \mathbb{Z}/60\mathbb{Z}$, where M is considered as a $\mathbb{Z}$-module. A element in M is denoted by $\overline{x}$, where x is an integer with $0 \leq x \leq 59$. Let $\alpha = 6$ and define

$$M(6) = \{\overline{x} \in M \mid 6\overline{x} = \overline{0}\} = \{\overline{0}, \overline{10}, \overline{20}, \overline{30}, \overline{40}, \overline{50}\}.$$

Since $6 = 2 \cdot 3$, Proposition 35.16 implies that $M(6) = M(2) \oplus M(3)$, where

$$\begin{aligned} M(2) &= \{\overline{x} \in M \mid 2\overline{x} = \overline{0}\} = \{\overline{0}, \overline{30}\} \\ M(3) &= \{\overline{x} \in M \mid 3\overline{x} = \overline{0}\} = \{\overline{0}, \overline{20}, \overline{40}\}. \end{aligned}$$

Recall that if M is a torsion module over a ring A which is an integral domain, then every finite set of elements $x_1, \ldots, x_n$ in M is annihilated by $a = a_1 \cdots a_n$, where each a_i annihilates x_i.

Since A is a PID, we can pick a set P of irreducible elements of A such that every nonzero nonunit of A has a unique factorization up to a unit. Then, we have the following structure theorem for torsion modules which holds even for modules that are not finitely generated.

Theorem 35.17. *(Primary Decomposition Theorem) Let M be a torsion-module over a PID. For every irreducible element $p \in P$, let M_p be the submodule of M annihilated by some power of p. Then, M is the (possibly infinite) direct sum*

$$M = \bigoplus_{p \in P} M_p.$$

Proof. Since M is a torsion-module, for every $x \in M$, there is some $\alpha \in A$ such that $x \in M(\alpha)$. By Proposition 35.16, if $\alpha = up_1^{n_1} \cdots p_r^{n_r}$ is a factorization of α into prime factors (where u is a unit), then the module $M(\alpha)$ is the direct sum

$$M(\alpha) = M(p_1^{n_1}) \oplus \cdots \oplus M(p_r^{n_r}).$$

This means that x can be written as

$$x = \sum_{p \in P} x_p, \quad x_p \in M_p,$$

with only finitely many x_p nonzero. If

$$\sum_{p \in P} x_p = \sum_{p \in P} y_p$$

for all $p \in P$, with only finitely many x_p and y_p nonzero, then x_p and y_p are annihilated by some common nonzero element $a \in A$, so $x_p, y_p \in M(a)$. By Proposition 35.16, we must have $x_p = y_p$ for all p, which proves that we have a direct sum. $\square$

It is clear that if p and p' are two irreducible elements such that $p = up'$ for some unit u, then $M_p = M_{p'}$. Therefore, M_p only depends on the ideal (p).

Definition 35.10. Given a torsion-module M over a PID, the modules M_p associated with irreducible elements in P are called the *p-primary components* of M.

The p-primary components of a torsion module uniquely determine the module, as shown by the next proposition.

Proposition 35.18. *Two torsion modules M and N over a PID are isomorphic iff for every every irreducible element $p \in P$, the p-primary components M_p and N_p of M and N are isomorphic.*

Proof. Let $f \colon M \to N$ be an isomorphism. For any $p \in P$, we have $x \in M_p$ iff $p^k x = 0$ for some $k \geq 1$, so

$$0 = f(p^k x) = p^k f(x),$$

which shows that $f(x) \in N_p$. Therefore, f restricts to a linear map $f \mid M_p$ from M_p to N_p. Since f is an isomorphism, we also have a linear map $f^{-1} \colon M \to N$, and our previous reasoning shows that f^{-1} restricts to a linear map $f^{-1} \mid N_p$ from N_p to M_p. But, $f \mid M_p$ and $f^{-1} \mid N_p$ are mutual inverses, so M_p and N_p are isomorphic.

Conversely, if $M_p \approx N_p$ for all $p \in P$, by Theorem 35.17, we get an isomorphism between $M = \bigoplus_{p \in P} M_p$ and $N = \bigoplus_{p \in P} N_p$. $\square$

In view of Proposition 35.18, the direct sum of Theorem 35.17 in terms of its p-primary components is called the *canonical primary decomposition* of M.

If M is a finitely generated torsion-module, then Theorem 35.17 takes the following form.

Theorem 35.19. *(Primary Decomposition Theorem for finitely generated torsion modules) Let M be a finitely generated torsion-module over a PID A. If $\mathrm{Ann}(M) = (a)$ and if $a = up_1^{n_1} \cdots p_r^{n_r}$ is a factorization of a into prime factors, then M is the finite direct sum*

$$M = \bigoplus_{i=1}^{r} M(p_i^{n_i}).$$

Furthermore, the projection of M over $M(p_i^{n_i})$ is of the form $x \mapsto \gamma_i x$, for some $\gamma_i \in A$.

Proof. This is an immediate consequence of Proposition 35.16. $\square$

Theorem 35.19 applies when $A = \mathbb{Z}$. In this case, M is a finitely generated torsion abelian group, and the theorem says that such a group is the direct sum of a finite number of groups whose elements have order some power of a prime number p. In particular, consider the $\mathbb{Z}$-module $\mathbb{Z}/10\mathbb{Z}$ where

$$\mathbb{Z}/10\mathbb{Z} = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}, \overline{6}, \overline{7}, \overline{8}, \overline{9}\}.$$

Clearly $\mathbb{Z}/10\mathbb{Z}$ is generated by $\overline{1}$ and $\mathrm{Ann}\,(\mathbb{Z}/10\mathbb{Z}) = 10$. Theorem 35.19 implies that

$$\mathbb{Z}/10\mathbb{Z} = M(2) \oplus M(5),$$

where

$$\begin{aligned} M(2) &= \{\overline{x} \in M \mid 2\overline{x} = \overline{0}\} = \{\overline{0}, \overline{5}\} \\ M(5) &= \{\overline{x} \in M \mid 5\overline{x} = \overline{0}\} = \{\overline{0}, \overline{2}, \overline{4}, \overline{6}, \overline{8}\}. \end{aligned}$$

Theorem 35.17 has several useful corollaries.

Proposition 35.20. *If M is a torsion module over a PID, for every submodule N of M, we have a direct sum*

$$N = \bigoplus_{p \in P} N \cap M_p.$$

Proof. It is easily verified that $N \cap M_p$ is the p-primary component of N. $\square$

Proposition 35.21. *If M is a torsion module over a PID, a submodule N of M is a direct factor of M iff N_p is a direct factor of M_p for every irreducible element $p \in A$.*

Proof. This is because if N and N' are two submodules of M, we have $M = N \oplus N'$ iff, by Proposition 35.20, $M_p = N_p \oplus N'_p$ for every irreducible elements $p \in A$. $\square$

Definition 35.11. An A-module M is said to be *semi-simple* iff for every submodule N of M, there is some submodule N' of M such that $M = N \oplus N'$.

Proposition 35.22. *Let A be a PID which is not a field, and let M be any A-module. Then M is semi-simple iff it is a torsion module and if $M_p = M(p)$ for every irreducible element $p \in A$ (in other words, if $x \in M$ is annihilated by a power of p, then it is already annihilated by p).*

Proof. Assume that M is semi-simple. Let $x \in M$ and pick any irreducible element $p \in A$. Then, the submodule pAx has a supplement N such that

$$M = pAx \oplus N,$$

so we can write $x = pax + y$, for some $y \in N$ and some $a \in A$. But then,

$$y = (1 - pa)x,$$

and since p is irreducible, p is not a unit, so $1 - pa \neq 0$. Observe that

$$p(1 - ap)x = py \in pAx \cap N = (0).$$

Since $p(1 - ap) \neq 0$, x is a torsion element, and thus M is a torsion module. The above argument shows that

$$p(1 - ap)x = 0,$$

which implies that $px = ap^2x$, and by induction,

$$px = a^n p^{n+1} x, \quad \text{for all } n \geq 1.$$

If we pick x in M_p, then there is some $m \geq 1$ such that $p^m x = 0$, and we conclude that

$$px = 0.$$

Therefore, $M_p = M(p)$, as claimed.

Conversely, assume that M is a torsion-module and that $M_p = M(p)$ for every irreducible element $p \in A$. By Proposition 35.21, it is sufficient to prove that a module annihilated by a an irreducible element is semi-simple. This is because such a module is a vector space over the field $A/(p)$ (recall that in a PID, an ideal (p) is maximal iff p is irreducible), and in a vector space, every subspace has a supplement. $\square$

Theorem 35.19 shows that a finitely generated torsion module is a direct sum of p-primary modules M_p. We can do better. In the next section we show that each primary module M_p is the direct sum of cyclic modules of the form $A/(p^n)$.

35.5 Finitely Generated Modules over a PID; Invariant Factor Decomposition

There are several ways of obtaining the decomposition of a finitely generated module as a direct sum of cyclic modules. One way to proceed is to first use the Primary Decomposition Theorem and then to show how each primary module M_p is the direct sum of cyclic modules of the form $A/(p^n)$. This is the approach followed by Lang [109] (Chapter III, section 7), among others. We prefer to use a proposition that produces a particular basis for a submodule of a finitely generated free module, because it yields more information. This is the approach followed in Dummitt and Foote [54] (Chapter 12) and Bourbaki [26] (Chapter VII). The proof that we present is due to Pierre Samuel.

Proposition 35.23. *Let F be a finitely generated free module over a PID A, and let M be any submodule of F. Then, M is a free module and there is a basis $(e_1, ..., e_n)$ of F, some $q \leq n$, and some nonzero elements $a_1, \ldots, a_q \in A$, such that $(a_1e_1, \ldots, a_qe_q)$ is a basis of M and a_i divides a_{i+1} for all i, with $1 \leq i \leq q-1$.*

Proof. The proposition is trivial when $M = \{0\}$, thus assume that M is nontrivial. Pick some basis $(u_1, \ldots, u_n)$ for F. Let $L(F, A)$ be the set of linear forms on F. For any $f \in L(F, A)$, it is immediately verified that $f(M)$ is an ideal in A. Thus, $f(M) = a_hA$, for some $a_h \in A$, since every ideal in A is a principal ideal. Since A is a PID, any nonempty family of ideals in A has a maximal element, so let f be a linear map such that a_hA is a maximal ideal in A. Let $\pi_i \colon F \to A$ be the i-th projection, i.e., π_i is defined such that $\pi_i(x_1u_1 + \cdots + x_nu_n) = x_i$. It is clear that π_i is a linear map, and since M is nontrivial, one of the $\pi_i(M)$ is nontrivial, and $a_h \neq 0$. There is some $e' \in M$ such that $f(e') = a_h$.

We claim that, for every $g \in L(F, A)$, the element $a_h \in A$ divides $g(e')$.

Indeed, if d is the gcd of a_h and $g(e')$, by the Bézout identity, we can write

$$d = ra_h + sg(e'),$$

for some $r, s \in A$, and thus

$$d = rf(e') + sg(e') = (rf + sg)(e').$$

However, $rf + sg \in L(F, A)$, and thus,

$$a_hA \subseteq dA \subseteq (rf + sg)(M),$$

since d divides a_h, and by maximality of a_hA, we must have $a_hA = dA$, which implies that $d = a_h$, and thus, a_h divides $g(e')$. In particular, a_h divides each $\pi_i(e')$ and let $\pi_i(e') = a_hb_i$, with $b_i \in A$.

Let $e = b_1u_1 + \cdots + b_nu_n$. Note that

$$e' = \pi_1(e')u_1 + \cdots + \pi_n(e')u_n = a_hb_1u_1 + \cdots + a_hb_nu_n,$$

and thus, $e' = a_h e$. Since $a_h = f(e') = f(a_h e) = a_h f(e)$, and since $a_h \neq 0$, we must have $f(e) = 1$.

Next, we claim that
$$F = Ae \oplus f^{-1}(0)$$
and
$$M = Ae' \oplus (M \cap f^{-1}(0)),$$
with $e' = a_h e$.

Indeed, every $x \in F$ can be written as
$$x = f(x)e + (x - f(x)e),$$
and since $f(e) = 1$, we have $f(x - f(x)e) = f(x) - f(x)f(e) = f(x) - f(x) = 0$. Thus, $F = Ae + f^{-1}(0)$. Similarly, for any $x \in M$, we have $f(x) = ra_h$, for some $r \in A$, and thus,
$$x = f(x)e + (x - f(x)e) = ra_h e + (x - f(x)e) = re' + (x - f(x)e),$$
we still have $x - f(x)e \in f^{-1}(0)$, and clearly, $x - f(x)e = x - ra_h e = x - re' \in M$, since $e' \in M$. Thus, $M = Ae' + (M \cap f^{-1}(0))$.

To prove that we have a direct sum, it is enough to prove that $Ae \cap f^{-1}(0) = \{0\}$. For any $x = re \in Ae$, if $f(x) = 0$, then $f(re) = rf(e) = r = 0$, since $f(e) = 1$ and, thus, $x = 0$. Therefore, the sums are direct sums.

We can now prove that M is a free module by induction on the size, q, of a maximal linearly independent family for M.

If $q = 0$, the result is trivial. Otherwise, since
$$M = Ae' \oplus (M \cap f^{-1}(0)),$$
it is clear that $M \cap f^{-1}(0)$ is a submodule of F and that every maximal linearly independent family in $M \cap f^{-1}(0)$ has at most $q - 1$ elements. By the induction hypothesis, $M \cap f^{-1}(0)$ is a free module, and by adding e' to a basis of $M \cap f^{-1}(0)$, we obtain a basis for M, since the sum is direct.

The second part is shown by induction on the dimension n of F.

The case $n = 0$ is trivial. Otherwise, since
$$F = Ae \oplus f^{-1}(0),$$
and since, by the previous argument, $f^{-1}(0)$ is also free, $f^{-1}(0)$ has dimension $n - 1$. By the induction hypothesis applied to its submodule $M \cap f^{-1}(0)$, there is a basis $(e_2, \ldots, e_n)$ of $f^{-1}(0)$, some $q \leq n$, and some nonzero elements $a_2, \ldots, a_q \in A$, such that, $(a_2e_2, \ldots, a_qe_q)$ is a basis of $M \cap f^{-1}(0)$, and a_i divides a_{i+1} for all i, with $2 \leq i \leq q - 1$. Let $e_1 = e$, and $a_1 = a_h$, as above. It is clear that $(e_1, \ldots, e_n)$ is a basis of F, and that that $(a_1e_1, \ldots, a_qe_q)$

is a basis of M, since the sums are direct, and $e' = a_1e_1 = a_he$. It remains to show that a_1 divides a_2. Consider the linear map $g\colon F \to A$ such that $g(e_1) = g(e_2) = 1$, and $g(e_i) = 0$, for all i, with $3 \leq i \leq n$. We have $a_h = a_1 = g(a_1e_1) = g(e') \in g(M)$, and thus $a_hA \subseteq g(M)$. Since a_hA is maximal, we must have $g(M) = a_hA = a_1A$. Since $a_2 = g(a_2e_2) \in g(M)$, we have $a_2 \in a_1A$, which shows that a_1 divides a_2. □

We need the following basic proposition.

Proposition 35.24. *For any commutative ring A, if F is a free A-module and if $(e_1, \ldots, e_n)$ is a basis of F, for any elements $a_1, \ldots, a_n \in A$, there is an isomorphism*

$$F/(Aa_1e_1 \oplus \cdots \oplus Aa_ne_n) \approx (A/a_1A) \oplus \cdots \oplus (A/a_nA).$$

Proof. Let $\sigma\colon F \to A/(a_1A) \oplus \cdots \oplus A/(a_nA)$ be the linear map given by

$$\sigma(x_1e_1 + \cdots + x_ne_n) = (\overline{x}_1, \ldots, \overline{x}_n),$$

where $\overline{x}_i$ is the equivalence class of x_i in A/a_iA. The map σ is clearly surjective, and its kernel consists of all vectors $x_1e_1 + \cdots + x_ne_n$ such that $x_i \in a_iA$, for $i = 1, \ldots, n$, which means that

$$\mathrm{Ker}\,(\sigma) = Aa_1e_1 \oplus \cdots \oplus Aa_ne_n.$$

Since $M/\mathrm{Ker}\,(\sigma)$ is isomorphic to $\mathrm{Im}(\sigma)$, we get the desired isomorphism. □

We can now prove the existence part of the structure theorem for finitely generated modules over a PID.

Theorem 35.25. *Let M be a finitely generated nontrivial A-module, where A a PID. Then, M is isomorphic to a direct sum of cyclic modules*

$$M \approx A/\mathfrak{a}_1 \oplus \cdots \oplus A/\mathfrak{a}_m,$$

where the $\mathfrak{a}_i$ are proper ideals of A (possibly zero) such that

$$\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_m \neq A.$$

More precisely, if $\mathfrak{a}_1 = \cdots = \mathfrak{a}_r = (0)$ and $(0) \neq \mathfrak{a}_{r+1} \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$, then

$$M \approx A^r \oplus (A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m),$$

where $A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m$ is the torsion submodule of M. The module M is free iff $r = m$, and a torsion-module iff $r = 0$. In the latter case, the annihilator of M is $\mathfrak{a}_1$.

Proof. Since M is finitely generated and nontrivial, there is a surjective homomorphism $\varphi\colon A^n \to M$ for some $n \geq 1$, and M is isomorphic to $A^n/\mathrm{Ker}\,(\varphi)$. Since $\mathrm{Ker}\,(\varphi)$ is a submodule of the free module A^n, by Proposition 35.23, $\mathrm{Ker}\,(\varphi)$ is a free module and there is a basis $(e_1, \ldots, e_n)$ of A^n and some nonzero elements $a_1, \ldots, a_q$ $(q \leq n)$ such that $(a_1e_1, \ldots, a_qe_q)$ is a basis of $\mathrm{Ker}\,(\varphi)$ and $a_1 \mid a_2 \mid \cdots \mid a_q$. Let $a_{q+1} = \ldots = a_n = 0$.

By Proposition 35.24, we have an isomorphism

$$A^n/\mathrm{Ker}\,(\varphi) \approx A/a_1A \oplus \cdots \oplus A/a_nA.$$

Whenever a_i is unit, the factor $A/a_iA = (0)$, so we can weed out the units. Let $r = n - q$, and let $s \in \mathbb{N}$ be the smallest index such that a_{s+1} is not a unit. Note that $s = 0$ means that there are no units. Also, as $M \neq (0)$, $s < n$. Then,

$$M \approx A^n/\mathrm{Ker}\,(\varphi) \approx A/a_{s+1}A \oplus \cdots \oplus A/a_nA.$$

Let $m = r + q - s = n - s$. Then, we have the sequence

$$\underbrace{a_{s+1}, \ldots, a_q}_{q-s}, \underbrace{a_{q+1}, \ldots, a_n}_{r=n-q},$$

where $a_{s+1} \mid a_{s+2} \mid \cdots \mid a_q$ are nonzero and nonunits and $a_{q+1} = \cdots = a_n = 0$, so we define the m ideals $\mathfrak{a}_i$ as follows:

$$\mathfrak{a}_i = \begin{cases} (0) & \text{if } 1 \leq i \leq r \\ a_{r+q+1-i}A & \text{if } r+1 \leq i \leq m. \end{cases}$$

With these definitions, the ideals $\mathfrak{a}_i$ are proper ideals and we have

$$\mathfrak{a}_i \subseteq \mathfrak{a}_{i+1}, \quad i = 1, \ldots, m-1.$$

When $r = 0$, since $a_{s+1} \mid a_{s+2} \mid \cdots \mid a_n$, it is clear that $\mathfrak{a}_1 = a_nA$ is the annihilator of M. The other statements of the theorem are clear. $\square$

Example 35.1. Here is an example of Theorem 35.25. Let M be a $\mathbb{Z}$-module with generators $\{e_1, e_2, e_3, e_4\}$ subject to the relations $6e_3 = 0$, $2e_4 = 0$. Then

$$M \cong \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}/6\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z},$$

where

$$\mathfrak{a}_1 = (0), \qquad \mathfrak{a}_2 = (0), \qquad \mathfrak{a}_3 = (6), \qquad \mathfrak{a}_4 = (2).$$

The natural number r is called the *free rank* or *Betti number* of the module M. The generators $\alpha_1, \ldots, \alpha_m$ of the ideals $\mathfrak{a}_1, \ldots, \mathfrak{a}_m$ (defined up to a unit) are often called the *invariant factors* of M (in the notation of Theorem 35.25, the generators of the ideals $\mathfrak{a}_1, \ldots, \mathfrak{a}_m$ are denoted by $a_q, \ldots, a_{s+1}$, $s \leq q$).

As corollaries of Theorem 35.25, we obtain again the following facts established in Section 35.1:

1. A finitely generated module over a PID is the direct sum of its torsion module and a free module.

2. A finitely generated torsion-free module over a PID is free.

It turns out that the ideals $\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$ are uniquely determined by the module M. Uniqueness proofs found in most books tend to be intricate and not very intuitive. The shortest proof that we are aware of is from Bourbaki [26] (Chapter VII, Section 4), and uses wedge products.

The following preliminary results are needed.

Proposition 35.26. *If A is a commutative ring and if $\mathfrak{a}_1, \ldots, \mathfrak{a}_m$ are ideals of A, then there is an isomorphism*

$$A/\mathfrak{a}_1 \otimes \cdots \otimes A/\mathfrak{a}_m \approx A/(\mathfrak{a}_1 + \cdots + \mathfrak{a}_m).$$

Sketch of proof. We proceed by induction on m. For $m = 2$, we define the map $\varphi \colon A/\mathfrak{a}_1 \times A/\mathfrak{a}_2 \to A/(\mathfrak{a}_1 + \mathfrak{a}_2)$ by

$$\varphi(\overline{a}, \overline{b}) = ab \pmod{\mathfrak{a}_1 + \mathfrak{a}_2}.$$

It is well-defined because if $a' = a + a_1$ and $b' = b + a_2$ with $a_1 \in \mathfrak{a}_1$ and $a_2 \in \mathfrak{a}_2$, then

$$a'b' = (a + a_1)(b + a_2) = ab + ba_1 + aa_2 + a_1a_2,$$

and so

$$a'b' \equiv ab \pmod{\mathfrak{a}_1 + \mathfrak{a}_2}.$$

It is also clear that this map is bilinear, so it induces a linear map $\varphi \colon A/\mathfrak{a}_1 \otimes A/\mathfrak{a}_2 \to A/(\mathfrak{a}_1 + \mathfrak{a}_2)$ such that $\varphi(\overline{a} \otimes \overline{b}) = ab \pmod{\mathfrak{a}_1 + \mathfrak{a}_2}$.

Next, observe that any arbitrary tensor

$$\overline{a}_1 \otimes \overline{b}_1 + \cdots + \overline{a}_n \otimes \overline{b}_n$$

in $A/\mathfrak{a}_1 \otimes A/\mathfrak{a}_2$ can be rewritten as

$$\overline{1} \otimes (\overline{a_1 b_1} + \cdots + \overline{a_n b_n}),$$

which is of the form $\overline{1} \otimes \overline{s}$, with $s \in A$. We can use this fact to show that φ is injective and surjective, and thus an isomorphism.

For example, if $\varphi(\overline{1} \otimes \overline{s}) = 0$, because $\varphi(\overline{1} \otimes \overline{s}) = s \pmod{\mathfrak{a}_1 + \mathfrak{a}_2}$, we have $s \in \mathfrak{a}_1 + \mathfrak{a}_2$, so we can write $s = a + b$ with $a \in \mathfrak{a}_1$ and $b \in \mathfrak{a}_2$. Then

$$\begin{aligned}\overline{1} \otimes \overline{s} &= \overline{1} \otimes \overline{a+b} \\ &= \overline{1} \otimes (\overline{a} + \overline{b}) \\ &= \overline{1} \otimes \overline{a} + \overline{1} \otimes \overline{b} \\ &= \overline{a} \otimes \overline{1} + \overline{1} \otimes \overline{b} \\ &= 0 + 0 = 0,\end{aligned}$$

since $a \in \mathfrak{a}_1$ and $b \in \mathfrak{a}_2$, which proves injectivity. $\square$

Recall that the exterior algebra of an A-module M is defined by

$$\bigwedge M = \bigoplus_{k \geq 0} \bigwedge^k (M).$$

Proposition 35.27. *If A is a commutative ring, then for any n modules M_i, there is an isomorphism*

$$\bigwedge (\bigoplus_{i=1}^n M_i) \approx \bigotimes_{i=1}^n \bigwedge M_i.$$

A proof can be found in Bourbaki [25] (Chapter III, Section 7, No 7, Proposition 10).

Proposition 35.28. *Let A be a commutative ring and let $\mathfrak{a}_1, \ldots, \mathfrak{a}_n$ be n ideals of A. If the module M is the direct sum of n cyclic modules*

$$M = A/\mathfrak{a}_1 \oplus \cdots \oplus A/\mathfrak{a}_n,$$

then for every $p > 0$, the exterior power $\bigwedge^p M$ is isomorphic to the direct sum of the modules $A/\mathfrak{a}_H$, where H ranges over all subsets $H \subseteq \{1, \ldots, n\}$ with p elements, and with

$$\mathfrak{a}_H = \sum_{h \in H} \mathfrak{a}_h.$$

Proof. If u_i is the image of 1 in $A/\mathfrak{a}_i$, then $A/\mathfrak{a}_i$ is equal to Au_i. By Proposition 35.27, we have

$$\bigwedge M \approx \bigotimes_{i=1}^n \bigwedge (Au_i).$$

We also have

$$\bigwedge (Au_i) = \bigoplus_{k \geq 0} \bigwedge^k (Au_i) \approx A \oplus Au_i,$$

since $au_i \wedge bu_i = 0$, and it follows that

$$\bigwedge^p M \approx \bigoplus_{\substack{H\subseteq\{1,\ldots,n\}\\ H=\{k_1,\ldots,k_p\}}} (Au_{k_1}) \otimes \cdots \otimes (Au_{k_p}).$$

However, by Proposition 35.26, we have

$$(Au_{k_1}) \otimes \cdots \otimes (Au_{k_p}) = A/\mathfrak{a}_{k_1} \otimes \cdots \otimes A/\mathfrak{a}_{k_p} \approx A/(\mathfrak{a}_{k_1} + \cdots + \mathfrak{a}_{k_p}) = A/\mathfrak{a}_H.$$

Therefore,

$$\bigwedge^p M \approx \bigoplus_{\substack{H\subseteq\{1,\ldots,n\}\\ |H|=p}} A/\mathfrak{a}_H,$$

as claimed. □

Example 35.1 continued: Recall that M is the $\mathbb{Z}$-module generated by $\{e_1, e_2, e_3, e_4\}$ subject to $6e_3 = 0$, $2e_2 = 0$. Then

$$\begin{aligned}
\bigwedge^1 M &= \mathrm{span}\{e_1, e_2, e_3, e_4\}\\
\bigwedge^2 M &= \mathrm{span}\{e_1\wedge e_2, e_1\wedge e_3, e_1\wedge e_4, e_2\wedge e_3, e_2\wedge e_4, e_3\wedge e_4\}\\
\bigwedge^3 M &= \mathrm{span}\{e_1\wedge e_2\wedge e_3, e_1\wedge e_2\wedge e_4, e_1\wedge e_3\wedge e_4, e_2\wedge e_3\wedge e_4\}\\
\bigwedge^3 M &= \mathrm{span}\{e_1\wedge e_2\wedge e_3\wedge e_4\}.
\end{aligned}$$

Since $6e_3 = 0$, each element of $\{e_1\wedge e_3, e_2\wedge e_3, e_1\wedge e_2\wedge e_3\}$ is annihilated by $6\mathbb{Z} = (6)$. Since $2e_4 = 0$, each element of $\{e_1\wedge e_4, e_2\wedge e_4, e_3\wedge e_4, e_1\wedge e_2\wedge e_4, e_1\wedge e_3\wedge e_4, e_2\wedge e_3\wedge e_4, e_1\wedge e_2\wedge e_3\wedge e_4\}$ is annihilated by $2\mathbb{Z} = (2)$. We have shown that

$$M \cong \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}/(6) \oplus \mathbb{Z}/(2),$$

where $\mathfrak{a}_1 = (0) = \mathfrak{a}_2$, $\mathfrak{a}_3 = (6)$,and $\mathfrak{a}_4 = (2)$. Then Proposition 35.28 implies that

$$\begin{aligned}
\bigwedge^1 M &\cong \mathbb{Z}/\mathfrak{a}_1 \oplus \mathbb{Z}/\mathfrak{a}_2 \oplus \mathbb{Z}/\mathfrak{a}_3 \oplus \mathbb{Z}/\mathfrak{a}_4 = \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}/(6) \oplus \mathbb{Z}/(2)\\
\bigwedge^2 M &\cong \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_2) \oplus \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_3) \oplus \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_4) \oplus \mathbb{Z}/(\mathfrak{a}_2+\mathfrak{a}_3) \oplus \mathbb{Z}/(\mathfrak{a}_2+\mathfrak{a}_3)\\
&\qquad \oplus \mathbb{Z}/(\mathfrak{a}_3+\mathfrak{a}_4) = \mathbb{Z} \oplus \mathbb{Z}/(6) \oplus \mathbb{Z}/(2) \oplus \mathbb{Z}/(6) \oplus \mathbb{Z}/(2) \oplus \mathbb{Z}/(2)\\
\bigwedge^3 M &\cong \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_2+\mathfrak{a}_3) \oplus \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_2+\mathfrak{a}_4) \oplus \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_3+\mathfrak{a}_4) \oplus \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_3+\mathfrak{a}_4)\\
&= \mathbb{Z}/(6) \oplus \mathbb{Z}/(2) \oplus \mathbb{Z}/(2) \oplus \mathbb{Z}(2)\\
\bigwedge^4 M &\cong \mathbb{Z}/(\mathfrak{a}_1+\mathfrak{a}_2+\mathfrak{a}_3+\mathfrak{a}_4) = \mathbb{Z}/(2).
\end{aligned}$$

When the ideals $\mathfrak{a}_i$ form a chain of inclusions $\mathfrak{a}_1 \subseteq \cdots \subseteq \mathfrak{a}_n$, we get the following remarkable result.

Proposition 35.29. *Let A be a commutative ring and let $\mathfrak{a}_1, \ldots, \mathfrak{a}_n$ be n ideals of A such that $\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_n$. If the module M is the direct sum of n cyclic modules*

$$M = A/\mathfrak{a}_1 \oplus \cdots \oplus A/\mathfrak{a}_n,$$

then for every p with $1 \leq p \leq n$, the ideal $\mathfrak{a}_p$ is the annihilator of the exterior power $\bigwedge^p M$. If $\mathfrak{a}_n \neq A$, then $\bigwedge^p M \neq (0)$ for $p = 1, \ldots, n$, and $\bigwedge^p M = (0)$ for $p > n$.

Proof. With the notation of Proposition 35.28, we have $\mathfrak{a}_H = \mathfrak{a}_{\max(H)}$, where $\max(H)$ is the greatest element in the set H. Since $\max(H) \geq p$ for any subset with p elements and since $\max(H) = p$ when $H = \{1, \ldots, p\}$, we see that

$$\mathfrak{a}_p = \bigcap_{\substack{H \subseteq \{1,\ldots,n\} \\ |H| = p}} \mathfrak{a}_H.$$

By Proposition 35.28, we have

$$\bigwedge^p M \approx \bigoplus_{\substack{H \subseteq \{1,\ldots,n\} \\ |H| = p}} A/\mathfrak{a}_H$$

which proves that $\mathfrak{a}_p$ is indeed the annihilator of $\bigwedge^p M$. The rest is clear. $\square$

Example 35.1 continued: Recall that M is the $\mathbb{Z}$-module generated by $\{e_1, e_2, e_3, e_4\}$ subject to $6e_3 = 0$, $2e_2 = 0$. Then

$$\begin{aligned}
\bigwedge^1 M &= \mathrm{span}\{e_1, e_2, e_3, e_4\} \\
\bigwedge^2 M &= \mathrm{span}\{e_1 \wedge e_2, e_1 \wedge e_3, e_1 \wedge e_4, e_2 \wedge e_3, e_2 \wedge e_4, e_3 \wedge e_4\} \\
\bigwedge^3 M &= \mathrm{span}\{e_1 \wedge e_2 \wedge e_3, e_1 \wedge e_2 \wedge e_4, e_1 \wedge e_3 \wedge e_4, e_2 \wedge e_3 \wedge e_4\} \\
\bigwedge^3 M &= \mathrm{span}\{e_1 \wedge e_2 \wedge e_3 \wedge e_4\}.
\end{aligned}$$

Since e_1 and e_2 are free, $e_1 \wedge e_2$ is also free. Since $6e_3 = 0$, each element of $\{e_1 \wedge e_3, e_2 \wedge e_3, e_1 \wedge e_2 \wedge e_3\}$ is annihilated by $6\mathbb{Z} = (6)$. Since $2e_4 = 0$, each element of $\{e_1 \wedge e_4, e_2 \wedge e_4, e_3 \wedge e_4, e_1 \wedge e_2 \wedge e_4, e_1 \wedge e_3 \wedge e_4, e_2 \wedge e_3 \wedge e_4, e_1 \wedge e_2 \wedge e_3 \wedge e_4\}$ is annihilated by $2\mathbb{Z} = (2)$.

Then

$$\begin{aligned}
\mathrm{Ann}(\bigwedge^1 M) &= \mathrm{Ann}\, e_1 = (0) \\
\mathrm{Ann}(\bigwedge^2 M) &= \mathrm{Ann}\, e_1 \wedge e_2 = (0) \\
\mathrm{Ann}(\bigwedge^3 M) &= \mathrm{Ann}\, e_1 \wedge e_2 \wedge e_3 = (6) \\
\mathrm{Ann}(\bigwedge^4 M) &= \mathrm{Ann}\, e_1 \wedge e_2 \wedge e_3 \wedge e_4 = (2),
\end{aligned}$$

and Proposition 35.29 provides another verification of

$$M \cong \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}/(6) \oplus \mathbb{Z}/(2).$$

Propostion 35.29 immediately implies the following crucial fact.

Proposition 35.30. *Let A be a commutative ring and let $\mathfrak{a}_1, \ldots, \mathfrak{a}_m$ be m ideals of A and $\mathfrak{a}'_1, \ldots, \mathfrak{a}'_n$ be n ideals of A such that $\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$ and $\mathfrak{a}'_1 \subseteq \mathfrak{a}'_2 \subseteq \cdots \subseteq \mathfrak{a}'_n \neq A$ If we have an isomorphism*

$$A/\mathfrak{a}_1 \oplus \cdots \oplus A/\mathfrak{a}_m \approx A/\mathfrak{a}'_1 \oplus \cdots \oplus A/\mathfrak{a}'_n,$$

then $m = n$ and $\mathfrak{a}_i = \mathfrak{a}'_i$ for $i = 1, \ldots, n$.

Proposition 35.30 yields the uniqueness of the decomposition in Theorem 35.25.

Theorem 35.31. *(Invariant Factors Decomposition) Let M be a finitely generated nontrivial A-module, where A a PID. Then, M is isomorphic to a direct sum of cyclic modules*

$$M \approx A/\mathfrak{a}_1 \oplus \cdots \oplus A/\mathfrak{a}_m,$$

where the $\mathfrak{a}_i$ are proper ideals of A (possibly zero) such that

$$\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_m \neq A.$$

More precisely, if $\mathfrak{a}_1 = \cdots = \mathfrak{a}_r = (0)$ and $(0) \neq \mathfrak{a}_{r+1} \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$, then

$$M \approx A^r \oplus (A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m),$$

where $A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m$ is the torsion submodule of M. The module M is free iff $r = m$, and a torsion-module iff $r = 0$. In the latter case, the annihilator of M is $\mathfrak{a}_1$. Furthermore, the integer r and ideals $\mathfrak{a}_1 \subseteq \mathfrak{a}_2 \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$ are uniquely determined by M.

Proof. By Theorem 35.7, since $M_{\mathrm{tor}} = A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m$, we know that the dimension r of the free summand only depends on M. The uniqueness of the sequence of ideals follows from Proposition 35.30. $\square$

In view of the uniqueness part of Theorem 35.31, we make the following definition.

Definition 35.12. Given a finitely generated module M over a PID A as in Theorem 35.31, the ideals $\mathfrak{a}_i = \alpha_i A$ are called the *invariant factors* of M. The generators α_i of these ideals (uniquely defined up to a unit) are also called the *invariant factors* of M.

Proposition 35.23 can be sharpened as follows:

Proposition 35.32. *Let F be a finitely generated free module over a PID A, and let M be any submodule of F. Then, M is a free module and there is a basis $(e_1, ..., e_n)$ of F, some $q \leq n$, and some nonzero elements $a_1, \ldots, a_q \in A$, such that $(a_1 e_1, \ldots, a_q e_q)$ is a basis of M and a_i divides a_{i+1} for all i, with $1 \leq i \leq q-1$. Furthermore, the free module M' with basis $(e_1, \ldots, e_q)$ and the ideals $a_1 A, \ldots, a_q A$ are uniquely determined by M; the quotient module M'/M is the torsion module of F/M, and we have an isomorphism*

$$M'/M \approx A/a_1 A \oplus \cdots \oplus A/a_q A.$$

Proof. Since $a_i \neq 0$ for $i = 1, \ldots, q$, observe that

$$M' = \{x \in F \mid (\exists \beta \in A,\ \beta \neq 0)(\beta x \in M)\},$$

which shows that M'/M is the torsion module of F/M. Therefore, M' is uniquely determined. Since

$$M = A a_1 e_1 \oplus \cdots \oplus A a_q e_q,$$

by Proposition 35.24 we have an isomorphism

$$M'/M \approx A/a_1 A \oplus \cdots \oplus A/a_q A.$$

Now, it is possible that the first s elements a_i are units, in which case $A/a_i A = (0)$, so we can eliminate such factors and we get

$$M'/M \approx A/a_{s+1} A \oplus \cdots \oplus A/a_q A,$$

with $a_q A \subseteq a_{q-1} A \subseteq \cdots \subseteq a_{s+1} A \neq A$. By Proposition 35.30, $q - s$ and the ideals $a_j A$ are uniquely determined for $j = s+1, \ldots, q$, and since $a_1 A = \cdots = a_s A = A$, the q ideals $a_i A$ are uniquely determined. $\square$

The ideals $a_1 A, \ldots, a_q A$ of Proposition 35.32 are called the *invariant factors of M with respect to F*. They *should not be confused* with the invariant factors of a module M.

It turns out that $a_1, \ldots, a_q$ can also be computed in terms of gcd's of minors of a certain matrix. Recall that if X is an $m \times n$ matrix, then a $k \times k$ minor of X is the determinant of any $k \times k$ matrix obtained by picking k columns of X, and then k rows from these k columns.

Proposition 35.33. *Let F be a free module of finite dimension over a PID, $(u_1, \ldots, u_n)$ be a basis of F, M be a submodule of F, and $(x_1, \ldots, x_m)$ be a set of generators of M. If $a_1A, \ldots, a_qA$ are the invariant factors of M with respect to F as in Proposition 35.32, then for $k = 1, \ldots, q$, the product $a_1 \cdots a_k$ is a gcd of the $k \times k$ minors of the $n \times m$ matrix X_U whose columns are the coordinates of the x_j over the u_i.*

Proof. Proposition 35.23 shows that $M \subseteq a_1F$. Consequently, the coordinates of any element of M are multiples of a_1. On the other hand, we know that there is a linear form f for which a_1A is a maximal ideal and some $e' \in M$ such that $f(e') = a_1$. If we write e' as a linear combination of the x_i, we see that a_1 belongs to the ideal spanned by the coordinates of the x_i over the basis $(u_1, \ldots, u_n)$. Since these coordinates are all multiples of a_1, it follows that a_1 is their gcd, which proves the case $k = 1$.

For any $k \geq 2$, consider the exterior power $\bigwedge^k M$. Using the notation of the proof of Proposition 35.23, the module M has the basis $(a_1e_1, \ldots, a_qe_q)$, so $\bigwedge^k M$ has a basis consisting of elements of the form

$$a_{i_1}e_{i_1} \wedge \cdots \wedge a_{i_k}e_{i_k} = a_{i_1} \cdots a_{i_k}\, e_{i_1} \wedge \cdots \wedge e_{i_k},$$

for all sequences $(i_1, \ldots, i_k)$ such that $1 \leq i_1 < i_2 < \cdots < i_k \leq q$. However, the vectors $e_{i_1} \wedge \cdots \wedge e_{i_k}$ form a basis of $\bigwedge^k F$. Thus, the map from $\bigwedge^k M$ into $\bigwedge^k F$ induced by the inclusion $M \subseteq F$ defines an isomorphism of $\bigwedge^k M$ onto the submodule of $\bigwedge^k F$ having the elements $a_{i_1} \cdots a_{i_k}\, e_{i_1} \wedge \cdots \wedge e_{i_k}$ as a basis. Since a_j is a multiple of the a_i for $i < j$, the products $a_{i_1} \cdots a_{i_k}$ are all multiples of $\delta_k = a_1 \cdots a_k$, and one of these is equal to δ_k. The reasoning used for $k = 1$ shows that δ_k is a gcd of the set of coordinates of any spanning set of $\bigwedge^k M$ over any basis of $\bigwedge^k F$. If we pick as basis of $\bigwedge^k F$ the wedge products $u_{i_1} \wedge \cdots \wedge u_{i_k}$, and as generators of $\bigwedge^k M$ the wedge products $x_{i_1} \wedge \cdots \wedge x_{i_k}$, it is easy to see that the coordinates of the $x_{i_1} \wedge \cdots \wedge x_{i_k}$ are indeed determinants which are the $k \times k$ minors of the matrix X_U. ☐

Proposition 35.33 yields $a_1, \ldots, a_q$ (up to units) as follows: First, a_1 is a gcd of the entries in X_U. Having computed $a_1, \ldots, a_k$, let $b_k = a_1 \cdots, a_k$, compute $b_{k+1} = a_1 \cdots a_k a_{k+1}$ as a gcd of all the $(k+1) \times (k+1)$ minors of X_U, and then a_{k+1} is obtained by dividing b_{k+1} by b_k (recall that a PID is an integral domain).

We also have the following interesting result about linear maps between free modules over a PID.

Proposition 35.34. *Let A be a PID, let F be a free module of dimension n, F' be a free module of dimension m, and $f\colon F \to F'$ be a linear map from F to F'. Then, there exist a basis $(e_1, \ldots, e_n)$ of F, a basis $(e'_1, \ldots, e'_m)$ of F', and some nonzero elements $\alpha_1, \ldots \alpha_r \in A$ such that*

$$f(e_i) = \begin{cases} \alpha_i e'_i & \text{if } 1 \leq i \leq r \\ 0 & \text{if } r+1 \leq i \leq n, \end{cases}$$

and $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$. Furthermore, the ideals $\alpha_1 A, \ldots, \alpha_r A$ are the invariant factors of $f(F)$ with respect F'.

Proof. Let F_0 be the kernel of f. Since $M' = f(F)$ is a submodule of the free module F', it is free, and similarly F_0 is free as a submodule of the free module F (by Proposition 35.23). By Proposition 35.2, we have

$$F = F_0 \oplus F_1,$$

where F_1 is a free module, and the restriction of f to F_1 is an isomorphism onto $M' = f(F)$. Proposition 35.32 applied to F' and M' yields a basis $(e'_1, \ldots, e'_m)$ of F' such that $(\alpha_1 e'_1, \ldots, \alpha_r e'_r)$ is a basis of M', where $\alpha_1 A, \ldots, \alpha_r A$ are the invariant factors for M' with respect to F'. Since the restriction of f to F_1 is and isomorphism, there is a basis $(e_1, \ldots, e_r)$ of F_1 such that

$$f(e_i) = \alpha_i e'_i, \quad i = 1, \ldots, r.$$

We can extend this basis to a basis of F by picking a basis of F_0 (a free module), which yields the desired result. □

The matrix version of Proposition 35.34 is the following proposition.

Proposition 35.35. *If X is an $m \times n$ matrix of rank r over a PID A, then there exist some invertible $n \times n$ matrix P, some invertible $m \times m$ matrix Q, and a $m \times n$ matrix D of the form*

$$D = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \alpha_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

for some nonzero $\alpha_i \in A$, such that

(1) $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$,

(2) $X = QDP^{-1}$, *and*

(3) *The α_is are uniquely determined up to a unit.*

The ideals $\alpha_1 A, \ldots, \alpha_r A$ are called the *invariant factors* of the matrix X. Recall that two $m \times n$ matrices X and Y are *equivalent* iff

$$Y = QXP^{-1},$$

for some invertible matrices, P and Q. Then, Proposition 35.35 implies the following fact.

Proposition 35.36. *Two $m \times n$ matrices X and Y are equivalent iff they have the same invariant factors.*

If X is the matrix of a linear map $f\colon F \to F'$ with respect to some basis $(u_1, \ldots, u_n)$ of F and some basis $(u'_1, \ldots, u'_m)$ of F', then the columns of X are the coordinates of the $f(u_j)$ over the u'_i, where the $f(u_j)$ generate $f(F)$, so Proposition 35.33 applies and yields the following result:

Proposition 35.37. *If X is a $m \times n$ matrix or rank r over a PID A, and if $\alpha_1 A, \ldots, \alpha_r A$ are its invariant factors, then α_1 is a gcd of the entries in X, and for $k = 2, \ldots, r$, the product $\alpha_1 \cdots \alpha_k$ is a gcd of all $k \times k$ minors of X.*

There are algorithms for converting a matrix X over a PID to the form $X = QDP^{-1}$ as described in Proposition 35.35. For Euclidean domains, this can be achieved by using the elementary row and column operations $P(i, k)$, $E_{i,j;\beta}$, and $E_{i,\lambda}$ described in Chapter 8, where we require the scalar λ used in $E_{i,\lambda}$ to be a unit. For an arbitrary PID, another kind of elementary matrix (containing some 2×2 submatrix in addition to diagonal entries) is needed. These procedures involve computing gcd's and use the Bezout identity to mimic division. Such methods are presented in D. Serre [156], Jacobson [98], and Van Der Waerden [179], and sketched in Artin [7]. We describe and justify several of these methods in Section 36.5.

Proposition 35.32 has the following two applications.

First, consider a finitely presented module M over a PID given by some $m \times n$ matrix R. By Proposition 35.35, the matrix R can be diagonalized as $R = QDP^{-1}$ where D is a diagonal matrix. Then, we see that M has a presentation with m generators and r relations of the form

$$\alpha_i e_i = 0,$$

where $\alpha_i \neq 0$ and $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$.

For the second application, let F be a free module with basis $(e_1, \ldots, e_n)$, and let M be a submodule of F generated by m vectors $v_1, \ldots, v_m$ in F. The module M can be viewed as the set of linear combinations of the columns of the $n \times m$ matrix also denoted M consisting of the coordinates of the vectors $v_1, \ldots, v_m$ over the basis $(e_1, \ldots, e_n)$. Then by Proposition 35.35, the matrix R can be diagonalized as $R = QDP^{-1}$ where D is a diagonal matrix. The columns of Q form a basis $(e'_1, \ldots, e'_n)$ of F, and since $RP = QD$, the nonzero columns of RP form the basis $(a_1 e'_1, \ldots, a_q e'_q)$ of M.

When the ring A is a Euclidean domain, Theorem 36.18 shows that P and Q are products of elementary row and column operations. In particular, when $A = \mathbb{Z}$, in which cases our $\mathbb{Z}$-modules are abelian groups, we can find P and Q using Euclidean division.

If $A = \mathbb{Z}$, a finitely generated submodule M of $\mathbb{Z}^n$ is called a *lattice*. It is given as the set of integral linear combinations of a finite set of integral vectors.

Here is an example taken from Artin [7] (Chapter 12, Section 4). Let F be the free $\mathbb{Z}$-module $\mathbb{Z}^2$, and let M be the lattice generated by the columns of the matrix

$$R = \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}.$$

The columns (u_1, u_2) of R are linearly independent, but they are not a basis of $\mathbb{Z}^2$. For example, in order to obtain e_1 as a linear combination of these columns, we would need to solve the linear system

$$\begin{aligned} 2x - y &= 1 \\ x + 2y &= 0. \end{aligned}$$

From the second equation, we get $x = -2y$, which yields

$$-5y = 1.$$

But, $y = -1/5$ is not an integer. We leave it as an exercise to check that

$$\begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix},$$

which means that

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix},$$

so $R = QDP^{-1}$ with

$$Q = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

The new basis (u_1', u_2') for $\mathbb{Z}^2$ consists of the columns of Q and the new basis for M consists of the columns $(u_1', 5u_2')$ of QD, where

$$QD = \begin{pmatrix} 1 & 0 \\ 3 & 5 \end{pmatrix}.$$

A picture of the lattice and its generators (u_1, u_2) and of the same lattice with the new basis $(u_1', 5u_2')$ is shown in Figure 35.1, where the lattice points are displayed as stars.

The invariant factor decomposition of a finitely generated module M over a PID A given by Theorem 35.31 says that

$$M_{\mathrm{tor}} \approx A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m,$$

a direct sum of cyclic modules, with $(0) \neq \mathfrak{a}_{r+1} \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$. Using the Chinese Remainder Theorem (Theorem 32.15), we can further decompose each module $A/\alpha_i A$ into a direct sum of modules of the form $A/p^n A$, where p is a prime in A.

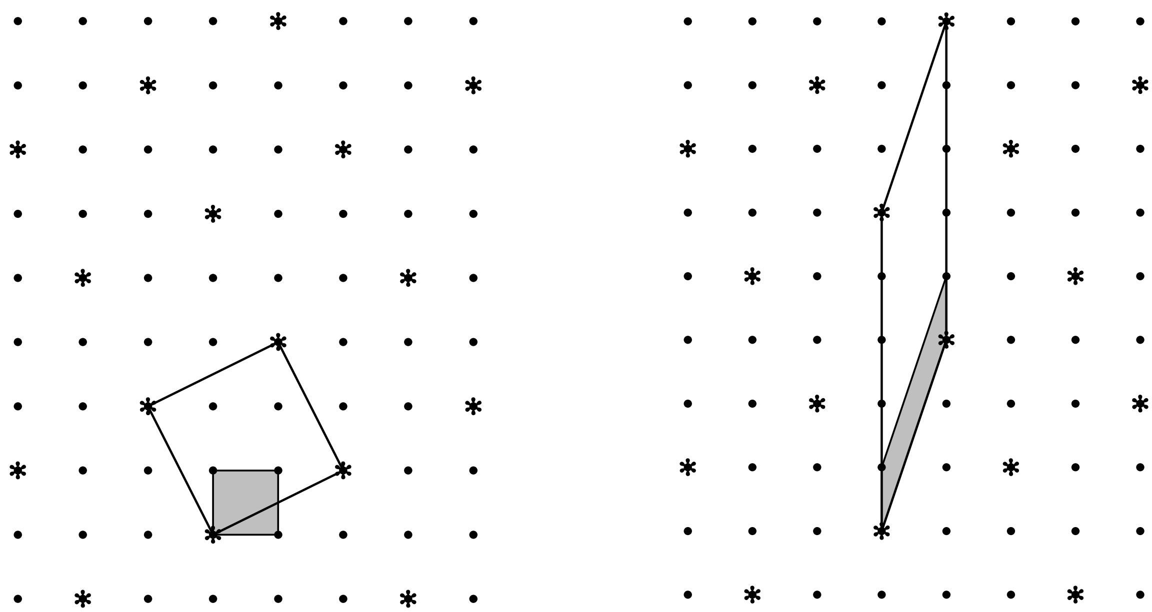


Figure 35.1: Diagonalization applied to a lattice

Theorem 35.38. *(Elementary Divisors Decomposition) Let M be a finitely generated nontrivial A-module, where A a PID. Then, M is isomorphic to the direct sum $A^r \oplus M_{\mathrm{tor}}$, where A^r is a free module and where the torsion module M_{tor} is a direct sum of cyclic modules of the form $A/p_i^{n_{i,j}}$, for some primes $p_1, \ldots, p_t \in A$ and some positive integers $n_{i,j}$, such that for each $i = 1, \ldots, t$, there is a sequence of integers*

$$1 \leq \underbrace{n_{i,1}, \ldots, n_{i,1}}_{m_{i,1}} < \underbrace{n_{i,2}, \ldots, n_{i,2}}_{m_{i,2}} < \cdots < \underbrace{n_{i,s_i}, \ldots, n_{i,s_i}}_{m_{i,s_i}},$$

with $s_i \geq 1$, and where $n_{i,j}$ occurs $m_{i,j} \geq 1$ times, for $j = 1, \ldots, s_i$. Furthermore, the irreducible elements p_i and the integers $r, t, n_{i,j}, s_i$, $m_{i,j}$ are uniquely determined.

Proof. By Theorem 35.31, we already know that $M \approx A^r \oplus M_{\mathrm{tor}}$, where r is uniquely determined, and where

$$M_{\mathrm{tor}} \approx A/\mathfrak{a}_{r+1} \oplus \cdots \oplus A/\mathfrak{a}_m,$$

a direct sum of cyclic modules, with $(0) \neq \mathfrak{a}_{r+1} \subseteq \cdots \subseteq \mathfrak{a}_m \neq A$. Then, each $\mathfrak{a}_i$ is a principal ideal of the form $\alpha_i A$, where $\alpha_i \neq 0$ and α_i is not a unit. Using the Chinese Remainder Theorem (Theorem 32.15), if we factor α_i into prime factors as

$$\alpha_i = u p_1^{k_1} \cdots p_h^{k_h},$$

with $k_j \geq 1$, we get an isomorphism

$$A/\alpha_i A \approx A/p_1^{k_1} A \oplus \cdots \oplus A/p_h^{k_h} A.$$

This implies that M_{tor} is the direct sum of modules of the form $A/p_i^{n_{i,j}}$, for some primes $p_i \in A$.

To prove uniqueness, observe that the p_i-primary component of M_{tor} is the direct sum

$$(A/p_i^{n_{i,1}}A)^{m_{i,1}} \oplus \cdots \oplus (A/p_i^{n_{i,s_i}}A)^{m_{i,s_i}},$$

and these are uniquely determined. Since $n_{i,1} < \cdots < n_{i,s_i}$, we have

$$p_i^{n_{i,s_i}}A \subseteq \cdots \subseteq p_i^{n_{i,1}}A \neq A,$$

Proposition 35.30 implies that the irreducible elements p_i and $n_{i,j}$, s_i, and $m_{i,j}$ are unique. $\square$

In view of Theorem 35.38, we make the following definition.

Definition 35.13. Given a finitely generated module M over a PID A as in Theorem 35.38, the ideals $p_i^{n_{i,j}}A$ are called the *elementary divisors* of M, and the $m_{i,j}$ are their *multiplicities*. The ideal (0) is also considered to be an elementary divisor and r is its multiplicity.

Remark: Theorem 35.38 shows how the elementary divisors are obtained from the invariant factors: the elementary divisors are the prime power factors of the invariant factors.

Conversely, we can get the invariant factors from the elementary divisors. We may assume that M is a torsion module. Let

$$m = \max_{1 \leq i \leq t}\{m_{i,1} + \cdots + m_{i,s_i}\},$$

and construct the $t \times m$ matrix $C = (c_{ij})$ whose ith row is the sequence

$$\underbrace{n_{i,s_i}, \ldots, n_{i,s_i}}_{m_{i,s_i}}, \ldots, \underbrace{n_{i,2}, \ldots, n_{i,2}}_{m_{i,2}}, \underbrace{n_{i,1}, \ldots, n_{i,1}}_{m_{i,1}}, 0, \ldots, 0,$$

padded with 0's if necessary to make it of length m. Then, the jth invariant factor is

$$\alpha_j A = p_1^{c_{1j}} p_2^{c_{2j}} \cdots p_t^{c_{tj}} A.$$

Observe that because the last column contains at least one prime, the α_i are not units, and $\alpha_m \mid \alpha_{m-1} \mid \cdots \mid \alpha_1$, so that $\alpha_1 A \subseteq \cdots \subseteq \alpha_{m-1}A \subseteq \alpha_m A \neq A$, as desired.

From a computational point of view, finding the elementary divisors is usually practically impossible, because it requires factoring. For example, if $A = K[X]$ where K is a field, such as $K = \mathbb{R}$ or $K = \mathbb{C}$, factoring amounts to finding the roots of a polynomial, but by Galois theory, in general, this is not algorithmically doable. On the other hand, the invariant factors can be computed using elementary row and column operations.

It can also be shown that A and the modules of the form A/p^nA are indecomposable (with $n > 0$). A module M is said to be *indecomposable* if M cannot be written as a direct sum of two nonzero proper submodules of M. For a proof, see Bourbaki [26] (Chapter VII, Section 4, No. 8, Proposition 8). Theorem 35.38 shows that a finitely generated module over a PID is a direct sum of indecomposable modules.

In Chapter 36 we apply the structure theorems for finitely generated (torsion) modules to the $K[X]$-module E_f associated with an endomorphism f on a vector space E. First, we need to understand the process of extension of the ring of scalars.

35.6 Extension of the Ring of Scalars

The need to extend the ring of scalars arises, in particular when dealing with eigenvalues. First we need to define how to restrict scalar multiplication to a subring. The situation is that we have two rings A and B, a B-module M, and a ring homomorphism $\rho\colon A \to B$. The special case that arises often is that A is a subring of B (B could be a field) and ρ is the inclusion map. Then we can make M into an A-module by defining the scalar multiplication $\cdot\colon A \times M \to M$ as follows.

Definition 35.14. Given two rings A and B and a ring homomorphism $\rho\colon A \to B$, any B-module M can made into an A-module denoted by $\rho_*(M)$, by defining scalar multiplication by any element of A as follows:

$$a \cdot x = \rho(a)x, \quad \text{for all } a \in A \text{ and all } x \in M.$$

In particular, viewing B as a B-module, we obtain the A-module $\rho_*(B)$.

If M and N are two B-modules and if $f\colon M \to N$ is a B-linear map, the map f is a homomorphism $f\colon \rho_*(M) \to \rho_*(N)$ of the abelian groups $\rho_*(M)$ and $\rho_*(N)$. This map is also A-linear, because for all $u \in M$ and all $a \in A$, by definition of the scalar multiplication by elements of A, we have

$$f(a \cdot u) = f(\rho(a)u) = \rho(a)f(u) = a \cdot f(u).$$

The map $f\colon \rho_*(M) \to \rho_*(N)$ viewed as an A-linear map is denoted by $\rho_*(f)$. As homomorphisms of abelian groups, the maps $f\colon M \to N$ and $\rho_*(f)\colon \rho_*(M) \to \rho_*(N)$ are identical, but f is a B-linear map whereas $\rho_*(f)$ is an A-linear map.

Now we can describe the process of scalar extension. Given any A-module M, we make $\rho_*(B) \otimes_A M$ into a (left) B-module as follows: for every $\beta \in B$, let $\mu_\beta\colon \rho_*(B) \times M \to \rho_*(B) \otimes_A M$ be given by

$$\mu_\beta(\beta', x) = (\beta\beta') \otimes x.$$

The map μ_β is bilinear so it induces a linear map $\mu_\beta\colon \rho_*(B) \otimes_A M \to \rho_*(B) \otimes_A M$ such that

$$\mu_\beta(\beta' \otimes x) = (\beta\beta') \otimes x.$$

If we define the scalar multiplication $\cdot\colon B \times (\rho_*(B) \otimes_A M) \to \rho_*(B) \otimes_A M$ by

$$\beta \cdot z = \mu_\beta(z), \quad \text{for all } \beta \in B \text{ and all } z \in \rho_*(B) \otimes_A M,$$

then it is easy to check that the axioms M1, M2, M3, M4 hold. Let us check M2 and M3. We have

$$\begin{aligned}
\mu_{\beta_1+\beta_2}(\beta' \otimes x) &= (\beta_1 + \beta_2)\beta' \otimes x \\
&= (\beta_1\beta' + \beta_2\beta') \otimes x \\
&= \beta_1\beta' \otimes x + \beta_2\beta' \otimes x \\
&= \mu_{\beta_1}(\beta' \otimes x) + \mu_{\beta_2}(\beta' \otimes x)
\end{aligned}$$

and

$$\begin{aligned}
\mu_{\beta_1\beta_2}(\beta' \otimes x) &= \beta_1\beta_2\beta' \otimes x \\
&= \mu_{\beta_1}(\beta_2\beta' \otimes x) \\
&= \mu_{\beta_1}(\mu_{\beta_2}(\beta' \otimes x)).
\end{aligned}$$

Definition 35.15. Given two rings A and B and a ring homomorphism $\rho\colon A \to B$, for any A-module M, with the scalar multiplication by elements of B given by

$$\beta \cdot (\beta' \otimes x) = (\beta\beta') \otimes x, \quad \beta, \beta' \in B,\ x \in M,$$

the tensor product $\rho_*(B) \otimes_A M$ is a B-module denoted by $\rho^*(M)$, or $M_{(B)}$ when ρ is the inclusion of A into B. The B-module $\rho^*(M)$ is sometimes called the *module induced from M by extension to B of the ring of scalars through ρ*.

Here is a specific example of Definition 35.15. Let $A = \mathbb{R}$, $B = \mathbb{C}$ and ρ be the inclusion map of $\mathbb{R}$ into $\mathbb{C}$, i.e. $\rho\colon \mathbb{R} \to \mathbb{C}$ with $\rho(a) = a$ for $a \in \mathbb{R}$. Let M be an $\mathbb{R}$-module. The field $\mathbb{C}$ is a $\mathbb{C}$-module, and when we restrict scalar multiplication by scalars $\lambda \in \mathbb{R}$, we obtain the $\mathbb{R}$-module $\rho_*(\mathbb{C})$ (which, as an abelian group, is just $\mathbb{C}$). Form $\rho_*(\mathbb{C}) \otimes_\mathbb{R} M$. This is an $\mathbb{R}$-module where typical elements have the form $\sum_{i=1}^n a_i(z_i \otimes m_i)$, $a_i \in \mathbb{R}$, $z_i \in \mathbb{C}$, and $m_i \in M$. Since

$$a_i(z_i \otimes m_i) = a_iz_i \otimes m_i$$

and since $a_iz_i \in \mathbb{C}$ and any element of $\mathbb{C}$ is obtained this way (let $a_i = 1$), the elements of $\rho_*(\mathbb{C}) \otimes_\mathbb{R} M$ can be written as

$$\sum_{i=1}^n z_i \otimes m_i, \quad z_i \in \mathbb{C},\ m_i \in M.$$

We want to make $\rho_*(\mathbb{C}) \otimes_\mathbb{R} M$ into a $\mathbb{C}$-module, denoted $\rho^*(M)$, and thus must describe how a complex number β acts on $\sum_{i=1}^n z_i \otimes m_i$. By linearity, it is enough to determine how $\beta = u + iv$ acts on a generator $z \otimes m$, where $z = x + iy$ and $m \in M$. The action is given by

$$\beta \cdot (z \otimes m) = \beta z \otimes m = (u + iv)(x + iy) \otimes m = (ux - vy + i(uy + vx)) \otimes m,$$

since complex multiplication only makes sense over $\mathbb{C}$.

We claim that $\rho^*(M)$ is isomorphic to the $\mathbb{C}$-module $M \times M$ with addition defined by

$$(u_1, v_1) + (u_2, v_2) = (u_1 + u_2, v_1 + v_2)$$

and scalar multiplication by $\lambda + i\mu \in \mathbb{C}$ as

$$(\lambda + i\mu) \cdot (u, v) = (\lambda u - \mu v, \lambda v + \mu u).$$

Define the map $\alpha_0 \colon \rho_*(\mathbb{C}) \times M \to M \times M$ by

$$\alpha_0(\lambda + i\mu, u) = (\lambda u, \mu u).$$

It is easy to check that α_0 is $\mathbb{R}$-linear, so we obtain an $\mathbb{R}$-linear map $\alpha \colon \rho_*(\mathbb{C}) \otimes_{\mathbb{R}} M \to M \times M$ such that

$$\alpha((\lambda + i\mu) \otimes u) = (\lambda u, \mu u).$$

We also define the map $\beta \colon M \times M \to \rho_*(\mathbb{C}) \otimes_{\mathbb{R}} M$ by

$$\beta(u, v) = 1 \otimes u + i \otimes v.$$

Again, it is clear that this map is $\mathbb{R}$-linear. We can now check that α and β are mutual inverses. We have

$$\alpha(\beta(u, v)) = \alpha(1 \otimes u + i \otimes v) = \alpha(1 \otimes u) + \alpha(i \otimes v) = (u, 0) + (0, v) = (u, v),$$

and on generators,

$$\beta(\alpha((\lambda + i\mu) \otimes u)) = \beta(\lambda u, \mu u) = 1 \otimes \lambda u + i \otimes \mu u = \lambda \otimes u + i\mu \otimes u = (\lambda + i\mu) \otimes u.$$

Therefore, $\rho_*(\mathbb{C}) \otimes_{\mathbb{R}} M$ and $M \times M$ are isomorphic as $\mathbb{R}$-module. However, the isomorphism α is also an isomorphism of $\mathbb{C}$-modules. This is because in $\rho_*(\mathbb{C}) \otimes_{\mathbb{R}} M$, on generators we have

$$(\lambda + i\mu) \cdot ((x + iy) \otimes u) = (\lambda + i\mu)(x + iy) \otimes u = (\lambda x - \mu y + i(\lambda y + \mu x)) \otimes u,$$

so

$$\begin{aligned}\alpha((\lambda + i\mu) \cdot ((x + iy) \otimes u) &= \alpha((\lambda x - \mu y + i(\lambda y + \mu x)) \otimes u) \\ &= ((\lambda x - \mu y)u, (\lambda y + \mu x)u),\end{aligned}$$

and by definition of the scalar multiplication by elements of $\mathbb{C}$ on $M \times M$

$$(\lambda + i\mu) \cdot \alpha((x + iy) \otimes u) = (\lambda + i\mu) \cdot (xu, yu) = ((\lambda x - \mu y)u, (\lambda y + \mu x)u).$$

Therefore, α is an isomorphism between the $\mathbb{C}$-modules $\rho^*(M) = \rho_*(\mathbb{C}) \otimes_{\mathbb{R}} M$ and $M \times M$.

The above process of ring extension can also be applied to linear maps. We have the following proposition whose proof is given in Bourbaki [25] (Chapter II, Section 5, Proposition 1).

Proposition 35.39. *Given a ring homomomorphism $\rho\colon A \to B$ and given any A-module M, the map $\varphi\colon M \to \rho_*(\rho^*(M))$ given by $\varphi(x) = 1 \otimes_A x$ is A-linear and $\varphi(M)$ spans the B-module $\rho^*(M)$. For every B-module N, and for every A-linear map $f\colon M \to \rho_*(N)$, there is a unique B-linear map*

$$\overline{f}\colon \rho^*(M) \to N$$

such that

$$\rho_*(\overline{f}) \circ \varphi = f$$

as in the following commutative diagram

$$\begin{tikzcd} M \arrow[r, "\varphi"] \arrow[dr, "f"'] & \rho_*(\rho^*(M)) \arrow[d, "\rho_*(\overline{f})"] \\ & \rho_*(N) \end{tikzcd}$$

or equivalently,

$$\overline{f}(1 \otimes_A x) = f(x), \quad \text{for all } x \in M.$$

As a consequence of Proposition 35.39, we obtain the following result.

Proposition 35.40. *Given a ring homomomorphism $\rho\colon A \to B$, for any two A-modules M and N, for every A-linear map $f\colon M \to N$, there is a unique B-linear map $\rho^*(f)\colon \rho^*(M) \to \rho^*(N)$ (also denoted $\overline{f}$) given by*

$$\rho^*(f) = \mathrm{id}_B \otimes f,$$

such that the following diagam commutes:

$$\begin{tikzcd} M \arrow[r, "\varphi_M"] \arrow[d, "f"'] & \rho_*(\rho^*(M)) \arrow[d, "\rho_*(\rho^*(f))"] \\ N \arrow[r, "\varphi_N"'] & \rho_*(\rho^*(N)) \end{tikzcd}$$

Proof. Apply Proposition 35.40 to the A-linear map $\varphi_N \circ f$. ∎

If S spans the module M, it is clear that $\varphi(S)$ spans $\rho^*(M)$. In particular, if M is finitely generated, so if $\rho^*(M)$. Bases of M also extend to bases of $\rho^*(M)$.

Proposition 35.41. *Given a ring homomomorphism $\rho\colon A \to B$, for any A-module M, if $(u_1, \ldots, u_n)$ is a basis of M, then $(\varphi(u_1), \ldots, \varphi(u_n))$ is a basis of $\rho^*(M)$, where φ is the A-linear map $\varphi\colon M \to \rho_*(\rho^*(M))$ given by $\varphi(x) = 1 \otimes_A x$. Furthermore, if ρ is injective, then so is φ.*

Proof. The first assertion follows immediately from Proposition 35.13, since it asserts that every element z of $\rho^*(M) = \rho_*(B) \otimes_A M$ can be written in a unique way as

$$z = b_1 \otimes u_1 + \cdots + b_n \otimes u_n = b_1(1 \otimes u_1) + \cdots + b_n(1 \otimes u_n),$$

and $\varphi(u_i) = 1 \otimes u_i$. Next, if ρ is injective, by definition of the scalar multiplication in the A-module $\rho_*(\rho^*(M))$, we have $\varphi(a_1u_1 + \cdots + a_nu_n) = 0$ iff

$$\rho(a_1)\varphi(u_1) + \cdots + \rho(a_n)\varphi(u_n) = 0,$$

and since $(\varphi(u_1), \ldots, \varphi(u_n))$ is a basis of $\rho^*(M)$, we must have $\rho(a_i) = 0$ for $i = 1, \ldots, n$, which (by injectivity of ρ) implies that $a_i = 0$ for $i = 1, \ldots, n$. Therefore, φ is injective. $\square$

In particular, if A is a subring of B, then ρ is the inclusion map and Proposition 35.41 shows that a basis of M becomes a basis of $M_{(B)}$ and that M is embedded into $M_{(B)}$. It is also easy to see that if M and N are two free A-modules and $f\colon M \to N$ is a linear map represented by the matrix X with respect to some bases $(u_1, \ldots, u_n)$ of M and $(v_1, \ldots, v_m)$ of N, then the B-linear map $\overline{f}$ is also represented by the matrix X over the bases $(\varphi(u_1), \ldots, \varphi(u_n))$ and $(\varphi(v_1), \ldots, \varphi(v_m))$.

Proposition 35.41 yields another proof of the fact that any two bases of a free A-module have the same cardinality. Indeed, if $\mathfrak{m}$ is a maximal ideal in the ring A, then we have the quotient ring homomorphism $\pi\colon A \to A/\mathfrak{m}$, and we get the $A/\mathfrak{m}$-module $\pi^*(M)$. If M is free, any basis $(u_1, \ldots, u_n)$ of M becomes the basis $(\varphi(u_1), \ldots, \varphi(u_n))$ of $\pi^*(M)$; but $A/\mathfrak{m}$ is a field, so the dimension n is uniquely determined. This argument also applies to an infinite basis $(u_i)_{i\in I}$. Observe that by Proposition 35.14, we have an isomorphism

$$\pi^*(M) = (A/\mathfrak{m}) \otimes_A M \approx M/\mathfrak{m}M,$$

so $M/\mathfrak{m}M$ is a vector space over the field $A/\mathfrak{m}$, which is the argument used in Theorem 35.1.

Proposition 35.42. *Given a ring homomomorphism $\rho\colon A \to B$, for any two A-modules M and N, there is a unique isomorphism*

$$\rho^*(M) \otimes_B \rho^*(N) \approx \rho^*(M \otimes_A N),$$

such that $(1 \otimes u) \otimes (1 \otimes v) \mapsto 1 \otimes (u \otimes v)$, for all $u \in M$ and all $v \in N$.

The proof uses identities from Proposition 33.13. It is not hard but it requires a little gymnastic; a good exercise for the reader.

Chapter 36

The Rational Canonical Form and Other Normal Forms

36.1 The Torsion Module Associated With An Endomorphism

We saw in Section 7.7 that given a linear map $f\colon E \to E$ from a K-vector space E into itself, we can define a scalar multiplication $\cdot\colon K[X] \times E \to E$ that makes E into a $K]X]$-module. If E is finite-dimensional, this $K[X]$-module denoted by E_f is a torsion module, and the main results of this chapter yield important direct sum decompositions of E into subspaces invariant under f.

Recall that given any polynomial $p(X) = a_0X^n + a_1X^{n-1} + \cdots + a_n$ with coefficients in the field K, we define the *linear map* $p(f)\colon E \to E$ by

$$p(f) = a_0 f^n + a_1 f^{n-1} + \cdots + a_n \mathrm{id},$$

where $f^k = f \circ \cdots \circ f$, the k-fold composition of f with itself. Note that

$$p(f)(u) = a_0 f^n(u) + a_1 f^{n-1}(u) + \cdots + a_n u,$$

for every vector $u \in E$. Then, we define the scalar multiplication $\cdot\colon K[X] \times E \to E$ by polynomials as follows: for every polynomial $p(X) \in K[X]$, for every $u \in E$,

$$p(X) \cdot u = p(f)(u).^1$$

It is easy to verify that this scalar multiplication satisfies the axioms M1, M2, M3, M4:

$$\begin{aligned} p \cdot (u + v) &= p \cdot u + p \cdot v \\ (p + q) \cdot u &= p \cdot u + q \cdot u \\ (pq) \cdot u &= p \cdot (q \cdot u) \\ 1 \cdot u &= u, \end{aligned}$$

[1]If necessary to avoid confusion, we use the notion $p(X) \cdot_f u$ instead of $p(X) \cdot u$.

for all $p, q \in K[X]$ and all $u, v \in E$. Thus, with this new scalar multiplication, E is a $K[X]$-module denoted by E_f.

If $p = \lambda$ is just a scalar in K (a polynomial of degree 0), then

$$\lambda \cdot u = (\lambda \mathrm{id})(u) = \lambda u,$$

which means that K acts on E by scalar multiplication as before. If $p(X) = X$ (the monomial X), then

$$X \cdot u = f(u).$$

Since K is a field, the ring $K[X]$ is a PID.

If E is finite-dimensional, say of dimension n, since K is a subring of $K[X]$ and since E is finitely generated over K, the $K[X]$-module E_f is finitely generated over $K[X]$. Furthermore, E_f is a torsion module. This follows from the Cayley-Hamilton Theorem (Theorem 7.15), but this can also be shown in an elementary fashion as follows. The space $\mathrm{Hom}(E, E)$ of linear maps of E into itself is a vector space of dimension n^2, therefore the n^2+1 linear maps

$$\mathrm{id}, f, f^2, \ldots, f^{n^2}$$

are linearly dependent, which yields a nonzero polynomial q such that $q(f) = 0$.

We can now translate notions defined for modules into notions for endomorphisms of vector spaces.

1. To say that U is a submodule of E_f means that U is a subspace of E invariant under f; that is, $f(U) \subseteq U$.

2. To say that V is a cyclic submodule of E_f means that there is some vector $u \in V$, such that V is spanned by $(u, f(u), \ldots, f^k(u), \ldots)$. If E has finite dimension n, then V is spanned by $(u, f(u), \ldots, f^k(u))$ for some $k \leq n-1$. We say that V is a *cyclic subspace for f with generator u*. Sometimes, V is denoted by $Z(u; f)$.

3. To say that the ideal $\mathfrak{a} = (p(X))$ (with $p(X)$ a monic polynomial) is the annihilator of the submodule V means that $p(f)(u) = 0$ for all $u \in V$, and we call p the *minimal polynomial* of V.

4. Suppose E_f is cyclic and let $\mathfrak{a} = (q)$ be its annihilator, where

$$q(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0.$$

Then, there is some vector u such that $(u, f(u), \ldots, f^k(u))$ span E_f, and because q is the minimal polynomial of E_f, we must have $k = n-1$. The fact that $q(f) = 0$ implies that

$$f^n(u) = -a_0 u - a_1 f(u) - \cdots - a_{n-1} f^{n-1}(u),$$

and so f is represented by the following matrix known as the *companion matrix* of $q(X)$:

$$U = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & -a_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}.$$

It is an easy exercise to prove that the characteristic polynomial $\chi_U(X)$ of U gives back $q(X)$:

$$\chi_U(X) = q(X).$$

We will need the following proposition to characterize when two linear maps are similar.

Proposition 36.1. *Let $f\colon E \to E$ and $f'\colon E' \to E'$ be two linear maps over the vector spaces E and E'. A linear map $g\colon E \to E'$ can be viewed as a linear map between the $K[X]$-modules E_f and $E_{f'}$ iff*

$$g \circ f = f' \circ g.$$

Proof. First, suppose g is $K[X]$-linear. Then, we have

$$g(p \cdot_f u) = p \cdot_{f'} g(u)$$

for all $p \in K[X]$ and all $u \in E$, so for $p = X$ we get

$$g(p \cdot_f u) = g(X \cdot_f u) = g(f(u))$$

and

$$p \cdot_{f'} g(u) = X \cdot_{f'} g(u) = f'(g(u)),$$

which means that $g \circ f = f' \circ g$.

Conversely, if $g \circ f = f' \circ g$, we prove by induction that

$$g \circ f^n = f'^n \circ g, \quad \text{for all } n \geq 1.$$

Indeed, we have

$$\begin{aligned} g \circ f^{n+1} &= g \circ f^n \circ f \\ &= f'^n \circ g \circ f \\ &= f'^n \circ f' \circ g \\ &= f'^{n+1} \circ g, \end{aligned}$$

establishing the induction step. It follows that for any polynomial $p(X) = \sum_{k=0}^n a_k X^k$, we have

$$\begin{aligned} g(p(X) \cdot_f u) &= g\Big(\sum_{k=0}^n a_k f^k(u)\Big) \\ &= \sum_{k=0}^n a_k g \circ f^k(u) \\ &= \sum_{k=0}^n a_k f'^k \circ g(u) \\ &= \Big(\sum_{k=0}^n a_k f'^k\Big)(g(u)) \\ &= p(X) \cdot_{f'} g(u), \end{aligned}$$

so, g is indeed $K[X]$-linear. $\square$

Definition 36.1. We say that the linear maps $f\colon E \to E$ and $f'\colon E' \to E'$ are *similar* iff there is an isomorphism $g\colon E \to E'$ such that

$$f' = g \circ f \circ g^{-1},$$

or equivalently,

$$g \circ f = f' \circ g.$$

Then, Proposition 36.1 shows the following fact:

Proposition 36.2. *With notation of Proposition 36.1, two linear maps f and f' are similar iff g is an isomorphism between E_f and $E'_{f'}$.*

Later on, we will see that the isomorphism of finitely generated torsion modules can be characterized in terms of invariant factors, and this will be translated into a characterization of similarity of linear maps in terms of so-called similarity invariants. If f and f' are represented by matrices A and A' over bases of E and E', then f and f' are similar iff the matrices A and A' are similar (there is an invertible matrix P such that $A' = PAP^{-1}$). Similar matrices (and endomorphisms) have the same characteristic polynomial.

It turns out that there is a useful relationship between E_f and the module $K[X] \otimes_K E$. Observe that the map $\cdot\colon K[X] \times E \to E$ given by

$$p \cdot u = p(f)(u)$$

is K-bilinear, so it yields a K-linear map $\sigma\colon K[X] \otimes_K E \to E$ such that

$$\sigma(p \otimes u) = p \cdot u = p(f)(u).$$

We know from Section 35.6 that $K[X] \otimes_K E$ is a $K[X]$-module (obtained from the inclusion $K \subseteq K[X]$), which we will denote by $E[X]$. Since E is a vector space, $E[X]$ is a free $K[X]$-module, and if $(u_1, \ldots, u_n)$ is a basis of E, then $(1 \otimes u_1, \ldots, 1 \otimes u_n)$ is a basis of $E[X]$.

The free $K[X]$-module $E[X]$ is not as complicated as it looks. Over the basis $(1 \otimes u_1, \ldots, 1 \otimes u_n)$, every element $z \in E[X]$ can be written uniquely as

$$z = p_1(1 \otimes u_1) + \cdots + p_n(1 \otimes u_n) = p_1 \otimes u_1 + \cdots + p_n \otimes u_n,$$

where $p_1, \ldots, p_n$ are polynomials in $K[X]$. For notational simplicity, we may write

$$z = p_1 u_1 + \cdots + p_n u_n,$$

where $p_1, \ldots, p_n$ are viewed as coefficients in $K[X]$. With this notation, we see that $E[X]$ is isomorphic to $(K[X])^n$, which is easy to understand.

Observe that σ is $K[X]$-linear, because

$$\begin{aligned} \sigma(q(p \otimes u)) &= \sigma((qp) \otimes u) \\ &= (qp) \cdot u \\ &= q(f)(p(f)(u)) \\ &= q \cdot (p(f)(u)) \\ &= q \cdot \sigma(p \otimes u). \end{aligned}$$

Therefore, σ is a linear map of $K[X]$-modules, $\sigma \colon E[X] \to E_f$. Using our simplified notation, if $z = p_1 u_1 + \cdots + p_n u_n \in E[X]$, then

$$\sigma(z) = p_1(f)(u_1) + \cdots + p_n(f)(u_n),$$

which amounts to plugging f for X and evaluating. Similarly, f is a $K[X]$-linear map of E_f, because

$$\begin{aligned} f(p \cdot u) &= f(p(f)(u)) \\ &= (fp(f))(u) \\ &= p(f)(f(u)) \\ &= p \cdot f(u), \end{aligned}$$

where we used the fact that $fp(f) = p(f)f$ because $p(f)$ is a polynomial in f. By Proposition 35.40, the linear map $f \colon E \to E$ induces a $K[X]$-linear map $\overline{f} \colon E[X] \to E[X]$ such that

$$\overline{f}(p \otimes u) = p \otimes f(u).$$

Observe that we have

$$f(\sigma(p \otimes u)) = f(p(f)(u)) = p(f)(f(u))$$

and

$$\sigma(\overline{f}(p \otimes u)) = \sigma(p \otimes f(u)) = p(f)(f(u)),$$

so we get

$$\sigma \circ \overline{f} = f \circ \sigma. \tag{$*$}$$

Using our simplified notation,

$$\overline{f}(p_1u_1 + \cdots + p_nu_n) = p_1f(u_1) + \cdots + p_nf(u_n).$$

Define the $K[X]$-linear map $\psi\colon E[X] \to E[X]$ by

$$\psi(p \otimes u) = (Xp) \otimes u - p \otimes f(u).$$

Observe that $\psi = X1_{E[X]} - \overline{f}$, which we abbreviate as $X1 - \overline{f}$. Using our simplified notation

$$\psi(p_1u_1 + \cdots + p_nu_n) = Xp_1u_1 + \cdots + Xp_nu_n - (p_1f(u_1) + \cdots + p_nf(u_n)).$$

It should be noted that everything we did in Section 36.1 applies to modules over a commutative ring A, except for the statements that assume that $A[X]$ is a PID. So, if M is an A-module, we can define the $A[X]$-modules M_f and $M[X] = A[X] \otimes_A M$, except that M_f is generally not a torsion module, and all the results showed above hold. Then, we have the following remarkable result.

Theorem 36.3. *(The Characteristic Sequence) Let A be a ring and let E be an A-module. The following sequence of $A[X]$-linear maps is exact:*

$$0 \longrightarrow E[X] \xrightarrow{\psi} E[X] \xrightarrow{\sigma} E_f \longrightarrow 0.$$

This means that ψ is injective, σ is surjective, and that $\mathrm{Im}(\psi) = \mathrm{Ker}\,(\sigma)$. *As a consequence, E_f is isomorphic to the quotient of $E[X]$ by* $\mathrm{Im}(X1 - \overline{f})$.

Proof. Because $\sigma(1 \otimes u) = u$ for all $u \in E$, the map σ is surjective. We have

$$\begin{aligned}\sigma(X(p \otimes u)) &= X \cdot \sigma(p \otimes u)\\ &= f(\sigma(p \otimes u)),\end{aligned}$$

which shows that

$$\sigma \circ X1 = f \circ \sigma = \sigma \circ \overline{f},$$

using $(*)$. This implies that

$$\begin{aligned}\sigma \circ \psi &= \sigma \circ (X1 - \overline{f})\\ &= \sigma \circ X1 - \sigma \circ \overline{f}\\ &= \sigma \circ \overline{f} - \sigma \circ \overline{f} = 0,\end{aligned}$$

and thus, $\mathrm{Im}(\psi) \subseteq \mathrm{Ker}\,(\sigma)$. It remains to prove that $\mathrm{Ker}\,(\sigma) \subseteq \mathrm{Im}(\psi)$.

Since the monomials X^k form a basis of $A[X]$, by Proposition 35.13 (with the roles of M and N exchanged), every $z \in E[X] = A[X] \otimes_A E$ has a unique expression as

$$z = \sum_k X^k \otimes u_k,$$

for a family (u_k) of finite support of $u_k \in E$. If $z \in \mathrm{Ker}\,(\sigma)$, then

$$0 = \sigma(z) = \sum_k f^k(u_k),$$

which allows us to write

$$\begin{aligned} z &= \sum_k X^k \otimes u_k - 1 \otimes 0 \\ &= \sum_k X^k \otimes u_k - 1 \otimes \left(\sum_k f^k(u_k) \right) \\ &= \sum_k (X^k \otimes u_k - 1 \otimes f^k(u_k)) \\ &= \sum_k (X^k(1 \otimes u_k) - \overline{f}^k(1 \otimes u_k)) \\ &= \sum_k (X^k 1 - \overline{f}^k)(1 \otimes u_k). \end{aligned}$$

Now, $X1$ and $\overline{f}$ commute, since

$$\begin{aligned} (X1 \circ \overline{f})(p \otimes u) &= (X1)(p \otimes f(u)) \\ &= (Xp) \otimes f(u) \end{aligned}$$

and

$$\begin{aligned} (\overline{f} \circ X1)(p \otimes u) &= \overline{f}((Xp) \otimes u) \\ &= (Xp) \otimes f(u), \end{aligned}$$

so we can write

$$X^k 1 - \overline{f}^k = (X1 - \overline{f}) \left(\sum_{j=0}^{k-1} (X1)^j \overline{f}^{k-j-1} \right),$$

and

$$z = (X1 - \overline{f}) \left(\sum_k \left(\sum_{j=0}^{k-1} (X1)^j \overline{f}^{k-j-1} \right) (1 \otimes u_k) \right),$$

which shows that $z = \psi(y)$ for some $y \in E[X]$.

Finally, we prove that ψ is injective as follows. We have

$$\begin{aligned}
\psi(z) &= \psi\biggl(\sum_k X^k \otimes u_k\biggr) \\
&= (X1 - \overline{f})\biggl(\sum_k X^k \otimes u_k\biggr) \\
&= \sum_k X^{k+1} \otimes (u_k - f(u_{k+1})),
\end{aligned}$$

where (u_k) is a family of finite support of $u_k \in E$. If $\psi(z) = 0$, then

$$\sum_k X^{k+1} \otimes (u_k - f(u_{k+1})) = 0,$$

and because the X^k form a basis of $A[X]$, we must have

$$u_k - f(u_{k+1}) = 0, \quad \text{for all } k.$$

Since (u_k) has finite support, there is a largest k, say $m+1$ so that $u_{m+1} = 0$, and then from

$$u_k = f(u_{k+1}),$$

we deduce that $u_k = 0$ for all k. Therefore, $z = 0$, and ψ is injective. $\square$

Remark: The exact sequence of Theorem 36.3 yields a *presentation* of M_f.

Since $A[X]$ is a free A-module, $A[X] \otimes_A M$ is a free A-module, but $A[X] \otimes_A M$ is generally not a free $A[X]$-module. However, if M is a free module, then $M[X]$ is a free $A[X]$-module, since if $(u_i)_{i \in I}$ is a basis for M, then $(1 \otimes u_i)_{i \in I}$ is a basis for $M[X]$. This allows us to define the characterisctic polynomial $\chi_f(X)$ of an endomorphism of a free module M as

$$\chi_f(X) = \det(X1 - \overline{f}).$$

Note that to have a correct definition, we need to define the determinant of a linear map allowing the indeterminate X as a scalar, and this is what the definition of $M[X]$ achieves (among other things). Theorem 36.3 can be used to give a short proof of the Cayley-Hamilton Theorem, see Bourbaki [25] (Chapter III, Section 8, Proposition 20). Proposition 7.10 is still the crucial ingredient of the proof.

36.2 The Rational Canonical Form

Let E be a finite-dimensional vector space over a field K, and let $f\colon E \to E$ be an endomorphism of E. We know from Section 36.1 that there is a $K[X]$-module E_f associated with f, and that E_f is a finitely generated torsion module over the PID $K[X]$. In this chapter, we show how Theorems from Sections 35.4 and 35.5 yield important results about the structure of the linear map f.

Recall that the annihilator of a subspace V is an ideal (p) uniquely defined by a monic polynomial p called the *minimal polynomial* of V.

Our first result is obtained by translating the primary decomposition theorem, Theorem 35.19. It is not too surprising that we obtain again Theorem 31.10!

Theorem 36.4. *(Primary Decomposition Theorem) Let $f\colon E \to E$ be a linear map on the finite-dimensional vector space E over the field K. Write the minimal polynomial m of f as*

$$m = p_1^{r_1} \cdots p_k^{r_k},$$

where the p_i are distinct irreducible monic polynomials over K, and the r_i are positive integers. Let

$$W_i = \mathrm{Ker}\,(p_i(f)^{r_i}), \quad i = 1, \ldots, k.$$

Then

(a) $E = W_1 \oplus \cdots \oplus W_k$.

(b) Each W_i is invariant under f and the projection from W onto W_i is given by a polynomial in f.

(c) The minimal polynomial of the restriction $f \mid W_i$ of f to W_i is $p_i^{r_i}$.

Example 36.1. Let $f\colon \mathbb{R}^4 \to \mathbb{R}^4$ be defined as $f(x, y, z, w) = (x + w, y + z, y + z, x + w)$. In terms of the standard basis, f has the matrix representation

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

A basic calculation shows that $\chi_f(X) = X^2(X-2)^2$ and that $m_f(X) = X(X-2)$. The primary decomposition theorem implies that

$$\mathbb{R}^4 = W_1 \oplus W_2, \qquad W_1 = \mathrm{Ker}\,(M), \qquad W_2 = \mathrm{Ker}\,(M - 2I).$$

Note that $\mathrm{Ker}\,(M)$ corresponds to the eigenspace associated with eigenvalue 0 and has basis $([-1, 0, 0, 1], [0, -1, 1, 0])$, while $\mathrm{Ker}\,(M - 2I)$ corresponds to the eigenspace associated with eigenvalue 2 and has basis $([1, 0, 0, 1], [0, 1, 1, 0])$.

Next we apply the Invariant Factors Decomposition Theorem, Theorem 35.31, to E_f. This theorem says that E_f is isomorphic to a direct sum

$$E_f \approx K[X]/(p_1) \oplus \cdots \oplus K[X]/(p_m)$$

of $m \leq n$ cyclic modules, where the p_j are uniquely determined monic polynomials of degree at least 1, such that

$$p_m \mid p_{m-1} \mid \cdots \mid p_1.$$

Each cyclic module $K[X]/(p_i)$ is isomorphic to a cyclic subspace for f, say V_i, whose minimal polynomial is p_i.

It is customary to renumber the polynomials p_i as follows. The n polynomials $q_1, \ldots, q_n$ are defined by:

$$q_j(X) = \begin{cases} 1 & \text{if } 1 \leq j \leq n-m \\ p_{n-j+1}(X) & \text{if } n-m+1 \leq j \leq n. \end{cases}$$

Then we see that

$$q_1 \mid q_2 \mid \cdots \mid q_n,$$

where the first $n-m$ polynomials are equal to 1, and we have the direct sum

$$E = E_1 \oplus \cdots \oplus E_n,$$

where E_i is a cyclic subspace for f whose minimal polynomial is q_i. In particular, $E_i = (0)$ for $i = 1, \ldots, n-m$. Theorem 35.31 also says that the minimal polynomial of f is $q_n = p_1$. We sum all this up in the following theorem.

Theorem 36.5. *(Cyclic Decomposition Theorem, First Version) Let $f \colon E \to E$ be an endomorphism on a K-vector space of dimension n. There exist n monic polynomials $q_1, \ldots, q_n \in K[X]$ such that*

$$q_1 \mid q_2 \mid \cdots \mid q_n,$$

and E is the direct sum

$$E = E_1 \oplus \cdots \oplus E_n$$

of cyclic subspaces $E_i = Z(u_i; f)$ for f, such that the minimal polynomial of the restriction of f to E_i is q_i. The polynomials q_i satisfying the above conditions are unique, and q_n is the minimal polynomial of f.

In view of translation point (4) at the beginning of Section 36.1, we know that over the basis

$$(u_i, f(u_i), \ldots, f^{n_i-1}(u_i))$$

of the cyclic subspace $E_i = Z(u_i; f)$, with $n_i = \deg(q_i)$, the matrix of the restriction of f to E_i is the *companion matrix* of $p_i(X)$, of the form

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & -a_{n_i-2} \\ 0 & 0 & 0 & \cdots & 1 & -a_{n_i-1} \end{pmatrix}.$$

If we put all these bases together, we obtain a block matrix whose blocks are of the above form. Therefore, we proved the following result.

Theorem 36.6. *(Rational Canonical Form, First Version) Let $f\colon E \to E$ be an endomorphism on a K-vector space of dimension n. There exist n monic polynomials $q_1, \ldots, q_n \in K[X]$ such that*

$$q_1 \mid q_2 \mid \cdots \mid q_n,$$

with $q_1 = \cdots = q_{n-m} = 1$, and a basis of E such that the matrix M of f is a block matrix of the form

$$M = \begin{pmatrix} M_{n-m+1} & 0 & \cdots & 0 & 0 \\ 0 & M_{n-m+2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & M_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & M_n \end{pmatrix},$$

where each M_i is the companion matrix of q_i. The polynomials q_i satisfying the above conditions are unique, and q_n is the minimal polynomial of f.

Definition 36.2. A matrix M as in Theorem 36.6 is called a matrix in *rational form*. The polynomials $q_1, \ldots, q_n$ arising in Theorems 36.5 and 36.6 are called the *similarity invariants* (or *invariant factors*) of f.

Theorem 36.6 shows that every matrix is similar to a matrix in rational form. Such a matrix is unique.

Example 1 continued: We will calculate the rational canonical form for $f(x, y, z, w) = (x + w, y + z, y + z, x + w)$. The difficulty in finding the rational canonical form lies in determining the invariant factors q_1, q_2, q_3, q_4. As we will shortly discover, the invariant factors of f correspond to the invariant factors of $XI - M$. See Propositions 36.8 and 36.11. The invariant factors of $XI - M$ are found by converting $XI - M$ to Smith normal form. Section 36.5 describes an algorithmic procedure for computing the Smith normal form of a matrix. By applying the methodology of Section 36.5, we find that Smith normal form for

$XI - M$ is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & X(X-2) & 0 \\ 0 & 0 & 0 & X(X-2) \end{pmatrix}.$$

Thus the invariant factors of f are $q_1 = 1 = q_2$, $q_3 = X(X-2) = q_4$, and Theorem 36.5 implies that

$$\mathbb{R}^4 = E_1 \oplus E_2,$$

where $E_1 = Z(u_1, f) \cong \mathbb{R}[X]/(X(X-2))$ and $E_2 = Z(u_2, f) \cong \mathbb{R}[X]/(X(X-2))$. The subspace E_1 has basis (u_1, Mu_1) where $u_1 = (1, 0, 1, 0)$ and $Mu_1 = (1, 1, 1, 1)$, while the subspace E_2 has basis (u_2, Mu_2) where $u_2 = (0, 0, 1, 0)$ and $Mu_2 = (0, 1, 1, 0)$. Theorem 36.6 implies that rational canonical form of $M(f)$ is

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}.$$

By Proposition 36.2, two linear maps f and f' are similar iff there is an isomorphism between E_f and $E'_{f'}$, and thus by the uniqueness part of Theorem 35.31, iff they have the same similarity invariants $q_1, \ldots, q_n$.

Proposition 36.7. *If E and E' are two finite-dimensional vector spaces and if $f \colon E \to E$ and $f' \colon E' \to E'$ are two linear maps, then f and f' are similar iff they have the same similarity invariants.*

The effect of extending the field K to a field L is the object of the next proposition.

Proposition 36.8. *Let $f \colon E \to E$ be a linear map on a K-vector space E, and let $(q_1, \ldots, q_n)$ be the similarity invariants of f. If L is a field extension of K (which means that $K \subseteq L$), and if $E_{(L)} = L \otimes_K E$ is the vector space obtained by extending the scalars, and $f_{(L)} = 1_L \otimes f$ the linear map of $E_{(L)}$ induced by f, then the similarity invariants of $f_{(L)}$ are $(q_1, \ldots, q_n)$ viewed as polynomials in $L[X]$.*

Proof. We know that E_f is isomorphic to the direct sum

$$E_f \approx K[X]/(q_1 K[X]) \oplus \cdots \oplus K[X]/(q_n K[X]),$$

so by tensoring with $L[X]$ and using Propositions 35.12 and 33.13, we get

$$\begin{aligned} L[X] \otimes_{K[X]} E_f &\approx L[X] \otimes_{K[X]} (K[X]/(q_1 K[X]) \oplus \cdots \oplus K[X]/(q_n K[X])) \\ &\approx L[X] \otimes_{K[X]} (K[X]/(q_1 K[X])) \oplus \cdots \oplus L[X] \otimes_{K[X]} (K[X]/(q_n K[X])) \\ &\approx (K[X]/(q_1 K[X])) \otimes_{K[X]} L[X] \oplus \cdots \oplus (K[X]/(q_n K[X])) \otimes_{K[X]} L[X]. \end{aligned}$$

However, by Proposition 35.14, we have isomorphisms

$$(K[X]/(q_iK[X])) \otimes_{K[X]} L[X] \approx L[X]/(q_iL[X]),$$

so we get

$$L[X] \otimes_{K[X]} E_f \approx L[X]/(q_1L[X]) \oplus \cdots \oplus L[X]/(q_nL[X]).$$

Since E_f is a $K[X]$-module, the $L[X]$ module $L[X] \otimes_{K[X]} E_f$ is the module obtained from E_f by the ring extension $K[X] \subseteq L[X]$. The L-module $E_{(L)} = L \otimes_K E$ becomes the $L[X]$-module $E_{(L)f_{(L)}}$ where

$$f_{(L)} = \mathrm{id}_L \otimes_K f.$$

We have the following proposition

Proposition 36.9. *For any field extension $K \subseteq L$, and any linear map $f \colon E \to E$ where E is a K-vector space, there is an isomorphism between the $L[X]$-modules $L[X] \otimes_{K[X]} E_f$ and $E_{(L)f_{(L)}}$.*

Proof. First we define the map $\alpha \colon L \times E \to L[X] \otimes_{K[X]} E_f$ by

$$\alpha(\lambda, u) = \lambda \otimes_{K[X]} u.$$

It is immediately verified that α is K-bilinear, so we obtain a K-linear map $\widetilde{\alpha} \colon L \otimes_K E \to L[X] \otimes_{K[X]} E_f$. Now $E_{(L)} = L \otimes_K E$ is a $L[X]$-module $(L \otimes_K E)_{f(L)}$, and let us denote this scalar multiplication by $\odot$. To describe $\odot$ it is enough to define how a monomial $aX^k \in L[X]$ acts on a generator $(\lambda \otimes_K u) \in L \otimes_K E$. We have

$$\begin{aligned} aX^k \odot (\lambda \otimes_K u) &= a(\mathrm{id}_L \otimes_K f)^k (\lambda \otimes_K u) \\ &= a(\lambda \otimes_K f^k(u)) \\ &= a\lambda \otimes_K f^k(u). \end{aligned}$$

We claim that $\widetilde{\alpha}$ is actually $L[X]$-linear. Indeed, we have

$$\begin{aligned} \widetilde{\alpha}(aX^k \odot (\lambda \otimes_K u)) &= \widetilde{\alpha}(a\lambda \otimes_K f^k(u)) \\ &= a\lambda \otimes_{K[X]} f^k(u), \end{aligned}$$

and by the definition of scalar multiplication in the $K[X]$-module E_f, we have $f^k(u) = X^k \cdot_f u$, so we have

$$\begin{aligned} \widetilde{\alpha}(aX^k \odot (\lambda \otimes_K u)) &= a\lambda \otimes_{K[X]} f^k(u) \\ &= a\lambda \otimes_{K[X]} X^k \cdot_f u \\ &= X^k \cdot (a\lambda \otimes_{K[X]} u) \\ &= aX^k \cdot (\lambda \otimes_{K[X]} u), \end{aligned}$$

which shows that $\widetilde{\alpha}$ is $L[X]$-linear.

We also define the map $\beta\colon L[X] \times E_f \to (L \otimes_K E)_{f_{(L)}}$ by

$$\beta(q(X), u) = q(X) \odot (1 \otimes_K u).$$

Using a computation similar to the computation that we just performed, we can check that β is $K[X]$-bilinear so we obtain a map $\widetilde{\beta}\colon L[X] \otimes_{K[X]} E_f \to (L \otimes_K E)_{f_{(L)}}$. To finish the proof, it suffices to prove that $\widetilde{\alpha} \circ \widetilde{\beta}$ and $\widetilde{\beta} \circ \widetilde{\alpha}$ are the identity on generators. We have

$$\widetilde{\alpha} \circ \widetilde{\beta}(q(X) \otimes_{K[X]} u) = \widetilde{\alpha}(q(X) \odot (1 \otimes_K u)) = q(X) \cdot (1 \otimes_{K[X]} u)) = q(X) \otimes_{K[X]} u,$$

and

$$\widetilde{\beta} \circ \widetilde{\alpha}(\lambda \otimes_K u) = \widetilde{\beta}(\lambda \otimes_{K[X]} u) = \lambda \odot (1 \otimes_K u) = \lambda \otimes_K u,$$

which finishes the proof. $\square$

By Proposition 36.9,

$$E_{(L)f_{(L)}} \approx L[X] \otimes_{K[X]} E_f \approx L[X]/(q_1 L[X]) \oplus \cdots \oplus L[X]/(q_n L[X]),$$

which shows that $(q_1, \ldots, q_n)$ are the similarity invariants of $f_{(L)}$. $\square$

Proposition 36.8 justifies the terminology "invariant" in similarity invariants. Indeed, under a field extension $K \subseteq L$, the similarity invariants of $f_{(L)}$ remain the same. This is not true of the elementary divisors, which depend on the field; indeed, an irreducible polynomial $p \in K[X]$ may split over $L[X]$. Since q_n is the minimal polynomial of f, the above reasoning also shows that the minimal polynomial of $f_{(L)}$ remains the same under a field extension.

Proposition 36.8 has the following corollary.

Proposition 36.10. *Let K be a field and let $L \supseteq K$ be a field extension of K. For any two square matrices A and B over K, if there is an invertible matrix Q over L such that $B = QAQ^{-1}$, then there is an invertible matrix P over K such that $B = PAP^{-1}$.*

Recall from Theorem 36.3 that the sequence of $K[X]$-linear maps

$$0 \longrightarrow E[X] \xrightarrow{\psi} E[X] \xrightarrow{\sigma} E_f \longrightarrow 0$$

is exact, and as a consequence, E_f is isomorphic to the quotient of $E[X]$ by $\mathrm{Im}(X1 - \overline{f})$. Furthermore, because E is a vector space, $E[X]$ is a free module with basis $(1 \otimes u_1, \ldots, 1 \otimes u_n)$, where $(u_1, \ldots, u_n)$ is a basis of E, and since ψ is injective, the module $\mathrm{Im}(X1 - \overline{f})$ has rank n. By Theorem 35.31, we have an isomorphism

$$E_f \approx K[X]/(q_1 K[X]) \oplus \cdots \oplus K[X]/(q_n K[X]),$$

and by Proposition 35.32, $E[X]/\mathrm{Im}(X1 - \overline{f})$ is isomorphic to a direct sum

$$E[X]/\mathrm{Im}(X1 - \overline{f}) \approx K[X]/(p_1 K[X]) \oplus \cdots \oplus K[X]/(p_n K[X]),$$

where $p_1, \ldots, p_n$ are the invariant factors of $\mathrm{Im}(X1 - \overline{f})$ with respect to $E[X]$. Since $E_f \approx E[X]/\mathrm{Im}(X1 - \overline{f})$, by the uniqueness part of Theorem 35.31 and because the polynomials are monic, we must have $p_i = q_i$, for $i = 1, \ldots, n$. Therefore, we proved the following crucial fact:

Proposition 36.11. *For any linear map $f \colon E \to E$ over a K-vector space E of dimension n, the similarity invariants of f are equal to the invariant factors of $\mathrm{Im}(X1 - \overline{f})$ with respect to $E[X]$.*

Proposition 36.11 is the key to computing the similarity invariants of a linear map. This can be done using a procedure to convert $XI - M$ to its *Smith normal form*. Propositions 36.11 and 35.37 yield the following result.

Proposition 36.12. *For any linear map $f \colon E \to E$ over a K-vector space E of dimension n, if $(q_1, \ldots, q_n)$ are the similarity invariants of f, for any matrix M representing f with respect to any basis, then for $k = 1, \ldots, n$ the product*

$$d_k(X) = q_1(X) \cdots q_k(X)$$

is the gcd of the $k \times k$-minors of the matrix $XI - M$.

Note that the matrix $XI - M$ is none other than the matrix that yields the characteristic polynomial $\chi_f(X) = \det(XI - M)$ of f.

Proposition 36.13. *For any linear map $f \colon E \to E$ over a K-vector space E of dimension n, if $(q_1, \ldots, q_n)$ are the similarity invariants of f, then the following properties hold:*

(1) If $\chi_f(X)$ is the characteristic polynomial of f, then

$$\chi_f(X) = q_1(X) \cdots q_n(X).$$

(2) The minimal polynomial $m(X) = q_n(X)$ of f divides the characteristic polynomial $\chi_f(X)$ of f.

(3) The characteristic polynomial $\chi_f(X)$ divides $m(X)^n$.

(4) E is cyclic for f iff $m(X) = \chi_f(X)$.

Proof. Property (1) follows from Proposition 36.12 for $k = n$. It also follows from Theorem 36.6 and the fact that for the companion matrix associated with q_i, the characteristic polynomial of this matrix is also q_i. Property (2) is obvious from (1). Since each q_i divides q_{i+1}, each q_i divides q_n, so their product $\chi_f(X)$ divides $m(X)^n = q_n(X)^n$. The last condition says that $q_1 = \cdots = q_{n-1} = 1$, which means that E_f has a single summand. $\square$

Observe that Proposition 36.13 yields another proof of the Cayley–Hamilton Theorem. It also implies that a linear map is nilpotent iff its characteristic polynomial is X^n.

36.3 The Rational Canonical Form, Second Version

Let us now translate the Elementary Divisors Decomposition Theorem, Theorem 35.38, in terms of E_f. We obtain the following result.

Theorem 36.14. *(Cyclic Decomposition Theorem, Second Version) Let $f\colon E \to E$ be an endomorphism on a K-vector space of dimension n. Then, E is the direct sum of of cyclic subspaces $E_j = Z(u_j; f)$ for f, such that the minimal polynomial of E_j is of the form $p_i^{n_{i,j}}$, for some irreducible monic polynomials $p_1, \ldots, p_t \in K[X]$ and some positive integers $n_{i,j}$, such that for each $i = 1, \ldots, t$, there is a sequence of integers*

$$1 \leq \underbrace{n_{i,1}, \ldots, n_{i,1}}_{m_{i,1}} < \underbrace{n_{i,2}, \ldots, n_{i,2}}_{m_{i,2}} < \cdots < \underbrace{n_{i,s_i}, \ldots, n_{i,s_i}}_{m_{i,s_i}},$$

with $s_i \geq 1$, and where $n_{i,j}$ occurs $m_{i,j} \geq 1$ times, for $j = 1, \ldots, s_i$. Furthermore, the monic polynomials p_i and the integers $r, t, n_{i,j}, s_i, m_{i,j}$ are uniquely determined.

Note that there are $\mu = \sum m_{i,j}$ cyclic subspaces $Z(u_j; f)$. Using bases for the cyclic subspaces $Z(u_j; f)$ as in Theorem 36.6, we get the following theorem.

Theorem 36.15. *(Rational Canonical Form, Second Version) Let $f\colon E \to E$ be an endomorphism on a K-vector space of dimension n. There exist t distinct irreducible monic polynomials $p_1, \ldots, p_t \in K[X]$ and some positive integers $n_{i,j}$, such that for each $i = 1, \ldots, t$, there is a sequence of integers*

$$1 \leq \underbrace{n_{i,1}, \ldots, n_{i,1}}_{m_{i,1}} < \underbrace{n_{i,2}, \ldots, n_{i,2}}_{m_{i,2}} < \cdots < \underbrace{n_{i,s_i}, \ldots, n_{i,s_i}}_{m_{i,s_i}},$$

with $s_i \geq 1$, and where $n_{i,j}$ occurs $m_{i,j} \geq 1$ times, for $j = 1, \ldots, s_i$, and there is a basis of E such that the matrix M of f is a block matrix of the form

$$M = \begin{pmatrix} M_1 & 0 & \cdots & 0 & 0 \\ 0 & M_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & M_{\mu-1} & 0 \\ 0 & 0 & \cdots & 0 & M_\mu \end{pmatrix},$$

where each M_j is the companion matrix of some $p_i^{n_{i,j}}$, and $\mu = \sum m_{i,j}$. The monic polynomials $p_1, \ldots, p_t$ and the integers $r, t, n_{i,j}, s_i, m_{i,j}$ are uniquely determined

The polynomials $p_i^{n_{i,j}}$ are called the *elementary divisors* of f (and M). These polynomials are factors of the minimal polynomial.

Example 1 continued: Recall that $f(x, y, z, w) = (x + w, y + z, y + z, x + w)$ has two nontrivial invariant factors $q_1 = x(x - 2) = q_2$. Thus the elementary factors of f are $p_1 = x = p_2$ and $p_3 = x - 2 = p_4$. Theorem 36.14 implies that

$$\mathbb{R}^4 = E_1 \oplus E_2 \oplus E_3 \oplus E_4,$$

where $E_1 = Z(u_1, f) \cong \mathbb{R}[X]/(X)$, $E_2 = Z(u_2, f) \cong \mathbb{R}[X]/(X)$, $E_3 = Z(u_3, f) \cong \mathbb{R}[X]/(X-2)$, and $E_4 = Z(u_4, f) \cong \mathbb{R}[X]/(X-2)$. The subspaces E_1 and E_2 correspond to one-dimensional spaces spanned by eigenvectors associated with eigenvalue 0, while E_3 and E_4 correspond to one-dimensional spaces spanned by eigenvectors associated with eigenvalue 2. If we let $u_1 = (-1, 0, 0, 1)$, $u_2 = (0, -1, 1, 0)$, $u_3 = (1, 0, 0, 1)$ and $u_4 = (0, 1, 1, 0)$, Theorem 36.15 gives

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

as the rational canonical form associated with the cyclic decomposition $\mathbb{R}^4 = E_1 \oplus E_2 \oplus E_3 \oplus E_4$.

As we pointed earlier, unlike the similarity invariants, the elementary divisors may change when we pass to a field extension.

We will now consider the special case where all the irreducible polynomials p_i are of the form $X - \lambda_i$; that is, when are the eigenvalues of f belong to K. In this case, we find again the Jordan form.

36.4 The Jordan Form Revisited

In this section, we assume that all the roots of the minimal polynomial of f belong to K. This will be the case if K is algebraically closed. The irreducible polynomials p_i of Theorem 36.14 are the polynomials $X - \lambda_i$, for the distinct eigenvalues λ_i of f. Then, each cyclic subspace $Z(u_j; f)$ has a minimal polynomial of the form $(X - \lambda)^m$, for some eigenvalue λ of f and some $m \geq 1$. It turns out that by choosing a suitable basis for the cyclic subspace $Z(u_j; f)$, the matrix of the restriction of f to $Z(u_j; f)$ is a Jordan block.

Proposition 36.16. *Let E be a finite-dimensional K-vector space and let $f\colon E \to E$ be a linear map. If E is a cyclic $K[X]$-module and if $(X - \lambda)^n$ is the minimal polynomial of f, then there is a basis of E of the form*

$$((f - \lambda \mathrm{id})^{n-1}(u), (f - \lambda \mathrm{id})^{n-2}(u), \ldots, (f - \lambda \mathrm{id})(u), u),$$

for some $u \in E$. With respect to this basis, the matrix of f is the Jordan block

$$J_n(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

Proof. Since E is a cyclic $K[X]$-module, there is some $u \in E$ so that E is generated by $u, f(u), f^2(u), \ldots$, which means that every vector in E is of the form $p(f)(u)$, for some polynomial, $p(X)$. We claim that $u, f(u), \ldots, f^{n-2}(u), f^{n-1}(u)$ generate E, which implies that the dimension of E is at most n.

This is because if $p(X)$ is any polynomial of degree at least n, then we can divide $p(X)$ by $(X - \lambda)^n$, obtaining

$$p = (X - \lambda)^n q + r,$$

where $0 \leq \deg(r) < n$, and as $(X - \lambda)^n$ annihilates E, we get

$$p(f)(u) = r(f)(u),$$

which means that every vector of the form $p(f)(u)$ with $p(X)$ of degree $\geq n$ is actually a linear combination of $u, f(u), \ldots, f^{n-2}(u), f^{n-1}(u)$.

We claim that the vectors

$$u, (f - \lambda \mathrm{id})(u), \ldots, (f - \lambda \mathrm{id})^{n-2}(u)(f - \lambda \mathrm{id})^{n-1}(u)$$

are linearly independent. Indeed, if we had a nontrivial linear combination

$$a_0(f - \lambda \mathrm{id})^{n-1}(u) + a_1(f - \lambda \mathrm{id})^{n-2}(u) + \cdots + a_{n-2}(f - \lambda \mathrm{id})(u) + a_{n-1}u = 0,$$

then the polynomial

$$a_0(X - \lambda)^{n-1} + a_1(X - \lambda)^{n-2} + \cdots + a_{n-2}(X - \lambda) + a_{n-1}$$

of degree at most $n - 1$ would annihilate E, contradicting the fact that $(X - \lambda)^n$ is the minimal polynomial of f (and thus, of smallest degree). Consequently, as the dimension of E is at most n,

$$((f - \lambda \mathrm{id})^{n-1}(u), (f - \lambda \mathrm{id})^{n-2}(u), \ldots, (f - \lambda \mathrm{id})(u), u),$$

is a basis of E and since $u, f(u), \ldots, f^{n-2}(u), f^{n-1}(u)$ span E,

$$(u, f(u), \ldots, f^{n-2}(u), f^{n-1}(u))$$

is also a basis of E.

Let us see how f acts on the basis

$$((f - \lambda \mathrm{id})^{n-1}(u), (f - \lambda \mathrm{id})^{n-2}(u), \ldots, (f - \lambda \mathrm{id})(u), u).$$

If we write $f = f - \lambda \mathrm{id} + \lambda \mathrm{id}$, as $(f - \lambda \mathrm{id})^n$ annihilates E, we get

$$f((f - \lambda \mathrm{id})^{n-1}(u)) = (f - \lambda \mathrm{id})^n(u) + \lambda(f - \lambda \mathrm{id})^{n-1}(u) = \lambda(f - \lambda \mathrm{id})^{n-1}(u)$$

and

$$f((f - \lambda \mathrm{id})^k(u)) = (f - \lambda \mathrm{id})^{k+1}(u) + \lambda(f - \lambda \mathrm{id})^k(u), \qquad 0 \leq k \leq n - 2.$$

But this means precisely that the matrix of f in this basis is the Jordan block $J_n(\lambda)$. $\square$

The basis

$$((f-\lambda \mathrm{id})^{n-1}(u), (f-\lambda \mathrm{id})^{n-2}(u), \ldots, (f-\lambda \mathrm{id})(u), u),$$

provided by Proposition 36.16 is known as a *Jordan chain.* Note that $(f-\lambda \mathrm{id})^{n-1}(u)$ is an eigenvector for f. To construct the Jordan chain, we must find u which is a generalized eigenvector of f. This is done by first finding an eigenvector x_1 of f and recursively solving the system $(f-\lambda \mathrm{id})x_{i+1} = x_i$ for $i \leq 1 \leq n-1$. For example suppose $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ where $f(x, y, z) = (x+y+z, y+z, z)$. In terms of the standard basis, the matrix representation for f is $M = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. By using M, it is readily verified that the minimal polynomial f equals the characteristic polynomial, namely $(X-1)^3$. Thus f has the eigenvalue $\lambda = 1$ with repeated three times. To find the eigenvector x_1 associated with $\lambda = 1$, we solve the system $(M-I)x_1 = 0$, or equivalently

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus $y = z = 0$ with $x = 1$ solves this system to provide the eigenvector $x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. We next solve the system $(M-I)x_2 = x_1$, namely

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

which implies that $z = 0$ and $y = 1$. Hence $x_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ will work. To finish constructing our Jordan chain, we must solve the system $(M-I)x_3 = x_2$, namely

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

from which we see that $z = 1$, $y = 0$, and $x_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$. By setting $x_3 = u$, we form the basis

$$((f-\lambda \mathrm{id})^2(u), (f-\lambda \mathrm{id})^1(u), \ldots, (f-\lambda \mathrm{id})(u), u) = (x_1, x_2, x_3).$$

In terms of the basis (x_1, x_2, x_3), the map $f(x, y, z) = (x + y + z, y + z, z)$ has the Jordan block matrix representation $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ since

$$\begin{aligned} f(x_1) &= f(1, 0, 0) = (1, 0, 0) = x_1 \\ f(x_2) &= f(1, 1, 0) = (2, 1, 0) = x_1 + x_2 \\ f(x_3) &= f(1, 0, 1) = (2, 1, 1) = x_2 + x_3. \end{aligned}$$

Combining Theorem 36.15 and Proposition 36.16, we obtain a strong version of the Jordan form.

Theorem 36.17. *(Jordan Canonical Form) Let E be finite-dimensional K-vector space. The following properties are equivalent:*

(1) The eigenvalues of f all belong to K.

(2) There is a basis of E in which the matrix of f is upper (or lower) triangular.

(3) There exist a basis of E in which the matrix A of f is Jordan matrix. Furthermore, the number of Jordan blocks $J_r(\lambda)$ appearing in A, for fixed r and λ, is uniquely determined by f.

Proof. The implication (1) $\Longrightarrow$ (3) follows from Theorem 36.15 and Proposition 36.16. The implications (3) $\Longrightarrow$ (2) and (2) $\Longrightarrow$ (1) are trivial. $\square$

Compared to Theorem 31.17, the new ingredient is the uniqueness assertion in (3), which is not so easy to prove.

Observe that the minimal polynomial of f is the least common multiple of the polynomials $(X - \lambda)^r$ associated with the Jordan blocks $J_r(\lambda)$ appearing in A, and the characteristic polynomial of A is the product of these polynomials.

We now return to the problem of computing effectively the similarity invariants of a matrix M. By Proposition 36.11, this is equivalent to computing the invariant factors of $XI - M$. In principle, this can be done using Proposition 35.35. A procedure to do this effectively for the ring $A = K[X]$ is to convert $XI - M$ to its Smith normal form. This will also yield the rational canonical form for M.

36.5 The Smith Normal Form

The Smith normal form is the special case of Proposition 35.35 applied to the PID $K[X]$ where K is a field, but it also says that the matrices P and Q are products of elementary matrices. It turns out that such a result holds for any Euclidean ring, and the proof is basically the same.

Recall from Definition 30.10 that a *Euclidean ring* is an integral domain A such that there exists a function $\sigma\colon A \to \mathbb{N}$ with the following property: For all $a, b \in A$ with $b \neq 0$, there are some $q, r \in A$ such that

$$a = bq + r \quad \text{and} \quad \sigma(r) < \sigma(b).$$

Note that the pair (q, r) is not necessarily unique.

We make use of the elementary row and column operations $P(i, k)$, $E_{i,j;\beta}$, and $E_{i,\lambda}$ described in Chapter 8, where we require the scalar λ used in $E_{i,\lambda}$ to be a unit.

Theorem 36.18. *If M is an $m \times n$ matrix over a Euclidean ring A, then there exist some invertible $n \times n$ matrix P and some invertible $m \times m$ matrix Q, where P and Q are products of elementary matrices, and a $m \times n$ matrix D of the form*

$$D = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \alpha_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

for some nonzero $\alpha_i \in A$, such that

(1) $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$, *and*

(2) $M = QDP^{-1}$.

Proof. We follow Jacobson's proof [98] (Chapter 3, Theorem 3.8). We proceed by induction on $m + n$.

If $m = n = 1$, let $P = (1)$ and $Q = (1)$.

For the induction step, if $M = 0$, let $P = I_n$ and $Q = I_m$. If $M \neq 0$, the stategy is to apply a sequence of elementary transformations that converts M to a matrix of the form

$$M' = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & Y & \\ 0 & & & \end{pmatrix}$$

where Y is a $(m-1) \times (n-1)$-matrix such that α_1 divides every entry in Y. Then, we proceed by induction on Y. To find M', we perform the following steps.

Step 1. Pick some nonzero entry a_{ij} in M such that $\sigma(a_{ij})$ is minimal. Then permute column j and column 1, and permute row i and row 1, to bring this entry in position $(1, 1)$. We denote this new matrix again by M.

Step 2a.

If $m = 1$ go to Step 2b.

If $m > 1$, then there are two possibilities:

(i) M is of the form

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

If $n = 1$, stop; else go to Step 2b.

(ii) There is some nonzero entry a_{i1} $(i > 1)$ below a_{11} in the first column.

(a) If there is some entry a_{k1} in the first column such that a_{11} does not divide a_{k1}, then pick such an entry (say, with the smallest index i such that $\sigma(a_{i1})$ is minimal), and divide a_{k1} by a_{11}; that is, find b_k and b_{k1} such that

$$a_{k1} = a_{11}b_k + b_{k1}, \quad \text{with} \quad \sigma(b_{k1}) < \sigma(a_{11}).$$

Subtract b_k times row 1 from row k and permute row k and row 1, to obtain a matrix of the form

$$M = \begin{pmatrix} b_{k1} & b_{k2} & \cdots & b_{kn} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Go back to Step 2a.

(b) If a_{11} divides every (nonzero) entry a_{i1} for $i \geq 2$, say $a_{i1} = a_{11}b_i$, then subtract b_i times row 1 from row i for $i = 2, \ldots, m$; go to Step 2b.

Observe that whenever we return to the beginning of Step 2a, we have $\sigma(b_{k1}) < \sigma(a_{11})$. Therefore, after a finite number of steps, we must exit Step 2a with a matrix in which all entries in column 1 but the first are zero and go to Step 2b.

Step 2b.

This step is reached only if $n > 1$ and if the only nonzero entry in the first column is a_{11}.

(a) If M is of the form

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

and $m = 1$ stop; else go to Step 3.

(b) If there is some entry a_{1k} in the first row such that a_{11} does not divide a_{1k}, then pick such an entry (say, with the smallest index j such that $\sigma(a_{1j})$ is minimal), and divide a_{1k} by a_{11}; that is, find b_k and b_{1k} such that

$$a_{1k} = a_{11}b_k + b_{1k}, \quad \text{with} \quad \sigma(b_{1k}) < \sigma(a_{11}).$$

Subtract b_k times column 1 from column k and permute column k and column 1, to obtain a matrix of the form

$$M = \begin{pmatrix} b_{1k} & a_{k2} & \cdots & a_{kn} \\ b_{2k} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{mk} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Go back to Step 2b.

(c) If a_{11} divides every (nonzero) entry a_{1j} for $j \geq 2$, say $a_{1j} = a_{11}b_j$, then subtract b_j times column 1 from column j for $j = 2, \ldots, n$; go to Step 3.

As in Step 2a, whenever we return to the beginning of Step 2b, we have $\sigma(b_{1k}) < \sigma(a_{11})$. Therefore, after a finite number of steps, we must exit Step 2b with a matrix in which all entries in row 1 but the first are zero.

Step 3. This step is reached only if the only nonzero entry in the first row is a_{11}.

(i) If

$$M = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & Y & \\ 0 & & & \end{pmatrix}$$

go to Step 4.

(ii) If Step 2b ruined column 1 which now contains some nonzero entry below a_{11}, go back to Step 2a.

We perform a sequence of alternating steps between Step 2a and Step 2b. Because the σ-value of the $(1,1)$-entry strictly decreases whenever we reenter Step 2a and Step 2b, such a sequence must terminate with a matrix of the form

$$M = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & Y & \\ 0 & & & \end{pmatrix}$$

Step 4. If a_{11} divides all entries in Y, stop.

Otherwise, there is some column, say j, such that a_{11} does not divide some entry a_{ij}, so add the jth column to the first column. This yields a matrix of the form

$$M = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ b_{2j} & & & \\ \vdots & & Y & \\ b_{mj} & & & \end{pmatrix}$$

where the ith entry in column 1 is nonzero, so go back to Step 2a,

Again, since the σ-value of the $(1,1)$-entry strictly decreases whenever we reenter Step 2a and Step 2b, such a sequence must terminate with a matrix of the form

$$M' = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & Y & \\ 0 & & & \end{pmatrix}$$

where α_1 divides every entry in Y. Then, we apply the induction hypothesis to Y. $\square$

If the PID A is the polynomial ring $K[X]$ where K is a field, the α_i are nonzero polynomials, so we can apply row operations to normalize their leading coefficients to be 1. We obtain the following theorem.

Theorem 36.19. *(Smith Normal Form) If M is an $m \times n$ matrix over the polynomial ring $K[X]$, where K is a field, then there exist some invertible $n \times n$ matrix P and some invertible $m \times m$ matrix Q, where P and Q are products of elementary matrices with entries in $K[X]$, and a $m \times n$ matrix D of the form*

$$D = \begin{pmatrix} q_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & q_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & q_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

for some nonzero monic polynomials $q_i \in k[X]$, such that

(1) $q_1 \mid q_2 \mid \cdots \mid q_r$, and

(2) $M = QDP^{-1}$.

In particular, if we apply Theorem 36.19 to a matrix M of the form $M = XI - A$, where A is a square matrix, then $\det(XI - A) = \chi_A(X)$ is never zero, and since $XI - A = QDP^{-1}$ with P, Q invertible, all the entries in D must be nonzero and we obtain the following result showing that the similarity invariants of A can be computed using elementary operations.

Theorem 36.20. *If A is an $n \times n$ matrix over the field K, then there exist some invertible $n \times n$ matrices P and Q, where P and Q are products of elementary matrices with entries in $K[X]$, and a $n \times n$ matrix D of the form*

$$D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & q_1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & q_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & q_m \end{pmatrix}$$

for some nonzero monic polynomials $q_i \in k[X]$ of degree ≥ 1, such that

(1) $q_1 \mid q_2 \mid \cdots \mid q_m$,

(2) $q_1, \ldots q_m$ are the similarity invariants of A, and

(3) $XI - A = QDP^{-1}$.

The matrix D in Theorem 36.20 is often called *Smith normal form* of A, even though this is confusing terminology since D is really the Smith normal form of $XI - A$.

Of course, we know from previous work that in Theorem 36.19, the $\alpha_1, \ldots, \alpha_r$ are unique, and that in Theorem 36.20, the $q_1, \ldots, q_m$ are unique. This can also be proved using some simple properties of minors, but we leave it as an exercise (for help, look at Jacobson [98], Chapter 3, Theorem 3.9).

The rational canonical form of A can also be obtained from Q^{-1} and D, but first, let us consider the generalization of Theorem 36.19 to PID's that are not necessarily Euclidean rings.

We need to find a "norm" that assigns a natural number $\sigma(a)$ to any nonzero element of a PID A, in such a way that $\sigma(a)$ decreases whenever we return to Step 2a and Step 2b. Since a PID is a UFD, we use the number

$$\sigma(a) = k_1 + \cdots + k_r$$

of prime factors in the factorization of a nonunit element

$$a = up_1^{k_1} \cdots p_r^{k_r},$$

and we set

$$\sigma(u) = 0$$

if u is a unit.

We can't divide anymore, but we can find gcd's and use Bezout to mimic division. The key ingredient is this: for any two nonzero elements $a, b \in A$, if a does not divide b then let $d \neq 0$ be a gcd of a and b. By Bezout, there exist $x, y \in A$ such that

$$ax + by = d.$$

We can also write $a = td$ and $b = -sd$, for some $s, t \in A$, so that $tdx - sdy = d$, which implies that

$$tx - sy = 1,$$

since A is an integral domain. Observe that

$$\begin{pmatrix} t & -s \\ -y & x \end{pmatrix} \begin{pmatrix} x & s \\ y & t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which shows that both matrices on the left of the equation are invertible, and so is the transpose of the second one,

$$\begin{pmatrix} x & y \\ s & t \end{pmatrix}$$

(they all have determinant 1). We also have

$$as + bt = tds - sdt = 0,$$

so

$$\begin{pmatrix} x & y \\ s & t \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} x & s \\ y & t \end{pmatrix} = \begin{pmatrix} d & 0 \end{pmatrix}.$$

Because a does not divide b, their gcd d has strictly fewer prime factors than a, so

$$\sigma(d) < \sigma(a).$$

Using matrices of the form

$$\begin{pmatrix} x & y & 0 & 0 & \cdots & 0 \\ s & t & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

with $xt - ys = 1$, we can modify Steps 2a and Step 2b to obtain the following theorem.

Theorem 36.21. *If M is an $m \times n$ matrix over a PID A, then there exist some invertible $n \times n$ matrix P and some invertible $m \times m$ matrix Q, where P and Q are products of elementary matrices and matrices of the form*

$$\begin{pmatrix} x & y & 0 & 0 & \cdots & 0 \\ s & t & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

with $xt - ys = 1$, and a $m \times n$ matrix D of the form

$$D = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \alpha_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

for some nonzero $\alpha_i \in A$, such that

(1) $\alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r$, and

(2) $M = QDP^{-1}$.

Proof sketch. In Step 2a, if a_{11} does not divide a_{k1}, then first permute row 2 and row k (if $k \neq 2$). Then, if we write $a = a_{11}$ and $b = a_{k1}$, if d is a gcd of a and b and if x, y, s, t are determined as explained above, multiply on the left by the matrix

$$\begin{pmatrix} x & y & 0 & 0 & \cdots & 0 \\ s & t & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

to obtain a matrix of the form

$$\begin{pmatrix} d & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

with $\sigma(d) < \sigma(a_{11})$. Then, go back to Step 2a.

In Step 2b, if a_{11} does not divide a_{1k}, then first permute column 2 and column k (if $k \neq 2$). Then, if we write $a = a_{11}$ and $b = a_{1k}$, if d is a gcd of a and b and if x, y, s, t are determined as explained above, multiply on the right by the matrix

$$\begin{pmatrix} x & s & 0 & 0 & \cdots & 0 \\ y & t & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

to obtain a matrix of the form

$$\begin{pmatrix} d & 0 & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mn} \end{pmatrix}$$

with $\sigma(d) < \sigma(a_{11})$. Then, go back to Step 2b. The other steps remain the same. Whenever we return to Step 2a or Step 2b, the σ-value of the $(1,1)$-entry strictly decreases, so the whole procedure terminates. $\square$

We conclude this section by explaining how the rational canonical form of a matrix A can be obtained from the canonical form QDP^{-1} of $XI - A$.

Let $f: E \to E$ be a linear map over a K-vector space of dimension n. Recall from Theorem 36.3 (see Section 36.1) that as a $K[X]$-module, E_f is the image of the free module $E[X]$ by the map $\sigma: E[X] \to E_f$, where $E[X]$ consists of all linear combinations of the form

$$p_1 e_1 + \cdots + p_n e_n,$$

where $(e_1, \ldots, e_n)$ is a basis of E and $p_1, \ldots, p_n \in K[X]$ are polynomials, and σ is given by

$$\sigma(p_1 e_1 + \cdots + p_n e_n) = p_1(f)(e_1) + \cdots + p_n(f)(e_n).$$

Furthermore, the kernel of σ is equal to the image of the map $\psi: E[X] \to E[X]$, where

$$\psi(p_1 e_1 + \cdots + p_n e_n) = Xp_1 e_1 + \cdots + Xp_n e_n - (p_1 f(e_1) + \cdots + p_n(e_n)).$$

The matrix A is the representation of a linear map f over the canonical basis $(e_1, \ldots, e_n)$ of $E = K^n$, and and $XI - A$ is the matrix of ψ with respect to the basis $(e_1, \ldots, e_n)$

(over $K[X]$). What Theorem 36.20 tells us is that there are $K[X]$-bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ of E_f with respect to which the matrix of ψ is D. Then

$$\psi(u_i) = v_i, \quad i = 1, \ldots, n-m,$$
$$\psi(u_{n-m+i}) = q_i v_{n-m+i}, \quad i = 1, \ldots, m,$$

and because $\mathrm{Im}(\psi) = \mathrm{Ker}\,(\sigma)$, this implies that

$$\sigma(v_i) = 0, \quad i = 1, \ldots, n-m.$$

Consequently, $w_1 = \sigma(v_{n-m+1}), \ldots, w_m = \sigma(v_n)$ span E_f as a $K[X]$-module, with $w_i \in E$, and we have

$$M(f) = K[X]w_1 \oplus \cdots \oplus K[X]w_m,$$

where $K[X]w_i \approx K[X]/(q_i)$ as a cyclic $K[X]$-module. Since $\mathrm{Im}(\psi) = \mathrm{Ker}\,(\sigma)$, we have

$$0 = \sigma(\psi(u_{n-m+i})) = \sigma(q_i v_{n-m+i}) = q_i \sigma(v_{n-m+i}) = q_i w_i,$$

so as a K-vector space, the cyclic subspace $Z(w_i; f) = K[X]w_i$ has q_i as annihilator, and by a remark from Section 36.1, it has the basis (over K)

$$(w_i, f(w_i), \ldots, f^{n_i - 1}(w_i)), \quad n_i = \deg(q_i).$$

Furthermore, over this basis, the restriction of f to $Z(w_i; f)$ is represented by the companion matrix of q_i. By putting all these bases together, we obtain a block matrix which is the canonical rational form of f (and A).

Now, $XI - A = QDP^{-1}$ is the matrix of ψ with respect to the canonical basis $(e_1, \ldots, e_n)$ (over $K[X]$), and D is the matrix of ψ with respect to the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ (over $K[X]$), which tells us that the columns of Q consist of the coordinates (in $K[X]$) of the basis vectors $(v_1, \ldots, v_n)$ with respect to the basis $(e_1, \ldots, e_n)$. Therefore, the coordinates (in K) of the vectors $(w_1, \ldots, w_m)$ spanning E_f over $K[X]$, where $w_i = \sigma(v_{n-m+i})$, are obtained by substituting the matrix A for X in the coordinates of the columns vectors of Q, and evaluating the resulting expressions.

Since

$$D = Q^{-1}(XI - A)P,$$

the matrix D is obtained from A by a sequence of elementary row operations whose product is Q^{-1} and a sequence of elementary column operations whose product is P. Therefore, to compute the vectors $w_1, \ldots, w_m$ from A, we simply have to figure out how to construct Q from the sequence of elementary row operations that yield Q^{-1}. The trick is to use column operations to gather a product of row operations in reverse order.

Indeed, if Q^{-1} is the product of elementary row operations

$$Q^{-1} = E_k \cdots E_2 E_1,$$

then

$$Q = E_1^{-1} E_2^{-1} \cdots E_k^{-1}.$$

Now, row operations operate on the left and column operations operate on the right, so the product $E_1^{-1} E_2^{-1} \cdots E_k^{-1}$ can be computed from left to right as a sequence of column operations.

Let us review the meaning of the elementary row and column operations $P(i,k)$, $E_{i,j;\beta}$, and $E_{i,\lambda}$.

1. As a row operation, $P(i,k)$ permutes row i and row k.
2. As a column operation, $P(i,k)$ permutes column i and column k.
3. The inverse of $P(i,k)$ is $P(i,k)$ itself.
4. As a row operation, $E_{i,j;\beta}$ adds β times row j to row i.
5. As a column operation, $E_{i,j;\beta}$ adds β times column i to column j (note the switch in the indices).
6. The inverse of $E_{i,j;\beta}$ is $E_{i,j;-\beta}$.
7. As a row operation, $E_{i,\lambda}$ multiplies row i by λ.
8. As a column operation, $E_{i,\lambda}$ multiplies column i by λ.
9. The inverse of $E_{i,\lambda}$ is $E_{i,\lambda^{-1}}$.

Given a square matrix A (over K), the row and column operations applied to $XI - A$ in converting it to its Smith normal form may involve coefficients that are polynomials and it is necessary to explain what is the action of an operation $E_{i,j;\beta}$ in this case. If the coefficient β in $E_{i,j;\beta}$ is a polynomial over K, as a row operation, the action of $E_{i,j;\beta}$ on a matrix X is to multiply the jth row of M by the matrix $\beta(A)$ obtained by substituting the matrix A for X and then to add the resulting vector to row i. Similarly, as a column operation, the action of $E_{i,j;\beta}$ on a matrix X is to multiply the ith column of M by the matrix $\beta(A)$ obtained by substituting the matrix A for X and then to add the resulting vector to column j. An algorithm to compute the rational canonical form of a matrix can now be given. We apply the elementary column operations E_i^{-1} for $i = 1, \ldots k$, starting with the identity matrix.

Algorithm for Converting an $n \times n$ matrix to Rational Canonical Form

While applying elementary row and column operations to compute the Smith normal form D of $XI - A$, keep track of the row operations and perform the following steps:

1. Let $P' = I_n$, and for every elementary row operation E do the following:

    (a) If $E = P(i,k)$, permute column i and column k of P'.

(b) If $E = E_{i,j;\beta}$, multiply the ith column of P' by the matrix $\beta(A)$ obtained by substituting the matrix A for X, and then subtract the resulting vector from column j.

(c) If $E = E_{i,\lambda}$ where $\lambda \in K$, then multiply the ith column of P' by λ^{-1}.

2. When step (1) terminates, the first $n-m$ columns of P' are zero and the last m are linearly independent. For $i = 1, \ldots, m$, multiply the $(n-m+i)$th column w_i of P' successively by I, A^1, A^2, A^{n_i-1}, where n_i is the degree of the polynomial q_i (appearing in D), and form the $n \times n$ matrix P consisting of the vectors
$$w_1, Aw_1, \ldots, A^{n_1-1}w_1, w_2, Aw_2, \ldots, A^{n_2-1}w_2, \ldots, w_m, Aw_m, \ldots, A^{n_m-1}w_m.$$
Then, $P^{-1}AP$ is the canonical rational form of A.

Here is an example taken from Dummit and Foote [54] (Chapter 12, Section 12.2). Let A be the matrix
$$A = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix}.$$
One should check that the following sequence of row and column operations produces the Smith normal form D of $XI - A$:
$$\begin{array}{llllll} \text{row } P(1,3) & \text{row } E_{1,-1} & \text{row } E_{2,1;2} & \text{row } E_{3,1;-(X-1)} & \text{column } E_{1,3;X-1} & \text{column } E_{1,4;2} \\ \text{row } P(2,4) & \text{row} E_{2,-1} & \text{row } E_{3,2;2} & \text{row } E_{4,2;-(X+1)} & \text{column } E_{2,3;2} & \text{column } E_{2,4;X-3}, \end{array}$$
with
$$D = \begin{pmatrix} 1 & 0 & 0 & \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (X-1)^2 & 0 \\ 0 & 0 & 0 & (X-1)^2 \end{pmatrix}.$$
Then, applying Step 1 of the above algorithm, we get the sequence of column operations:
$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{P(1,3)} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{E_{1,-1}} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{E_{2,1,-2}}$$
$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ -2 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{E_{3,1,A-I}} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{P(2,4)} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \xrightarrow{E_{2,-1}}$$
$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \xrightarrow{E_{3,2,-2}} \begin{pmatrix} 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \xrightarrow{E_{4,2;A+I}} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = P'.$$

Step 2 of the algorithm yields the vectors

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad A\begin{pmatrix}1\\0\\0\\0\end{pmatrix} = \begin{pmatrix}1\\2\\1\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad A\begin{pmatrix}0\\1\\0\\0\end{pmatrix} = \begin{pmatrix}2\\-1\\0\\1\end{pmatrix},$$

so we get

$$P = \begin{pmatrix}1 & 1 & 0 & 2\\0 & 2 & 1 & -1\\0 & 1 & 0 & 0\\0 & 0 & 0 & 1\end{pmatrix}.$$

We find that

$$P^{-1} = \begin{pmatrix}1 & 0 & -1 & -2\\0 & 0 & 1 & 0\\0 & 1 & -2 & 1\\0 & 0 & 0 & 1\end{pmatrix},$$

and thus, the rational canonical form of A is

$$P^{-1}AP = \begin{pmatrix}0 & -1 & 0 & 0\\1 & 2 & 0 & 0\\0 & 0 & 0 & -1\\0 & 0 & 1 & 2\end{pmatrix}.$$

Part V

Topology, Differential Calculus

Chapter 37

Topology

37.1 Metric Spaces and Normed Vector Spaces

This chapter contains a review of basic topological concepts. First metric spaces are defined. Next normed vector spaces are defined. Closed and open sets are defined, and their basic properties are stated. The general concept of a topological space is defined. The closure and the interior of a subset are defined. The subspace topology and the product topology are defined. Continuous maps and homeomorphisms are defined. Limits of sequences are defined. Continuous linear maps and multilinear maps are defined and studied briefly. The chapter ends with the definition of a normed affine space.

Most spaces considered in this book have a topological structure given by a metric or a norm, and we first review these notions. We begin with metric spaces. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Definition 37.1. A *metric space* is a set E together with a function $d\colon E \times E \to \mathbb{R}_+$, called a *metric, or distance*, assigning a nonnegative real number $d(x, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x, y) = d(y, x)$. (symmetry)

(D2) $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

Geometrically, Condition (D3) expresses the fact that in a triangle with vertices x, y, z, the length of any side is bounded by the sum of the lengths of the other two sides. From (D3), we immediately get

$$|d(x, y) - d(y, z)| \leq d(x, z).$$

Let us give some examples of metric spaces. Recall that the *absolute value* $|x|$ of a real number $x \in \mathbb{R}$ is defined such that $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$, and for a complex number $x = a + ib$, by $|x| = \sqrt{a^2 + b^2}$.

Example 37.1.

1. Let $E = \mathbb{R}$, and $d(x, y) = |x - y|$, the absolute value of $x - y$. This is the so-called natural metric on $\mathbb{R}$.

2. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the *Euclidean metric*

$$d_2(x, y) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{\frac{1}{2}},$$

the distance between the points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

3. For every set E, we can define the *discrete metric*, defined such that $d(x, y) = 1$ iff $x \neq y$, and $d(x, x) = 0$.

4. For any $a, b \in \mathbb{R}$ such that $a < b$, we define the following sets:

$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, (closed interval)

$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$, (open interval)

$[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$, (interval closed on the left, open on the right)

$(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$, (interval open on the left, closed on the right)

Let $E = [a, b]$, and $d(x, y) = |x - y|$. Then, $([a, b], d)$ is a metric space.

We will need to define the notion of proximity in order to define convergence of limits and continuity of functions. For this we introduce some standard "small neighborhoods."

Definition 37.2. Given a metric space E with metric d, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set

$$B(a, \rho) = \{x \in E \mid d(a, x) \leq \rho\}$$

is called the *closed ball of center a and radius ρ*, the set

$$B_0(a, \rho) = \{x \in E \mid d(a, x) < \rho\}$$

is called the *open ball of center a and radius ρ*, and the set

$$S(a, \rho) = \{x \in E \mid d(a, x) = \rho\}$$

is called the *sphere of center a and radius ρ*. It should be noted that ρ is finite (i.e., not $+\infty$). A subset X of a metric space E is *bounded* if there is a closed ball $B(a, \rho)$ such that $X \subseteq B(a, \rho)$.

Clearly, $B(a, \rho) = B_0(a, \rho) \cup S(a, \rho)$.

Example 37.2.

1. In $E = \mathbb{R}$ with the distance $|x - y|$, an open ball of center a and radius ρ is the open interval $(a - \rho, a + \rho)$.
2. In $E = \mathbb{R}^2$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the disk of center a and radius ρ, excluding the boundary points on the circle.
3. In $E = \mathbb{R}^3$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the sphere of center a and radius ρ, excluding the boundary points on the sphere.

One should be aware that intuition can be misleading in forming a geometric image of a closed (or open) ball. For example, if d is the discrete metric, a closed ball of center a and radius $\rho < 1$ consists only of its center a, and a closed ball of center a and radius $\rho \geq 1$ consists of the entire space!

If $E = [a, b]$, and $d(x, y) = |x - y|$, as in Example 37.1, an open ball $B_0(a, \rho)$, with $\rho < b - a$, is in fact the interval $[a, a + \rho)$, which is closed on the left.

We now consider a very important special case of metric spaces, normed vector spaces. Normed vector spaces have already been defined in Chapter 9 (Definition 9.1) but for the reader's convenience we repeat the definition.

Definition 37.3. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm on E* is a function $\| \, \| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \, \|$ is called a *normed vector space*.

From (N3), we easily get

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Given a normed vector space E, if we define d such that

$$d(x, y) = \|x - y\|,$$

it is easily seen that d is a metric. Thus, every normed vector space is immediately a metric space. Note that the metric associated with a norm is invariant under translation, that is,

$$d(x + u, y + u) = d(x, y).$$

For this reason, we can restrict ourselves to open or closed balls of center 0.

Examples of normed vector spaces were given in Example 9.1. We repeat the most important examples.

Example 37.3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the ℓ_p-*norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

We proved in Proposition 9.1 that the ℓ_p-norms are indeed norms. The closed unit balls centered at $(0, 0)$ for $\| \,\|_1$, $\| \,\|_2$, and $\| \,\|_\infty$, along with the containment relationships, are shown in Figures 37.1 and 37.2. Figures 37.3 and 37.4 illustrate the situation in $\mathbb{R}^3$.

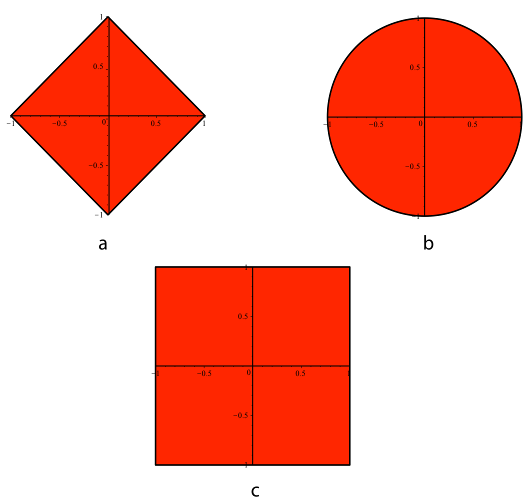


Figure 37.1: Figure (a) shows the diamond shaped closed ball associated with $\| \,\|_1$. Figure (b) shows the closed unit disk associated with $\| \,\|_2$, while Figure (c) illustrates the closed unit ball associated with $\| \,\|_\infty$.

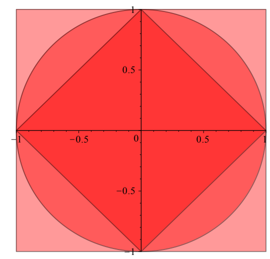


Figure 37.2: The relationship between the closed unit balls centered at $(0,0)$.

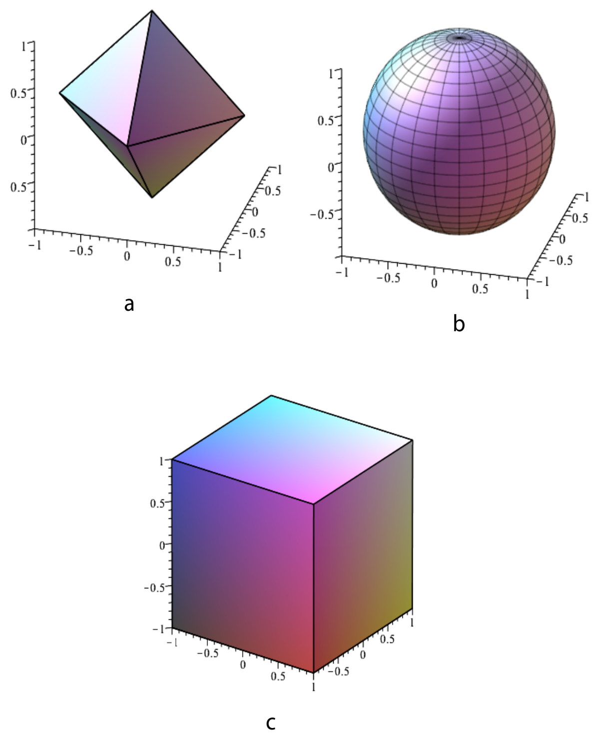


Figure 37.3: Figure (a) shows the octahedral shaped closed ball associated with $\| \|_1$. Figure (b) shows the closed spherical associated with $\| \|_2$, while Figure (c) illustrates the closed unit ball associated with $\| \|_\infty$.

In a normed vector space we define a closed ball or an open ball of radius ρ as a closed ball or an open ball of center 0. We may use the notation $B(\rho)$ and $B_0(\rho)$.

We will now define the crucial notions of open sets and closed sets, and of a topological space.

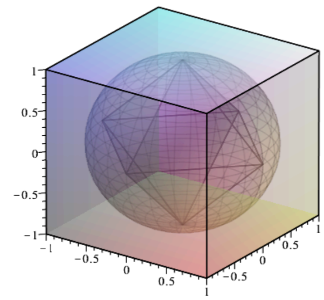


Figure 37.4: The relationship between the closed unit balls centered at $(0, 0, 0)$.

Definition 37.4. Let (E, d) be a metric space. A subset $U \subseteq E$ is an *open set* in E if either $U = \emptyset$, or for every $a \in U$, there is some open ball $B_0(a, \rho)$ such that, $B_0(a, \rho) \subseteq U$.[1] A subset $F \subseteq E$ is a *closed set* in E if its complement $E - F$ is open in E. See Figure 37.5.

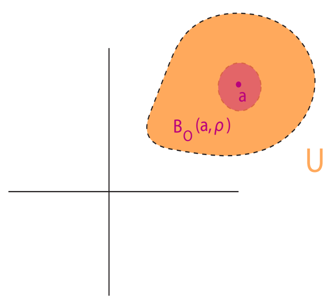


Figure 37.5: An open set U in $E = \mathbb{R}^2$ under the standard Euclidean metric. Any point in the peach set U is surrounded by a small raspberry open set which lies within U.

The set E itself is open, since for every $a \in E$, every open ball of center a is contained in E. In $E = \mathbb{R}^n$, given n intervals $[a_i, b_i]$, with $a_i < b_i$, it is easy to show that the open n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i < x_i < b_i,\ 1 \leq i \leq n\}$$

is an open set. In fact, it is possible to find a metric for which such open n-cubes are open balls! Similarly, we can define the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\},$$

which is a closed set.

The open sets satisfy some important properties that lead to the definition of a topological space.

[1]Recall that $\rho > 0$.

Proposition 37.1. *Given a metric space E with metric d, the family $\mathcal{O}$ of all open sets defined in Definition 37.4 satisfies the following properties:*

(O1) For every finite family $(U_i)_{1\leq i\leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(O2) For every arbitrary family $(U_i)_{i\in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i\in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(O3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

Furthermore, for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$.

Proof. It is straightforward. For the last point, letting $\rho = d(a, b)/3$ (in fact $\rho = d(a, b)/2$ works too), we can pick $U_a = B_0(a, \rho)$ and $U_b = B_0(b, \rho)$. By the triangle inequality, we must have $U_a \cap U_b = \emptyset$. ☐

The above proposition leads to the very general concept of a topological space.

One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x - y|$, letting $U_n = (-1/n, +1/n)$, each U_n is open, but $\bigcap_n U_n = \{0\}$, which is not open.

Later on, given any nonempty subset A of a metric space (E, d), we will need to know that certain special sets containing A are open.

Definition 37.5. Let (E, d) be a metric space. For any nonempty subset A of E and any $x \in E$, let

$$d(x, A) = \inf_{a\in A} d(x, a).$$

Proposition 37.2. *Let (E, d) be a metric space. For any nonempty subset A of E and for any two points $x, y \in E$, we have*

$$|d(x, A) - d(y, A)| \leq d(x, y).$$

Proof. For all $a \in A$ we have

$$d(x, a) \leq d(x, y) + d(y, a),$$

which implies

$$\begin{aligned} d(x, A) &= \inf_{a\in A} d(x, a) \\ &\leq \inf_{a\in A} (d(x, y) + d(y, a)) \\ &= d(x, y) + \inf_{a\in A} d(y, a) \\ &= d(x, y) + d(y, A). \end{aligned}$$

By symmetry, we also obtain $d(y, A) \leq d(x, y) + d(x, A)$, and thus

$$|d(x, A) - d(y, A)| \leq d(x, y),$$

as claimed. $\square$

Definition 37.6. Let (E, d) be a metric space. For any nonempty subset A of E, and any $r > 0$, let

$$V_r(A) = \{x \in E \mid d(x, A) < r\}.$$

Proposition 37.3. *Let (E, d) be a metric space. For any nonempty subset A of E, and any $r > 0$, the set $V_r(A)$ is an open set containing A.*

Proof. For any $y \in E$ such that $d(x, y) < r - d(x, A)$, by Proposition 37.2 we have

$$d(y, A) \leq d(x, A) + d(x, y) \leq d(x, A) + r - d(x, A) = r,$$

so $V_r(A)$ contains the open ball $B_0(x, r - d(x, A))$, which means that it is open. Obviously, $A \subseteq V_r(A)$. $\square$

37.2 Topological Spaces

Motivated by Proposition 37.1, a topological space is defined in terms of a family of sets satisfying the properties of open sets stated in that proposition.

Definition 37.7. Given a set E, a *topology on E (or a topological structure on E)*, is defined as a family $\mathcal{O}$ of subsets of E called *open sets*, and satisfying the following three properties:

(1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

A set E together with a topology $\mathcal{O}$ on E is called a *topological space*. Given a topological space $(E, \mathcal{O})$, a subset F of E is a *closed set* if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e., F is the complement of some open set.

It is possible that an open set is also a closed set. For example, $\emptyset$ and E are both open and closed. When a topological space contains a proper nonempty subset U which is both open and closed, the space E is said to be *disconnected*.

Definition 37.8. A topological space $(E, \mathcal{O})$ is said to satisfy the *Hausdorff separation axiom (or T_2-separation axiom)* if for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the T_2-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a *Hausdorff space*.

As shown by Proposition 37.1, any metric space is a topological Hausdorff space, the family of open sets being in fact the family of arbitrary unions of open balls. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of E is called the *discrete topology*.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough "small" open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed). Nevertheless, non-Hausdorff topological spaces arise naturally in algebraic geometry. But even there, some substitute for separation is used.

One of the reasons why topological spaces are important is that the definition of a topology only involves a certain family $\mathcal{O}$ of sets, and not **how** such family is generated from a metric or a norm. For example, different metrics or different norms can define the same family of open sets. Many topological properties only depend on the family $\mathcal{O}$ and not on the specific metric or norm. But the fact that a topology is definable from a metric or a norm is important, because it usually implies nice properties of a space. All our examples will be spaces whose topology is defined by a metric or a norm.

Definition 37.9. A topological space $(E, \mathcal{O})$ is *metrizable* if there is a distance on E defining the topology $\mathcal{O}$.

Note that in a metric space (E, d), the metric d is *explicitly given*. However, in general, the topology of a metrizable space $(E, \mathcal{O})$ is not specified by an explicitly given metric, but *some metric* defining the topology $\mathcal{O}$ exists. Obviously, a metrizable topological space must be Hausdorff. Actually, a stronger separation property holds, a metrizable space is normal; see Definition 37.30.

Remark: By taking complements we can state properties of the closed sets dual to those of Definition 37.7. Thus, $\emptyset$ and E are closed sets, and the closed sets are closed under finite unions and arbitrary intersections.

It is also worth noting that the Hausdorff separation axiom implies that for every $a \in E$, the set $\{a\}$ is closed. Indeed, if $x \in E - \{a\}$, then $x \neq a$, and so there exist open sets U_a and U_x such that $a \in U_a$, $x \in U_x$, and $U_a \cap U_x = \emptyset$. See Figure 37.6. Thus, for every $x \in E - \{a\}$, there is an open set U_x containing x and contained in $E - \{a\}$, showing by (O3) that $E - \{a\}$ is open, and thus that the set $\{a\}$ is closed.

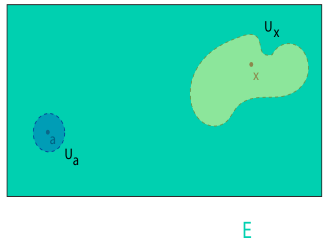


Figure 37.6: A schematic illustration of the Hausdorff separation property

Given a topological space $(E, \mathcal{O})$, given any subset A of E, since $E \in \mathcal{O}$ and E is a closed set, the family $\mathcal{C}_A = \{F \mid A \subseteq F,\ F \text{ a closed set}\}$ of closed sets containing A is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection $\bigcap \mathcal{C}_A$ of the sets in the family $\mathcal{C}_A$ is the smallest closed set containing A. By a similar reasoning, the union of all the open subsets contained in A is the largest open set contained in A.

Definition 37.10. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the smallest closed set containing A is denoted by $\overline{A}$, and is called the *closure, or adherence of* A. See Figure 37.7. A subset A of E is *dense in* E if $\overline{A} = E$. The largest open set contained in A is denoted by $\overset{\circ}{A}$, and is called the *interior of* A. See Figure 37.8. The set $\mathrm{Fr}\, A = \overline{A} \cap \overline{E - A}$ is called the *boundary (or frontier) of* A. We also denote the boundary of A by ∂A. See Figure 37.9.

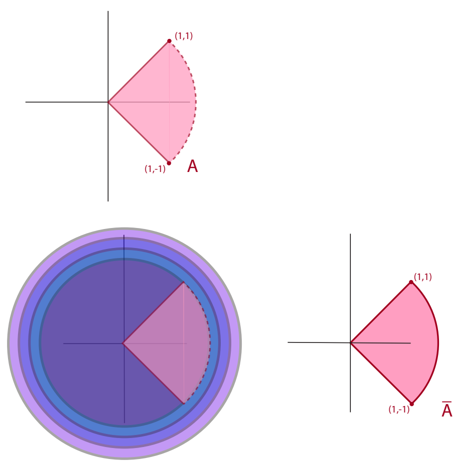


Figure 37.7: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The closure of A is obtained by the intersection of A with the closed unit ball.

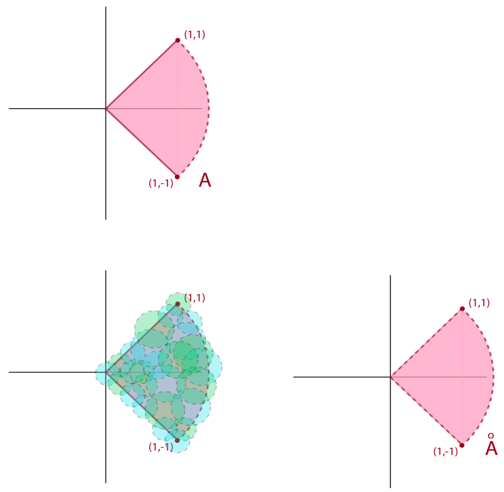


Figure 37.8: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The interior of A is obtained by the covering A with small open balls.

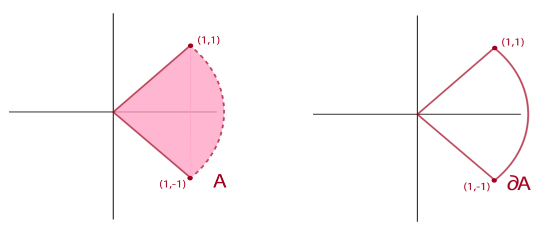


Figure 37.9: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The boundary of A is $\overline{A} - \mathring{A}$.

Remark: The notation $\overline{A}$ for the closure of a subset A of E is somewhat unfortunate, since $\overline{A}$ is often used to denote the set complement of A in E. Still, we prefer it to more cumbersome notations such as $\mathrm{clo}(A)$, and we denote the complement of A in E by $E - A$ (or sometimes, A^c).

By definition, it is clear that a subset A of E is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \mathring{A} \cup \partial A$ and $\mathring{A} \cap \partial A = \emptyset$. Another useful characterization of $\overline{A}$ is given by the following proposition.

Proposition 37.4. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, the closure $\overline{A}$ of A is the set of all points $x \in E$ such that for every open set U containing x, then $U \cap A \neq \emptyset$. See Figure 37.10.*

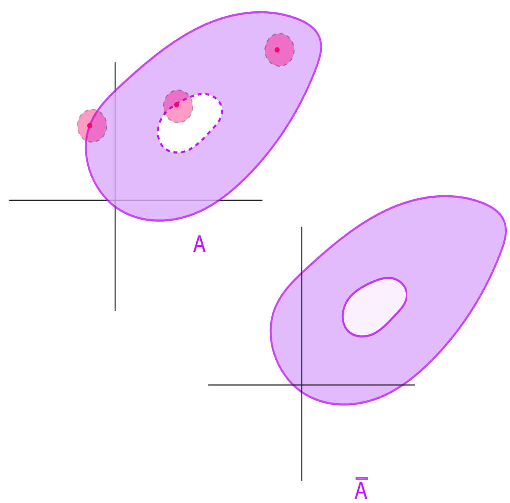


Figure 37.10: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The purple subset A is illustrated with three red points, each in its closure since the open ball centered at each point has nontrivial intersection with A.

Proof. If $A = \emptyset$, since $\emptyset$ is closed, the proposition holds trivially. Thus, assume that $A \neq \emptyset$. First assume that $x \in \overline{A}$. Let U be any open set such that $x \in U$. If $U \cap A = \emptyset$, since U is open, then $E - U$ is a closed set containing A, and since $\overline{A}$ is the intersection of all closed sets containing A, we must have $x \in E - U$, which is impossible. Conversely, assume that $x \in E$ is a point such that for every open set U containing x, then $U \cap A \neq \emptyset$. Let F be any closed subset containing A. If $x \notin F$, since F is closed, then $U = E - F$ is an open set such that $x \in U$, and $U \cap A = \emptyset$, a contradiction. Thus, we have $x \in F$ for every closed set containing A, that is, $x \in \overline{A}$. $\square$

Often it is necessary to consider a subset A of a topological space E, and to view the subset A as a topological space. The following proposition shows how to define a topology on a subset.

Proposition 37.5. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, let*

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

be the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. The following properties hold.

(1) *The space* $(A, \mathcal{U})$ *is a topological space.*

(2) *If* E *is a metric space with metric* d*, then the restriction* $d_A \colon A \times A \to \mathbb{R}_+$ *of the metric* d *to* A *defines a metric space. Furthermore, the topology induced by the metric* d_A *agrees with the topology defined by* $\mathcal{U}$*, as above.*

Proof. Left as an exercise. □

Proposition 37.5 suggests the following definition.

Definition 37.11. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the *subspace topology on* A *induced by* $\mathcal{O}$ is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

is the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. We say that $(A, \mathcal{U})$ has the *subspace topology*. If (E, d) is a metric space, the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A is called the *subspace metric*.

For example, if $E = \mathbb{R}^n$ and d is the Euclidean metric, we obtain the subspace topology on the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}.$$

See Figure 37.11,

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in A is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x - y|$, and $A = [a, b]$, then sets of the form $[a, c)$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x - y|$. However, there is agreement in the following situation.

Proposition 37.6. *Given a topological space* $(E, \mathcal{O})$*, given any subset* A *of* E*, if* $\mathcal{U}$ *is the subspace topology, then the following properties hold.*

(1) *If* A *is an open set* $A \in \mathcal{O}$*, then every open set* $U \in \mathcal{U}$ *is an open set* $U \in \mathcal{O}$*.*

(2) *If* A *is a closed set in* E*, then every closed set w.r.t. the subspace topology is a closed set w.r.t.* $\mathcal{O}$*.*

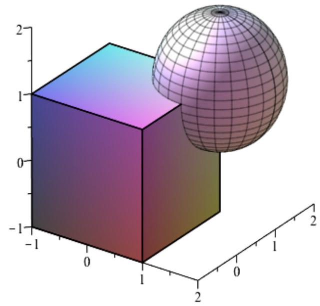
Proof. Left as an exercise. □

The concept of product topology is also useful. We have the following proposition.

Proposition 37.7. *Given* n *topological spaces* $(E_i, \mathcal{O}_i)$*, let* $\mathcal{B}$ *be the family of subsets of* $E_1 \times \cdots \times E_n$ *defined as follows:*

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

and let $\mathcal{P}$ *be the family consisting of arbitrary unions of sets in* $\mathcal{B}$*, including* $\emptyset$*. Then* $\mathcal{P}$ *is a topology on* $E_1 \times \cdots \times E_n$*.*



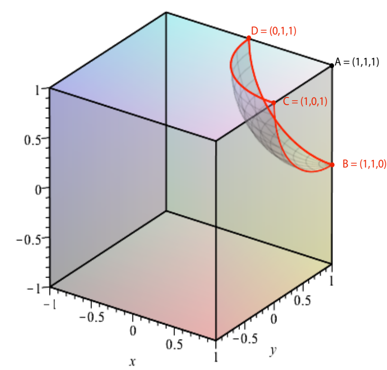


Figure 37.11: An example of an open set in the subspace topology for $\{(x, y, z) \in \mathbb{R}^3 \mid -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1\}$. The open set is the corner region $ABCD$ and is obtained by intersection the cube $B_0((1, 1, 1), 1)$.

Proof. Left as an exercise. $\square$

Definition 37.12. Given n topological spaces $(E_i, \mathcal{O}_i)$, the *product topology on* $E_1 \times \cdots \times E_n$ is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. See Figure 37.12.

If each (E_i, d_{E_i}) is a metric space, there are three natural metrics that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned}
d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= d_{E_1}(x_1, y_1) + \cdots + d_{E_n}(x_n, y_n), \\
d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \left((d_{E_1}(x_1, y_1))^2 + \cdots + (d_{E_n}(x_n, y_n))^2\right)^{\frac{1}{2}}, \\
d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \max\{d_{E_1}(x_1, y_1), \ldots, d_{E_n}(x_n, y_n)\}.
\end{aligned}$$

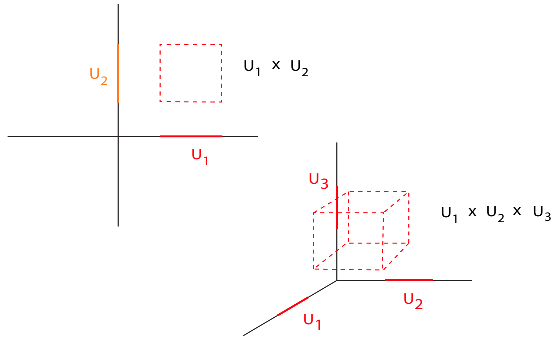


Figure 37.12: Examples of open sets in the product topology for $\mathbb{R}^2$ and $\mathbb{R}^3$ induced by the Euclidean metric.

It is easy to show that

$$\begin{aligned} d_\infty((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \leq d_2((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \leq d_1((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \\ \leq n d_\infty((x_1,\ldots,x_n),(y_1,\ldots,y_n)), \end{aligned}$$

so these distances define the same topology, which is the product topology.

If each $(E_i, \| \|_{E_i})$ is a normed vector space, there are three natural norms that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned} \|(x_1,\ldots,x_n)\|_1 &= \|x_1\|_{E_1} + \cdots + \|x_n\|_{E_n}, \\ \|(x_1,\ldots,x_n)\|_2 &= \left(\|x_1\|_{E_1}^2 + \cdots + \|x_n\|_{E_n}^2\right)^{\frac{1}{2}}, \\ \|(x_1,\ldots,x_n)\|_\infty &= \max\left\{\|x_1\|_{E_1},\ldots,\|x_n\|_{E_n}\right\}. \end{aligned}$$

It is easy to show that

$$\|(x_1,\ldots,x_n)\|_\infty \leq \|(x_1,\ldots,x_n)\|_2 \leq \|(x_1,\ldots,x_n)\|_1 \leq n\|(x_1,\ldots,x_n)\|_\infty,$$

so these norms define the same topology, which is the product topology. It can also be verified that when $E_i = \mathbb{R}$, with the standard topology induced by $|x-y|$, the topology product on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

Definition 37.13. Two metrics d and d' on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\| \|$ and $\| \|'$ on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E.

Given a topological space $(E, \mathcal{O})$, it is often useful, as in Proposition 37.7, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of E.

Definition 37.14. We say that a family $\mathcal{B}$ of subsets of E is a *basis for the topology* $\mathcal{O}$, if $\mathcal{B}$ is a subset of $\mathcal{O}$, and if every open set U in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set).

For example, given any metric space (E, d), $\mathcal{B} = \{B_0(a, \rho) \mid a \in E, \rho > 0\}$. In particular, if $d = \| \, \|_2$, the open intervals form a basis for $\mathbb{R}$, while the open disks form a basis for $\mathbb{R}^2$. The open rectangles also form a basis for $\mathbb{R}^2$ with the standard topology. See Figure 37.13.

It is immediately verified that if a family $\mathcal{B} = (U_i)_{i \in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i \in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

Definition 37.15. A *subbasis for* $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of E, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including E itself, in case of the empty intersection) is a basis of $\mathcal{O}$. See Figure 37.13.

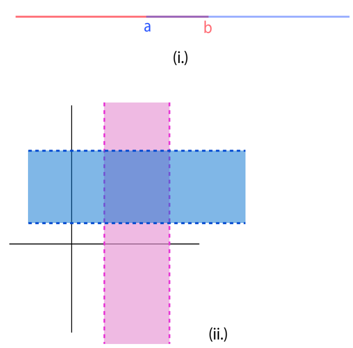


Figure 37.13: Figure $(i.)$ shows that the set of infinite open intervals forms a subbasis for $\mathbb{R}$. Figure $(ii.)$ shows that the infinite open strips form a subbasis for $\mathbb{R}^2$.

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.

Proposition 37.8. *Given a topological space $(E, \mathcal{O})$ and a family $\mathcal{B}$ of open subsets in $\mathcal{O}$ the following properties hold:*

(1) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff for every open set $U \in \mathcal{O}$ and every $x \in U$, there is some $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq U$. See Figure 37.14.

(2) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff

(a) For every $x \in E$, there is some $B \in \mathcal{B}$ such that $x \in B$.

(b) *For any two open subsets, $B_1, B_2 \in \mathcal{B}$, for every $x \in E$, if $x \in B_1 \cap B_2$, then there is some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. See Figure 37.15.*

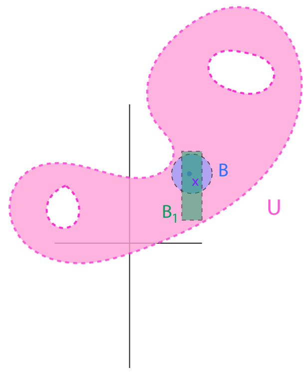


Figure 37.14: Given an open subset U of $\mathbb{R}^2$ and $x \in U$, there exists an open ball B containing x with $B \subset U$. There also exists an open rectangle B_1 containing x with $B_1 \subset U$.

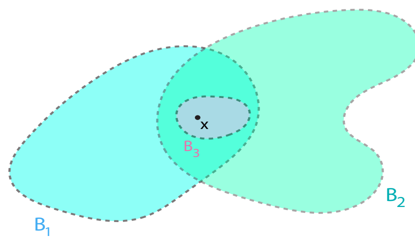


Figure 37.15: A schematic illustration of Condition (b) in Proposition 37.8.

We now consider the fundamental property of continuity.

37.3 Continuous Functions, Limits

Definition 37.16. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f \colon E \to F$ be a function. For every $a \in E$, we say that f *is continuous at* a, if for every open set $V \in \mathcal{O}_F$ containing $f(a)$, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U) \subseteq V$. See Figure 37.16. We say that f *is continuous* if it is continuous at every $a \in E$.

Define a *neighborhood of* $a \in E$ as any subset N of E containing some open set $O \in \mathcal{O}$ such that $a \in O$. If f is continuous at a and N is any neighborhood of $f(a)$, there is some open set $V \subseteq N$ containing $f(a)$, and since f is continuous at a, there is some open set U containing a, such that $f(U) \subseteq V$. Since $V \subseteq N$, the open set U is a subset of $f^{-1}(N)$

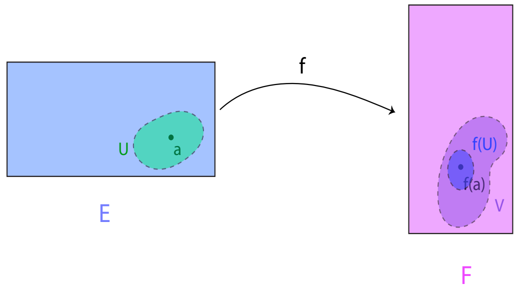


Figure 37.16: A schematic illustration of Definition 37.16.

containing a, and $f^{-1}(N)$ is a neighborhood of a. Conversely, if $f^{-1}(N)$ is a neighborhood of a whenever N is any neighborhood of $f(a)$, it is immediate that f is continuous at a. See Figure 37.17.

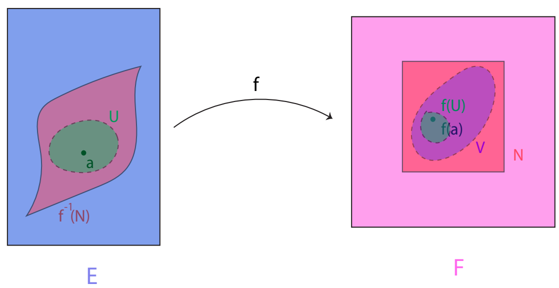


Figure 37.17: A schematic illustration of the neighborhood condition.

It is easy to see that Definition 37.16 is equivalent to the following statements.

Proposition 37.9. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f : E \to F$ be a function. For every $a \in E$, the function f is continuous at $a \in E$ iff for every neighborhood N of $f(a) \in F$, then $f^{-1}(N)$ is a neighborhood of a. The function f is continuous on E iff $f^{-1}(V)$ is an open set in $\mathcal{O}_E$ for every open set $V \in \mathcal{O}_F$.*

If E and F are metric spaces defined by metrics d_E and d_F, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } d_E(a, x) \leq \eta, \text{ then } d_F(f(a), f(x)) \leq \epsilon.$$

Similarly, if E and F are normed vector spaces defined by norms $\| \; \|_E$ and $\| \; \|_F$, we can show easily that f is continuous at a iff for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } \|x - a\|_E \le \eta, \text{ then } \|f(x) - f(a)\|_F \le \epsilon.$$

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

Definition 37.17. If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f\colon E \to F$ is a function, for every nonempty subset $A \subseteq E$ of E, we say that f *is continuous on* A if the restriction of f to A is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on A.

Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i\colon E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each π_i is continuous.

Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is *isolated* if $\{a\}$ is an open set in $\mathcal{O}$. Then if $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f\colon E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated.

In a nontrivial normed vector space $(E, \| \, \|)$ (with $E \neq \{0\}$), no point is isolated. To show this, we show that every open ball $B_0(u, \rho,)$ contains some vectors different from u. Indeed, since E is nontrivial, there is some $v \in E$ such that $v \neq 0$, and thus $\lambda = \|v\| > 0$ (by (N1)). Let

$$w = u + \frac{\rho}{\lambda + 1} v.$$

Since $v \neq 0$ and $\rho > 0$, we have $w \neq u$. Then,

$$\|w - u\| = \left\| \frac{\rho}{\lambda + 1} v \right\| = \frac{\rho\lambda}{\lambda + 1} < \rho,$$

which shows that $\|w - u\| < \rho$, for $w \neq u$.

The following proposition is easily shown.

Proposition 37.10. *Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f\colon E \to F$ and $g\colon F \to G$, if f is continuous at $a \in E$ and g is continuous at $f(a) \in F$, then $g \circ f\colon E \to G$ is continuous at $a \in E$. Given n topological spaces $(F_i, \mathcal{O}_i)$, for every function $f\colon E \to F_1 \times \cdots \times F_n$, then f is continuous at $a \in E$ iff every $f_i\colon E \to F_i$ is continuous at a, where $f_i = \pi_i \circ f$.*

One can also show that in a metric space (E, d), the distance $d\colon E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology. By the triangle inequality, we have

$$d(x, y) \le d(x, x_0) + d(x_0, y_0) + d(y_0, y) = d(x_0, y_0) + d(x_0, x) + d(y_0, y)$$

and
$$d(x_0, y_0) \leq d(x_0, x) + d(x, y) + d(y, y_0) = d(x, y) + d(x_0, x) + d(y_0, y).$$
Consequently,
$$|d(x, y) - d(x_0, y_0)| \leq d(x_0, x) + d(y_0, y),$$
which proves that d is continuous at (x_0, y_0). In fact this shows that d is uniformly continuous; see Definition 37.36.

Given any nonempty subset A of E, by Proposition 37.2, the map $x \mapsto d(x, A)$ is continuous (in fact, uniformy continuous).

Similarly, for a normed vector space $(E, \| \|)$, the norm $\| \| \colon E \to \mathbb{R}$ is (uniformly) continuous.

Given a function $f \colon E_1 \times \cdots \times E_n \to F$, we can fix $n-1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view f as a function of the remaining argument,
$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$
where $x_i \in E_i$. If f is continuous, it is clear that each f_i is continuous.

One should be careful that the converse is false! For example, consider the function $f \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined such that,
$$f(x, y) = \frac{xy}{x^2 + y^2} \quad \text{if } (x, y) \neq (0, 0), \quad \text{and} \quad f(0, 0) = 0.$$
The function f is continuous on $\mathbb{R} \times \mathbb{R} - \{(0, 0)\}$, but on the line $y = mx$, with $m \neq 0$, we have $f(x, y) = \frac{m}{1+m^2} \neq 0$, and thus, on this line, $f(x, y)$ does not approach 0 when (x, y) approaches $(0, 0)$. See Figure 37.18.

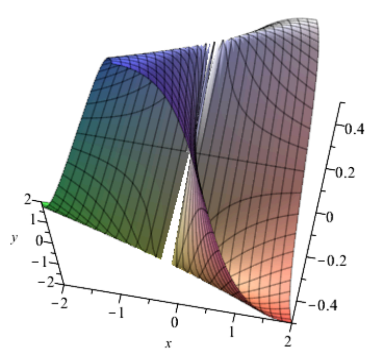


Figure 37.18: The graph of $f(x, y) = \frac{xy}{x^2+y^2}$ for $(x, y) \neq (0, 0)$. The bottom of this graph, which shows the approach along the line $y = -x$, does not have a z value of 0.

The following proposition is useful for showing that real-valued functions are continuous.

Proposition 37.11. *If E is a topological space, and $(\mathbb{R}, |x-y|)$ the reals under the standard topology, for any two functions $f\colon E \to \mathbb{R}$ and $g\colon E \to \mathbb{R}$, for any $a \in E$, for any $\lambda \in \mathbb{R}$, if f and g are continuous at a, then $f+g$, λf, $f \cdot g$, are continuous at a, and f/g is continuous at a if $g(a) \neq 0$.*

Proof. Left as an exercise. $\square$

Using Proposition 37.11, we can show easily that every real polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

Definition 37.18. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. We say that f *is a homeomorphism between E and F* if f is bijective, and both $f\colon E \to F$ and $f^{-1}\colon F \to E$ are continuous.

One should be careful that a bijective continuous function $f\colon E \to F$ is not necessarily a homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism. Another interesting example involving a parametric curve is given below. Let $L\colon \mathbb{R} \to \mathbb{R}^2$ be the function, defined such that,

$$\begin{aligned} L_1(t) &= \frac{t(1+t^2)}{1+t^4}, \\ L_2(t) &= \frac{t(1-t^2)}{1+t^4}. \end{aligned}$$

If we think of $(x(t), y(t)) = (L_1(t), L_2(t))$ as a geometric point in $\mathbb{R}^2$, the set of points $(x(t), y(t))$ obtained by letting t vary in $\mathbb{R}$ from $-\infty$ to $+\infty$, defines a curve having the shape of a "figure eight," with self-intersection at the origin, called the "lemniscate of Bernoulli." See Figure 37.19. The map L is continuous, and in fact bijective, but its inverse L^{-1} is not continuous. Indeed, when we approach the origin on the branch of the curve in the upper left quadrant (i.e., points such that, $x \leq 0$, $y \geq 0$), then t goes to $-\infty$, and when we approach the origin on the branch of the curve in the lower right quadrant (i.e., points such that, $x \geq 0$, $y \leq 0$), then t goes to $+\infty$.

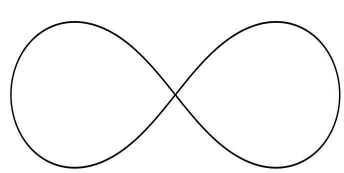


Figure 37.19: The lemniscate of Bernoulli.

We also review the concept of limit of a sequence. Given any set E, a *sequence* is any function $x\colon \mathbb{N} \to E$, usually denoted by $(x_n)_{n\in\mathbb{N}}$, or $(x_n)_{n\geq 0}$, or even by (x_n).

Definition 37.19. Given a topological space $(E, \mathcal{O})$, we say that *a sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to some* $a \in E$ if for every open set U containing a, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that *a is a limit of* $(x_n)_{n\in\mathbb{N}}$. See Figure 37.20.

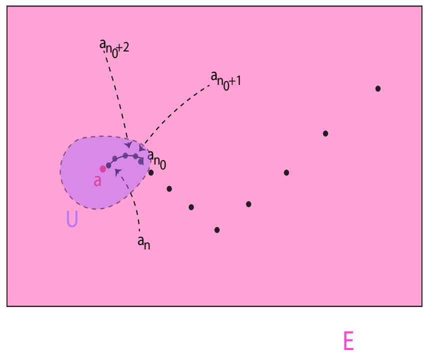


Figure 37.20: A schematic illustration of Definition 37.19.

When E is a metric space with metric d, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n, a) \leq \epsilon$, for all $n \geq n_0$.

When E is a normed vector space with norm $\| \ \|$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

Proposition 37.12. *Given a topological space* $(E, \mathcal{O})$*, if the Hausdorff separation axiom holds, then every sequence has at most one limit.*

Proof. Left as an exercise. $\square$

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit b iff it converges to the same limit b in any equivalent metric (and similarly for equivalent norms).

If E is a metric space and if A is a subset of E, there is a convenient way of showing that a point $x \in E$ belongs to the closure $\overline{A}$ of A in terms of sequences.

Proposition 37.13. *Given any metric space* (E, d)*, for any subset* A *of* E *and any point* $x \in E$*, we have* $x \in \overline{A}$ *iff there is a sequence* (a_n) *of points* $a_n \in A$ *converging to* x*.*

Proof. If the sequence (a_n) of points $a_n \in A$ converges to x, then for every open subset U of E containing x, there is some n_0 such that $a_n \in U$ for all $n \geq n_0$, so $U \cap A \neq \emptyset$, and Proposition 37.4 implies that $x \in \overline{A}$.

Conversely, assume that $x \in \overline{A}$. Then for every $n \geq 1$, consider the open ball $B_0(x, 1/n)$. By Proposition 37.4, we have $B_0(x, 1/n) \cap A \neq \emptyset$, so we can pick some $a_n \in B_0(x, 1/n) \cap A$. This, way, we define a sequence (a_n) of points in A, and by construction $d(x, a_n) < 1/n$ for all $n \geq 1$, so the sequence (a_n) converges to x. ☐

We still need one more concept of limit for functions.

Definition 37.20. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, we say that $f(x)$ *approaches* b *as* x *approaches* a *with values in* A if for every open set $V \in \mathcal{O}_F$ containing b, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U \cap A) \subseteq V$. See Figure 37.21. This is denoted by

$$\lim_{x \to a, x \in A} f(x) = b.$$

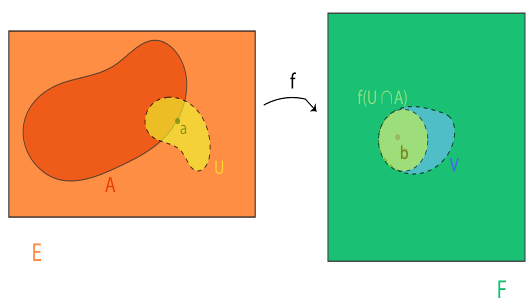


Figure 37.21: A schematic illustration of Definition 37.20.

First, note that by Proposition 37.4, since $a \in \overline{A}$, for every open set U containing a, we have $U \cap A \neq \emptyset$, and the definition is nontrivial. Also, even if $a \in A$, the value $f(a)$ of f at a plays no role in this definition. When E and F are metric space with metrics d_E and d_F, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } d_E(x, a) \leq \eta, \text{ then } d_F(f(x), b) \leq \epsilon.$$

When E and F are normed vector spaces with norms $\| \ \|_E$ and $\| \ \|_F$, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } \|x - a\|_E \leq \eta, \text{ then } \|f(x) - b\|_F \leq \epsilon.$$

We have the following result relating continuity at a point and the previous notion.

Proposition 37.14. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function. For any $a \in E$, the function f is continuous at a iff $f(x)$ approaches $f(a)$ when x approaches a (with values in E).*

Proof. Left as a trivial exercise. ☐

Another important proposition relating the notion of convergence of a sequence to continuity, is stated without proof.

Proposition 37.15. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function.*

(1) If f is continuous, then for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, if (x_n) converges to a, then $(f(x_n))$ converges to $f(a)$.

(2) If E is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever (x_n) converges to a, for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, then f is continuous.

A special case of Definition 37.20 will be used when E and F are (nontrivial) normed vector spaces with norms $\| \, \|_E$ and $\| \, \|_F$. Let U be any nonempty open subset of E. We showed earlier that E has no isolated points and that every set $\{v\}$ is closed, for every $v \in E$. Since E is nontrivial, for every $v \in U$, there is a nontrivial open ball contained in U (an open ball not reduced to its center). Then, for every $v \in U$, $A = U - \{v\}$ is open and nonempty, and clearly, $v \in \overline{A}$. For any $v \in U$, if $f(x)$ approaches b when x approaches v with values in $A = U - \{v\}$, we say that $f(x)$ *approaches b when x approaches v with values $\neq v$ in U*. This is denoted by

$$\lim_{x\to v, x\in U, x\neq v} f(x) = b.$$

Remark: Variations of the above case show up in the following case: $E = \mathbb{R}$, and F is some arbitrary topological space. Let A be some nonempty subset of $\mathbb{R}$, and let $f\colon A \to F$ be some function. For any $a \in A$, we say that f *is continuous on the right at a* if

$$\lim_{x\to a, x\in A\cap[a, +\infty)} f(x) = f(a).$$

We can define *continuity on the left* at a in a similar fashion.

Let us consider another variation. Let A be some nonempty subset of $\mathbb{R}$, and let $f\colon A \to F$ be some function. For any $a \in A$, we say that f *has a discontinuity of the first kind at a* if

$$\lim_{x\to a, x\in A\cap\, (-\infty, a)} f(x) = f(a_-)$$

and

$$\lim_{x \to a, x \in A \cap (a, +\infty)} f(x) = f(a_+)$$

both exist, and either $f(a_-) \neq f(a)$, or $f(a_+) \neq f(a)$.

Note that it is possible that $f(a_-) = f(a_+)$, but f is still discontinuous at a if this common value differs from $f(a)$. Functions defined on a nonempty subset of $\mathbb{R}$, and that are continuous, except for some points of discontinuity of the first kind, play an important role in analysis.

We now turn to connectivity properties of topological spaces.

37.4 Connected Sets

Connectivity properties of topological spaces play a very important role in understanding the topology of surfaces. This section gathers the facts needed to have a good understanding of the classification theorem for compact surfaces (with boundary). The main references are Ahlfors and Sario [2] and Massey [121, 122]. For general background on topology, geometry, and algebraic topology, we also highly recommend Bredon [30] and Fulton [67].

Definition 37.21. A topological space $(E, \mathcal{O})$ is *connected* if the only subsets of E that are both open and closed are the empty set and E itself. Equivalently, $(E, \mathcal{O})$ is connected if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty open sets U, V, or if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty closed sets. A subset, $S \subseteq E$, is *connected* if it is connected in the subspace topology on S induced by $(E, \mathcal{O})$. See Figure 37.22. A connected open set is called a *region* and a closed set is a *closed region* if its interior is a connected (open) set.

The definition of connectivity is meant to capture the fact that a connected space S is "one piece." Given the metric space $(\mathbb{R}^n, \| \, \|_2)$, the quintessential examples of connected spaces are $B_0(a, \rho)$ and $B(a, \rho)$. In particular, the following standard proposition characterizing the connected subsets of $\mathbb{R}$ can be found in most topology texts (for example, Munkres [131], Schwartz [150]). For the sake of completeness, we give a proof.

Proposition 37.16. *A subset of the real line, $\mathbb{R}$, is connected iff it is an interval, i.e., of the form $[a, b]$, $(a, b]$, where $a = -\infty$ is possible, $[a, b)$, where $b = +\infty$ is possible, or (a, b), where $a = -\infty$ or $b = +\infty$ is possible.*

Proof. Assume that A is a connected nonempty subset of $\mathbb{R}$. The cases where $A = \emptyset$ or A consists of a single point are trivial. Otherwise, we show that whenever $a, b \in A$, $a < b$, then the entire interval $[a, b]$ is a subset of A. Indeed, if this was not the case, there would be some $c \in (a, b)$ such that $c \notin A$, and then we could write $A = ((-\infty, c) \cap A) \cup ((c, +\infty) \cap A)$, where $(-\infty, c) \cap A$ and $(c, +\infty) \cap A$ are nonempty and disjoint open subsets of A, contradicting the fact that A is connected. It follows easily that A must be an interval.

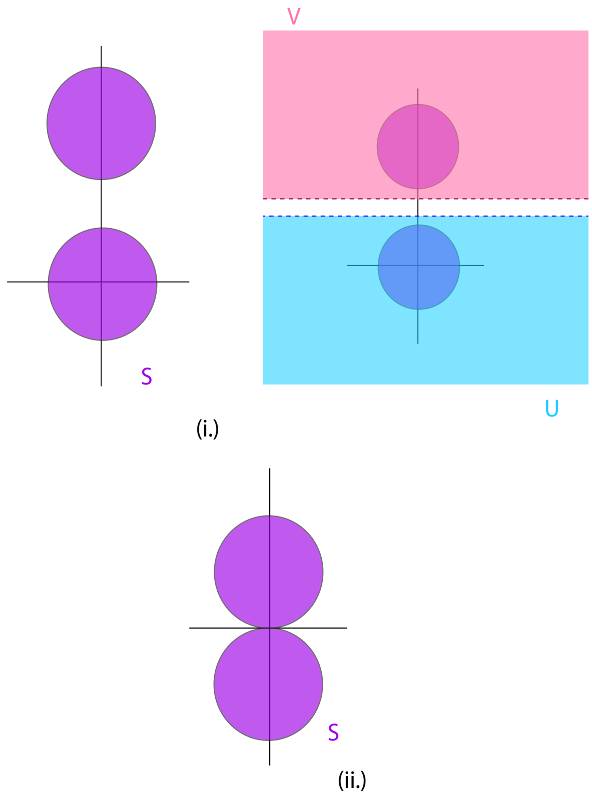


Figure 37.22: Figure (i) shows that the union of two disjoint disks in $\mathbb{R}^2$ is a disconnected set since each circle can be separated by open half regions. Figure (ii) is an example of a connected subset of $\mathbb{R}^2$ since the two disks can not separated by open sets.

Conversely, we show that an interval, I, must be connected. Let A be any nonempty subset of I which is both open and closed in I. We show that $I = A$. Fix any $x \in A$ and consider the set, R_x, of all y such that $[x, y] \subseteq A$. If the set R_x is unbounded, then $R_x = [x, +\infty)$. Otherwise, if this set is bounded, let b be its least upper bound. We claim that b is the right boundary of the interval I. Because A is closed in I, unless I is open on the right and b is its right boundary, we must have $b \in A$. In the first case, $A \cap [x, b) = I \cap [x, b) = [x, b)$. In the second case, because A is also open in I, unless b is the right boundary of the interval I (closed on the right), there is some open set $(b - \eta, b + \eta)$ contained in A, which implies that $[x, b + \eta/2] \subseteq A$, contradicting the fact that b is the least upper bound of the set R_x. Thus, b must be the right boundary of the interval I (closed on the right). A similar argument applies to the set, L_y, of all x such that $[x, y] \subseteq A$ and either L_y is unbounded, or its greatest lower bound a is the left boundary of I (open or closed on the left). In all cases, we showed that $A = I$, and the interval must be connected. $\square$

Intuitively, if a space is not connected, it is possible to define a continuous function which is constant on disjoint "connected components" and which takes possibly distinct values on disjoint components. This can be stated in terms of the concept of a locally constant function.

Definition 37.22. Given two topological spaces X, Y, a function $f\colon X \to Y$ is *locally constant* if for every $x \in X$, there is an open set $U \subseteq X$ such that $x \in U$ and f is constant on U.

We claim that a locally constant function is continuous. In fact, we will prove that $f^{-1}(V)$ is open for every subset, $V \subseteq Y$ (not just for an open set V). It is enough to show that $f^{-1}(y)$ is open for every $y \in Y$, since for every subset $V \subseteq Y$,

$$f^{-1}(V) = \bigcup_{y \in V} f^{-1}(y),$$

and open sets are closed under arbitrary unions. However, either $f^{-1}(y) = \emptyset$ if $y \in Y - f(X)$ or f is constant on $U = f^{-1}(y)$ if $y \in f(X)$ (with value y), and since f is locally constant, for every $x \in U$, there is some open set, $W \subseteq X$, such that $x \in W$ and f is constant on W, which implies that $f(w) = y$ for all $w \in W$ and thus, that $W \subseteq U$, showing that U is a union of open sets and thus, is open. The following proposition shows that a space is connected iff every locally constant function is constant:

Proposition 37.17. *A topological space is connected iff every locally constant function is constant. See Figure 37.23.*

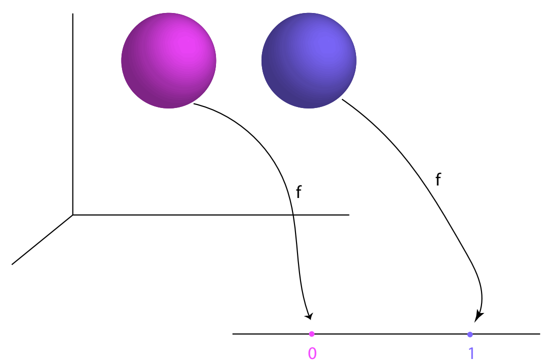


Figure 37.23: An example of a locally constant, but not constant, real-valued function f over the disconnected set consisting of the disjoint union of the two solid balls. On the pink ball, f is 0, while on the purple ball, f is 1.

Proof. First, assume that X is connected. Let $f\colon X \to Y$ be a locally constant function to some space Y and assume that f is not constant. Pick any $y \in f(X)$. Since f is not constant, $U_1 = f^{-1}(y) \neq X$, and of course, $U_1 \neq \emptyset$. We proved just before Proposition

37.17 that $f^{-1}(V)$ is open for every subset $V \subseteq Y$, and thus $U_1 = f^{-1}(y) = f^{-1}(\{y\})$ and $U_2 = f^{-1}(Y - \{y\})$ are both open, nonempty, and clearly $X = U_1 \cup U_2$ and U_1 and U_2 are disjoint. This contradicts the fact that X is connected and f must be constant.

Assume that every locally constant function $f\colon X \to Y$ is constant. If X is not connected, we can write $X = U_1 \cup U_2$, where both U_1, U_2 are open, disjoint, and nonempty. We can define the function, $f\colon X \to \mathbb{R}$, such that $f(x) = 1$ on U_1 and $f(x) = 0$ on U_2. Since U_1 and U_2 are open, the function f is locally constant, and yet not constant, a contradiction. □

A characterization on the connected subsets of $\mathbb{R}^n$ is harder and requires the notion of arcwise connectedness. One of the most important properties of connected sets is that they are preserved by continuous maps.

Proposition 37.18. *Given any continuous map, $f\colon E \to F$, if $A \subseteq E$ is connected, then $f(A)$ is connected.*

Proof. If $f(A)$ is not connected, then there exist some nonempty open sets, U, V, in F such that $f(A) \cap U$ and $f(A) \cap V$ are nonempty and disjoint, and

$$f(A) = (f(A) \cap U) \cup (f(A) \cap V).$$

Then, $f^{-1}(U)$ and $f^{-1}(V)$ are nonempty and open since f is continuous and

$$A = (A \cap f^{-1}(U)) \cup (A \cap f^{-1}(V)),$$

with $A \cap f^{-1}(U)$ and $A \cap f^{-1}(V)$ nonempty, disjoint, and open in A, contradicting the fact that A is connected. □

An important corollary of Proposition 37.18 is that for every continuous function, $f\colon E \to \mathbb{R}$, where E is a connected space, $f(E)$ is an interval. Indeed, this follows from Proposition 37.16. Thus, if f takes the values a and b where $a < b$, then f takes all values $c \in [a, b]$. This is a very important property known as the intermediate value theorem.

Even if a topological space is not connected, it turns out that it is the disjoint union of maximal connected subsets and these connected components are closed in E. In order to obtain this result, we need a few lemmas.

Lemma 37.19. *Given a topological space, E, for any family, $(A_i)_{i\in I}$, of (nonempty) connected subsets of E, if $A_i \cap A_j \neq \emptyset$ for all $i, j \in I$, then the union, $A = \bigcup_{i\in I} A_i$, of the family, $(A_i)_{i\in I}$, is also connected.*

Proof. Assume that $\bigcup_{i\in I} A_i$ is not connected. There exists two nonempty open subsets, U and V, of E such that $A \cap U$ and $A \cap V$ are disjoint and nonempty and such that

$$A = (A \cap U) \cup (A \cap V).$$

Now, for every $i \in I$, we can write

$$A_i = (A_i \cap U) \cup (A_i \cap V),$$

where $A_i \cap U$ and $A_i \cap V$ are disjoint, since $A_i \subseteq A$ and $A \cap U$ and $A \cap V$ are disjoint. Since A_i is connected, either $A_i \cap U = \emptyset$ or $A_i \cap V = \emptyset$. This implies that either $A_i \subseteq A \cap U$ or $A_i \subseteq A \cap V$. However, by assumption, $A_i \cap A_j \neq \emptyset$, for all $i, j \in I$, and thus, either both $A_i \subseteq A \cap U$ and $A_j \subseteq A \cap U$, or both $A_i \subseteq A \cap V$ and $A_j \subseteq A \cap V$, since $A \cap U$ and $A \cap V$ are disjoint. Thus, we conclude that either $A_i \subseteq A \cap U$ for all $i \in I$, or $A_i \subseteq A \cap V$ for all $i \in I$. But this proves that either

$$A = \bigcup_{i \in I} A_i \subseteq A \cap U,$$

or

$$A = \bigcup_{i \in I} A_i \subseteq A \cap V,$$

contradicting the fact that both $A \cap U$ and $A \cap V$ are disjoint and nonempty. Thus, A must be connected. ☐

In particular, the above lemma applies when the connected sets in a family $(A_i)_{i \in I}$ have a point in common.

Lemma 37.20. *If A is a connected subset of a topological space, E, then for every subset, B, such that $A \subseteq B \subseteq \overline{A}$, where $\overline{A}$ is the closure of A in E, the set B is connected.*

Proof. If B is not connected, then there are two nonempty open subsets, U, V, of E such that $B \cap U$ and $B \cap V$ are disjoint and nonempty, and

$$B = (B \cap U) \cup (B \cap V).$$

Since $A \subseteq B$, the above implies that

$$A = (A \cap U) \cup (A \cap V),$$

and since A is connected, either $A \cap U = \emptyset$, or $A \cap V = \emptyset$. Without loss of generality, assume that $A \cap V = \emptyset$, which implies that $A \subseteq A \cap U \subseteq B \cap U$. However, $B \cap U$ is closed in the subspace topology for B and since $B \subseteq \overline{A}$ and $\overline{A}$ is closed in E, the closure of A in B w.r.t. the subspace topology of B is clearly $B \cap \overline{A} = B$, which implies that $B \subseteq B \cap U$ (since the closure is the smallest closed set containing the given set). Thus, $B \cap V = \emptyset$, a contradiction. ☐

In particular, Lemma 37.20 shows that if A is a connected subset, then its closure, $\overline{A}$, is also connected. We are now ready to introduce the connected components of a space.

Definition 37.23. Given a topological space, $(E, \mathcal{O})$, we say that two points, $a, b \in E$, are *connected* if there is some connected subset, A, of E such that $a \in A$ and $b \in A$.

It is immediately verified that the relation "a and b are connected in E" is an equivalence relation. Only transitivity is not obvious, but it follows immediately as a special case of Lemma 37.19. Thus, the above equivalence relation defines a partition of E into nonempty disjoint *connected components*. The following proposition is easily proved using Lemma 37.19 and Lemma 37.20:

Proposition 37.21. *Given any topological space, E, for any $a \in E$, the connected component containing a is the largest connected set containing a. The connected components of E are closed.*

The notion of a locally connected space is also useful.

Definition 37.24. A topological space, $(E, \mathcal{O})$, is *locally connected* if for every $a \in E$, for every neighborhood, V, of a, there is a connected neighborhood, U, of a such that $U \subseteq V$. See Figure 37.24.

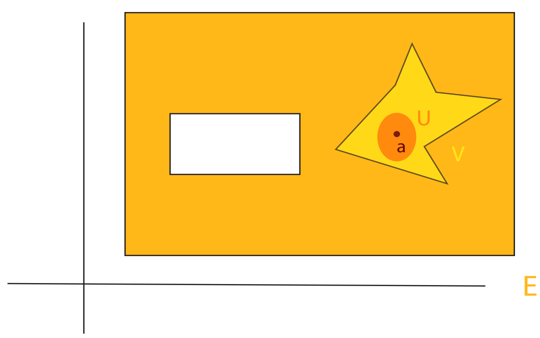


Figure 37.24: The topological space E, which is homeomorphic to an annulus, is locally connected since each point is surrounded by a small disk contained in E.

As we shall see in a moment, it would be equivalent to require that E has a basis of connected open sets.

There are connected spaces that are not locally connected and there are locally connected spaces that are not connected. The two properties are independent. For example, the subspace of $\mathbb{R}^2$ $S = \{(x, \sin(1/x)), \mid x > 0\} \cup \{(0, y) \mid -1 \leq y \leq 1\}$ is connected but not locally connected. See Figure 37.25. The subspace S of $\mathbb{R}$ consisting $[0, 1] \cup [2, 3]$ is locally connected but not connected.

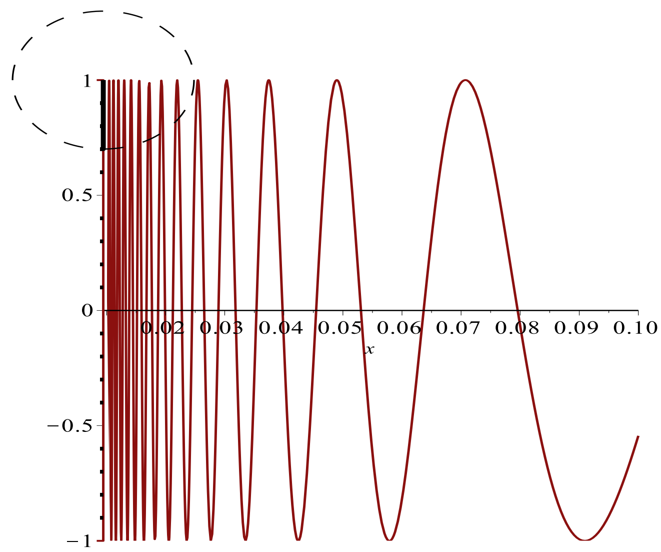


Figure 37.25: Let S be the graph of $f(x) = \sin(1/x)$ union the y-axis between -1 and 1. This space is connected, but not locally connected.

Proposition 37.22. *A topological space, E, is locally connected iff for every open subset, A, of E, the connected components of A are open.*

Proof. Assume that E is locally connected. Let A be any open subset of E and let C be one of the connected components of A. For any $a \in C \subseteq A$, there is some connected neighborhood, U, of a such that $U \subseteq A$ and since C is a connected component of A containing a, we must have $U \subseteq C$. This shows that for every $a \in C$, there is some open subset containing a contained in C, so C is open.

Conversely, assume that for every open subset, A, of E, the connected components of A are open. Then, for every $a \in E$ and every neighborhood, U, of a, since U contains some open set A containing a, the interior, $\overset{\circ}{U}$, of U is an open set containing a and its connected components are open. In particular, the connected component C containing a is a connected open set containing a and contained in U. $\square$

Proposition 37.22 shows that in a locally connected space, the connected open sets form a basis for the topology. It is easily seen that $\mathbb{R}^n$ is locally connected. Another very important property of surfaces and more generally, manifolds, is to be arcwise connected. The intuition is that any two points can be joined by a continuous arc of curve. This is formalized as follows.

Definition 37.25. Given a topological space, $(E, \mathcal{O})$, an *arc (or path)* is a continuous map, $\gamma\colon [a, b] \to E$, where $[a, b]$ is a closed interval of the real line, $\mathbb{R}$. The point $\gamma(a)$ is the *initial point* of the arc and the point $\gamma(b)$ is the *terminal point* of the arc. We say that γ *is an arc joining* $\gamma(a)$ *and* $\gamma(b)$. See Figure 37.26. An arc is a *closed curve* if $\gamma(a) = \gamma(b)$. The set $\gamma([a, b])$ is the *trace* of the arc γ.

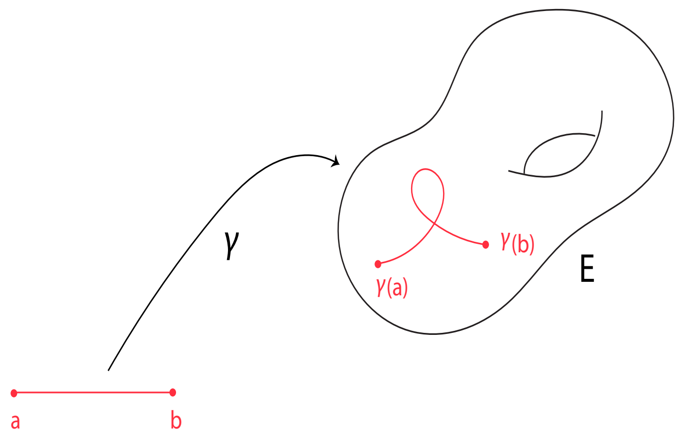


Figure 37.26: Let E be the torus with subspace topology induced from $\mathbb{R}^3$ with red arc $\gamma([a, b])$. The torus is both arcwise connected and locally arcwise connected.

Typically, $a = 0$ and $b = 1$.

One should not confuse an arc, $\gamma\colon [a, b] \to E$, with its trace. For example, γ could be constant, and thus, its trace reduced to a single point.

An arc is a *Jordan arc* if γ is a homeomorphism onto its trace. An arc, $\gamma\colon [a, b] \to E$, is a *Jordan curve* if $\gamma(a) = \gamma(b)$ and γ is injective on $[a, b)$. Since $[a, b]$ is connected, by Proposition 37.18, the trace $\gamma([a, b])$ of an arc is a connected subset of E.

Given two arcs $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$ such that $\gamma(1) = \delta(0)$, we can form a new arc defined as follows:

Definition 37.26. Given two arcs, $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$, such that $\gamma(1) = \delta(0)$, we can form their *composition (or product)*, $\gamma\delta$,, defined such that

$$\gamma\delta(t) = \begin{cases} \gamma(2t) & \text{if } 0 \leq t \leq 1/2; \\ \delta(2t - 1) & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

The *inverse*, γ^{-1}, *of the arc*, γ, is the arc defined such that $\gamma^{-1}(t) = \gamma(1-t)$, for all $t \in [0, 1]$.

It is trivially verified that Definition 37.26 yields continuous arcs.

Definition 37.27. A topological space, E, is *arcwise connected* if for any two points, $a, b \in E$, there is an arc, $\gamma\colon [0,1] \to E$, joining a and b, i.e., such that $\gamma(0) = a$ and $\gamma(1) = b$. A topological space, E, is *locally arcwise connected* if for every $a \in E$, for every neighborhood, V, of a, there is an arcwise connected neighborhood, U, of a such that $U \subseteq V$. See Figure 37.26.

The space $\mathbb{R}^n$ is locally arcwise connected, since for any open ball, any two points in this ball are joined by a line segment. Manifolds and surfaces are also locally arcwise connected. Proposition 37.18 also applies to arcwise connectedness (this is a simple exercise). The following theorem is crucial to the theory of manifolds and surfaces:

Theorem 37.23. *If a topological space, E, is arcwise connected, then it is connected. If a topological space, E, is connected and locally arcwise connected, then E is arcwise connected.*

Proof. First, assume that E is arcwise connected. Pick any point, a, in E. Since E is arcwise connected, for every $b \in E$, there is a path, $\gamma_b\colon [0,1] \to E$, from a to b and so,

$$E = \bigcup_{b \in E} \gamma_b([0,1])$$

a union of connected subsets all containing a. By Lemma 37.19, E is connected.

Now assume that E is connected and locally arcwise connected. For any point $a \in E$, let F_a be the set of all points, b, such that there is an arc, $\gamma_b\colon [0,1] \to E$, from a to b. Clearly, F_a contains a. We show that F_a is both open and closed. For any $b \in F_a$, since E is locally arcwise connected, there is an arcwise connected neighborhood U containing b (because E is a neighborhood of b). Thus, b can be joined to every point $c \in U$ by an arc, and since by the definition of F_a, there is an arc from a to b, the composition of these two arcs yields an arc from a to c, which shows that $c \in F_a$. But then $U \subseteq F_a$ and thus, F_a is open. See Figure 37.27 (*i.*). Now assume that b is in the complement of F_a. As in the previous case, there is some arcwise connected neighborhood U containing b. Thus, every point $c \in U$ can be joined to b by an arc. If there was an arc joining a to c, we would get an arc from a to b, contradicting the fact that b is in the complement of F_a. Thus, every point $c \in U$ is in the complement of F_a, which shows that U is contained in the complement of F_a, and thus, that the the complement of F_a is open. See Figure 37.27 (*ii.*). Consequently, we have shown that F_a is both open and closed and since it is nonempty, we must have $E = F_a$, which shows that E is arcwise connected. $\square$

If E is locally arcwise connected, the above argument shows that the connected components of E are arcwise connected.

It is not true that a connected space is arcwise connected. For example, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the y-axis, for which $-1 \leq y \leq 1$, is connected, but not arcwise connected. See Figure 37.25.

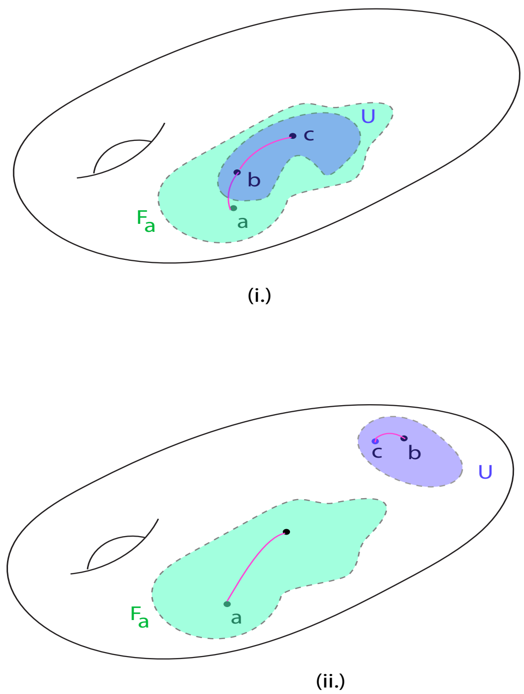


Figure 37.27: Schematic illustrations of the proof techniques that show F_a is both open and closed.

A trivial modification of the proof of Theorem 37.23 shows that in a normed vector space, E, a connected open set is arcwise connected by polygonal lines (i.e., arcs consisting of line segments). This is because in every open ball, any two points are connected by a line segment. Furthermore, if E is finite dimensional, these polygonal lines can be forced to be parallel to basis vectors.

We now consider compactness.

37.5 Compact Sets and Locally Compact Spaces

The property of compactness is very important in topology and analysis. We provide a quick review geared towards the study of manifolds, and for details, we refer the reader to Munkres [131], Schwartz [150]. In this section we will need to assume that the topological spaces are Hausdorff spaces. This is not a luxury, as many of the results are false otherwise.

We begin this section by providing the definition of compactness and describing a collection of compact spaces in $\mathbb{R}$. There are various equivalent ways of defining compactness. For our purposes, the most convenient way involves the notion of open cover.

Definition 37.28. Given a topological space E, for any subset A of E, an *open cover* $(U_i)_{i\in I}$ *of* A is a family of open subsets of E such that $A \subseteq \bigcup_{i\in I} U_i$. An *open subcover* of an open cover $(U_i)_{i\in I}$ of A is any subfamily $(U_j)_{j\in J}$ which is an open cover of A, with $J \subseteq I$. An open cover $(U_i)_{i\in I}$ of A is *finite* if I is finite. See Figure 37.28. The topological space E is *compact* if it is Hausdorff and for every open cover $(U_i)_{i\in I}$ of E, there is a finite open subcover $(U_j)_{j\in J}$ of E. Given any subset A of E, we say that A is *compact* if it is compact with respect to the subspace topology. We say that A is *relatively compact* if its closure $\overline{A}$ is compact.

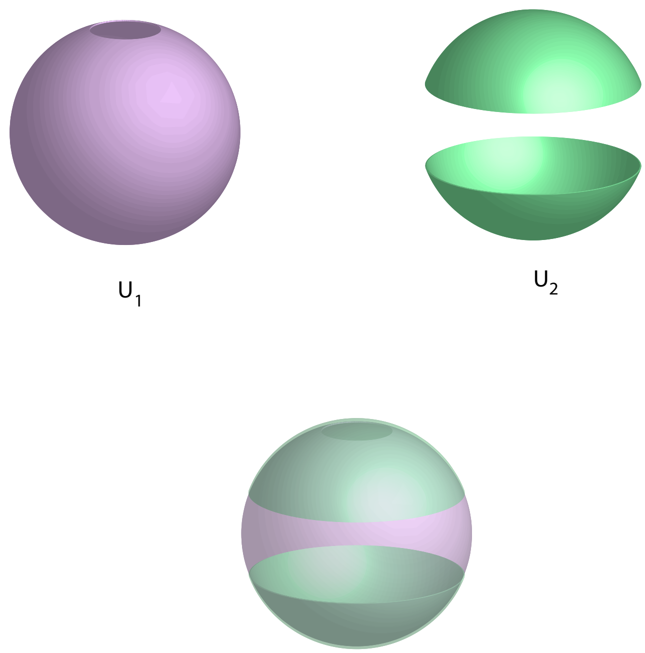


Figure 37.28: An open cover of S^2 using two open sets induced by the Euclidean topology of $\mathbb{R}^3$.

It is immediately verified that a subset A of E is compact in the subspace topology relative to A iff for every open cover $(U_i)_{i\in I}$ of A by open subsets of E, there is a finite open subcover $(U_j)_{j\in J}$ of A. The property that every open cover contains a finite open subcover is often called the *Heine-Borel-Lebesgue* property. By considering complements, a Hausdorff space is compact iff for every family $(F_i)_{i\in I}$ of closed sets, if $\bigcap_{i\in I} F_i = \emptyset$, then $\bigcap_{j\in J} F_j = \emptyset$ for some finite subset J of I.

Definition 37.28 requires that a compact space be Hausdorff. There are books in which a compact space is not necessarily required to be Hausdorff. Following Schwartz, we prefer calling such a space *quasi-compact*.

Another equivalent and useful characterization can be given in terms of families having the finite intersection property.

Definition 37.29. A family $(F_i)_{i\in I}$ of sets has the *finite intersection property* if $\bigcap_{j\in J} F_j \neq \emptyset$ for every finite subset J of I.

Proposition 37.24. *A topological Hausdorff space E is compact iff for every family $(F_i)_{i\in I}$ of closed sets having the finite intersection property, then $\bigcap_{i\in I} F_i \neq \emptyset$.*

Proof. If E is compact and $(F_i)_{i\in I}$ is a family of closed sets having the finite intersection property, then $\bigcap_{i\in I} F_i$ cannot be empty, since otherwise we would have $\bigcap_{j\in J} F_j = \emptyset$ for some finite subset J of I, a contradiction. The converse is equally obvious. $\square$

Another useful consequence of compactness is as follows. For any family $(F_i)_{i\in I}$ of closed sets such that $F_{i+1} \subseteq F_i$ for all $i \in I$, if $\bigcap_{i\in I} F_i = \emptyset$, then $F_i = \emptyset$ for some $i \in I$. Indeed, there must be some finite subset J of I such that $\bigcap_{j\in J} F_j = \emptyset$, and since $F_{i+1} \subseteq F_i$ for all $i \in I$, we must have $F_j = \emptyset$ for the smallest F_j in $(F_j)_{j\in J}$. Using this fact, we note that $\mathbb{R}$ is *not* compact. Indeed, the family of closed sets, $([n, +\infty))_{n\geq 0}$, is decreasing and has an empty intersection.

It is immediately verified that every finite union of compact subsets is compact. Similarly, every finite union of relatively compact subsets is relatively compact (use the fact that $\overline{A \cup B} = \overline{A} \cap \overline{B}$).

Given a metric space, if we define a *bounded subset* to be a subset that can be enclosed in some closed ball (of finite radius), then any nonbounded subset of a metric space is not compact. However, a closed interval $[a, b]$ of the real line is compact.

Proposition 37.25. *Every closed interval, $[a, b]$, of the real line is compact.*

Proof. We proceed by contradiction. Let $(U_i)_{i\in I}$ be any open cover of $[a, b]$ and assume that there is no finite open subcover. Let $c = (a + b)/2$. If both $[a, c]$ and $[c, b]$ had some finite open subcover, so would $[a, b]$, and thus, either $[a, c]$ does not have any finite subcover, or $[c, b]$ does not have any finite open subcover. Let $[a_1, b_1]$ be such a bad subinterval. The same argument applies and we split $[a_1, b_1]$ into two equal subintervals, one of which must be bad. Thus, having defined $[a_n, b_n]$ of length $(b - a)/2^n$ as an interval having no finite open subcover, splitting $[a_n, b_n]$ into two equal intervals, we know that at least one of the two has no finite open subcover and we denote such a bad interval by $[a_{n+1}, b_{n+1}]$. See Figure 37.29. The sequence (a_n) is nondecreasing and bounded from above by b, and thus, by a fundamental property of the real line, it converges to its least upper bound, α. Similarly, the sequence (b_n) is nonincreasing and bounded from below by a and thus, it converges to its greatest lowest bound, β. Since $[a_n, b_n]$ has length $(b-a)/2^n$, we must have $\alpha = \beta$. However, the common limit $\alpha = \beta$ of the sequences (a_n) and (b_n) must belong to some open set, U_i, of the open cover and since U_i is open, it must contain some interval $[c, d]$ containing α. Then, because α is the common limit of the sequences (a_n) and (b_n), there is some N such that the intervals $[a_n, b_n]$ are all contained in the interval $[c, d]$ for all $n \geq N$, which contradicts the fact that none of the intervals $[a_n, b_n]$ has a finite open subcover. Thus, $[a, b]$ is indeed compact. $\square$

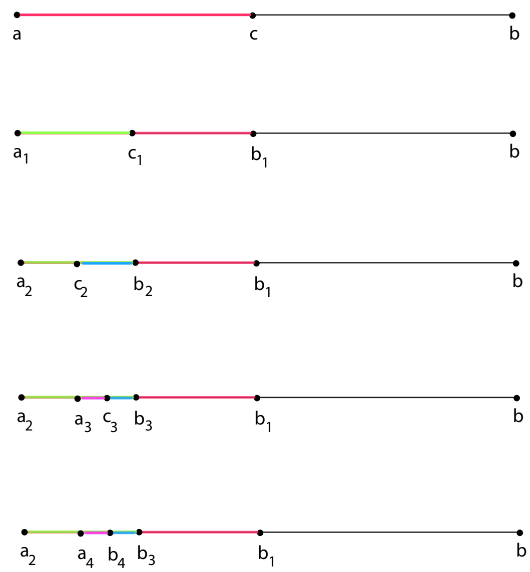


Figure 37.29: The first four stages of the nested interval construction utilized in the proof of Proposition 37.25.

The argument of Proposition 37.25 can be adapted to show that in $\mathbb{R}^m$, every closed set, $[a_1, b_1] \times \cdots \times [a_m, b_m]$, is compact. At every stage, we need to divide into 2^m subpieces instead of 2.

We next discuss some important properties of compact spaces. We begin with two separations axioms which only hold for Hausdorff spaces:

Proposition 37.26. *Given a topological Hausdorff space, E, for every compact subset, A, and every point, b, not in A, there exist disjoint open sets, U and V, such that $A \subseteq U$ and $b \in V$. See Figure 37.30. As a consequence, every compact subset is closed.*

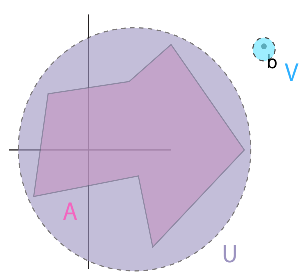


Figure 37.30: The compact set of $\mathbb{R}^2$, A, is separated by any point in its complement.

Proof. Since E is Hausdorff, for every $a \in A$, there are some disjoint open sets, U_a and V_a, containing a and b respectively. Thus, the family, $(U_a)_{a \in A}$, forms an open cover of A. Since A is compact there is a finite open subcover, $(U_j)_{j \in J}$, of A, where $J \subseteq A$, and then $\bigcup_{j \in J} U_j$ is an open set containing A disjoint from the open set $\bigcap_{j \in J} V_j$ containing b. This shows that every point, b, in the complement of A belongs to some open set in this complement and thus, that the complement is open, i.e., that A is closed. See Figure 37.31. ☐

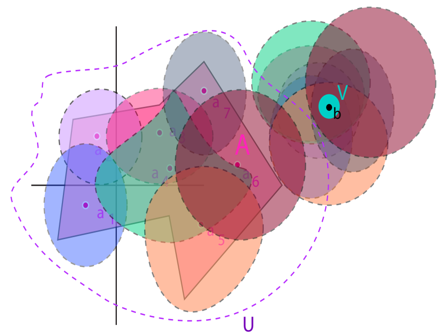


Figure 37.31: For the pink compact set A, U is the union of the seven disks which cover A, while V is the intersection of the seven open sets containing b.

Actually, the proof of Proposition 37.26 can be used to show the following useful property:

Proposition 37.27. *Given a topological Hausdorff space E, for every pair of compact disjoint subsets A and B, there exist disjoint open sets U and V, such that $A \subseteq U$ and $B \subseteq V$.*

Proof. We repeat the argument of Proposition 37.26 with B playing the role of b and use Proposition 37.26 to find disjoint open sets U_a containing $a \in A$, and V_a containing B. ☐

The following proposition shows that in a compact topological space, every closed set is compact:

Proposition 37.28. *Given a compact topological space, E, every closed set is compact.*

Proof. Since A is closed, $E - A$ is open and from any open cover, $(U_i)_{i \in I}$, of A, we can form an open cover of E by adding $E - A$ to $(U_i)_{i \in I}$ and, since E is compact, a finite subcover, $(U_j)_{j \in J} \cup \{E - A\}$, of E can be extracted such that $(U_j)_{j \in J}$ is a finite subcover of A. See Figure 37.32. ☐

Remark: Proposition 37.28 also holds for quasi-compact spaces, i.e., the Hausdorff separation property is not needed.

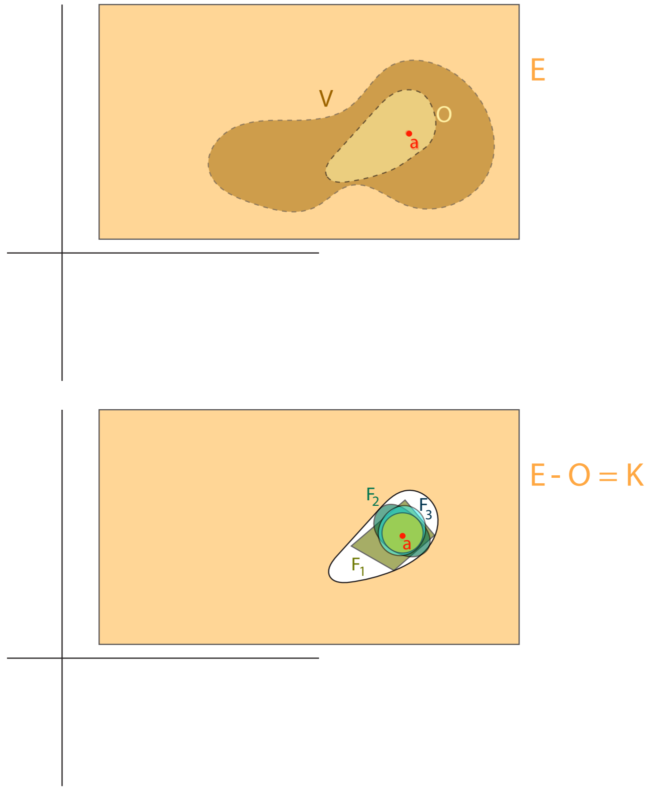


Figure 37.32: An illustration of the proof of Proposition 37.28. Both E and A are closed squares in $\mathbb{R}^2$. Note that an open cover of A, namely the green circles, when combined with the yellow square annulus $E - A$ covers all of the yellow square E.

Putting Proposition 37.27 and Proposition 37.28 together, we note that if X is compact, then for every pair of disjoint closed sets A and B, there exist disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$.

Definition 37.30. A topological space E is *normal* if every one-point set is closed, and for every pair of disjoint closed sets A and B, there exist disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$. A topological space E is *regular* if every one-point set is closed, and for every point $a \in E$ and every closed subset B of E, if $a \notin B$, then there exist disjoint open sets U and V such that $a \in U$ and $B \subseteq V$.

It is clear that a normal space is regular, and a regular space is Hausdorff. There are examples of Hausdorff spaces that are not regular, and of regular spaces that are not normal.

We just observed that a compact space is normal. An important property of metrizable spaces is that they are normal.

Proposition 37.29. *Every metrizable space E is normal.*

Proof. Assume the topology of E is given by the metric d. Since B is closed and $A \cap B = \emptyset$, for every $a \in A$ since $a \notin \overline{B} = B$, there is some open ball $B_0(a, \epsilon_a)$ of radius $\epsilon_a > 0$ such that $B_0(a, \epsilon_a) \cap B = \emptyset$. Similarly, since A is closed and $A \cap B = \emptyset$, for every $b \in B$ there is some open ball $B_0(b, \epsilon_b)$ of radius $\epsilon_b > 0$ such that $B_0(b, \epsilon_b) \cap A = \emptyset$. Let

$$U = \bigcup_{a \in A} B_0(a, \epsilon_a/2), \quad V = \bigcup_{b \in B} B_0(b, \epsilon_b/2).$$

Then A and B are open sets such that $A \subseteq U$ and $B \subseteq V$, and we claim that $U \cap V = \emptyset$.

If not, then there is some $z \in U \cap V$, which implies that for some $a \in A$ and some $b \in B$, we have

$$z \in B_0(a, \epsilon_a/2) \cap B_0(b, \epsilon_b/2).$$

It follows that

$$d(a, b) \leq d(a, z) + d(z, b) < (\epsilon_a + \epsilon_b)/2.$$

If $\epsilon_a \leq \epsilon_b$,then $d(a, b) < \epsilon_b$, so $a \in B_0(b, \epsilon_b)$, contradicting the fact that $B_0(b, \epsilon_b) \cap A = \emptyset$. If $\epsilon_b \leq \epsilon_a$,then $d(a, b) < \epsilon_a$, so $b \in B_0(a, \epsilon_a)$, contradicting the fact that $B_0(a, \epsilon_a) \cap B = \emptyset$. ☐

Compact spaces also have the following property.

Proposition 37.30. *Given a compact topological space, E, for every $a \in E$, for every neighborhood, V, of a, there exists a compact neighborhood, U, of a such that $U \subseteq V$. See Figure 37.33.*

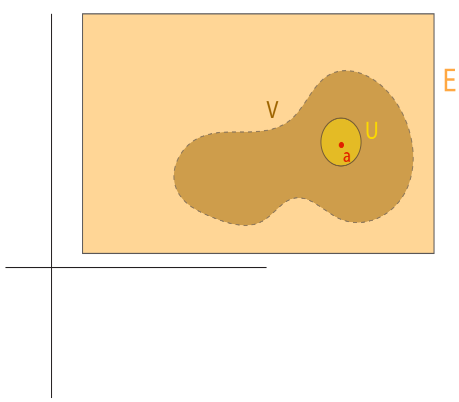


Figure 37.33: Let E be the peach square of $\mathbb{R}^2$. Each point of E is contained in a compact neighborhood U, in this case the small closed yellow disk.

Proof. Since V is a neighborhood of a, there is some open subset, O, of V containing a. Then the complement, $K = E - O$, of O is closed and since E is compact, by Proposition 37.28, K is compact. Now, if we consider the family of all closed sets of the form, $K \cap F$, where F is any closed neighborhood of a, since $a \notin K$, this family has an empty intersection and thus, there is a finite number of closed neighborhood, $F_1, \ldots, F_n$, of a, such that $K \cap F_1 \cap \cdots \cap F_n = \emptyset$. Then, $U = F_1 \cap \cdots \cap F_n$ is closed and hence by Proposition 37.28, a compact neigborhood of a contained in $O \subseteq V$. See Figure 37.34. $\square$

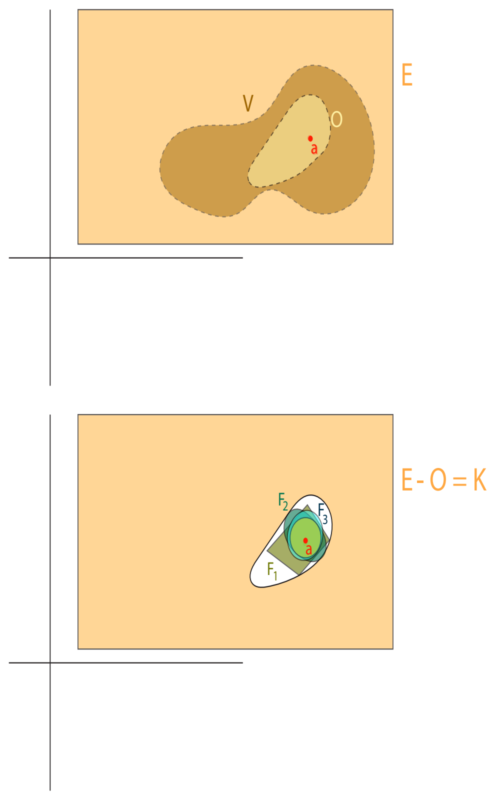


Figure 37.34: Let E be the peach square of $\mathbb{R}^2$. The compact neighborhood of a, U, is the intersection of the closed sets F_1, F_2, F_3, each of which are contained in the complement of K.

It can be shown that in a normed vector space of finite dimension, a subset is compact iff it is closed and bounded. For $\mathbb{R}^n$ the proof is simple.

In a normed vector space of infinite dimension, there are closed and bounded sets that are not compact!

More could be said about compactness in metric spaces but we will only need the notion of Lebesgue number, which will be discussed a little later. Another crucial property of compactness is that it is preserved under continuity.

Proposition 37.31. *Let E be a topological space and let F be a topological Hausdorff space. For every compact subset, A, of E, for every continuous map, $f\colon E \to F$, the subspace $f(A)$ is compact.*

Proof. Let $(U_i)_{i\in I}$ be an open cover of $f(A)$. We claim that $(f^{-1}(U_i))_{i\in I}$ is an open cover of A, which is easily checked. Since A is compact, there is a finite open subcover, $(f^{-1}(U_j))_{j\in J}$, of A, and thus, $(U_j)_{j\in J}$ is an open subcover of $f(A)$. ☐

As a corollary of Proposition 37.31, if E is compact, F is Hausdorff, and $f\colon E \to F$ is continuous and bijective, then f is a homeomorphism. Indeed, it is enough to show that f^{-1} is continuous, which is equivalent to showing that f maps closed sets to closed sets. However, closed sets are compact and Proposition 37.31 shows that compact sets are mapped to compact sets, which, by Proposition 37.26, are closed.

Another important corollary of Proposition 37.31 is the following result.

Proposition 37.32. *If E is a compact nonempty topological space and if $f\colon E \to \mathbb{R}$ is a continuous function, then there are points $a, b \in E$ such that $f(a)$ is the minimum of $f(E)$ and $f(b)$ is the maximum of $f(E)$.*

Proof. The set $f(E)$ is a compact subset of $\mathbb{R}$ and thus, a closed and bounded set which contains its greatest lower bound and its least upper bound. ☐

The following property also holds.

Proposition 37.33. *Let (E, d) be a metric space. For any nonempty subset A of E, if A is compact, then for every open subset U such that $A \subseteq U$, there is some $r > 0$ such that $V_r(A) \subseteq U$.*

Proof. The function $x \mapsto d(x, E - U)$ is continuous and $d(x, E - U) > 0$ for $x \in A$ (since $A \subseteq U$). By Proposition 37.32, there is some $a \in A$ such that

$$d(a, E - U) = \inf_{x\in A} d(x, E - U).$$

But $d(a, E - U) = r > 0$, which implies that $V_r(A) \subseteq U$. ☐

Another useful notion is that of local compactness. Indeed manifolds and surfaces are locally compact.

Definition 37.31. A topological space E is *locally compact* if it is Hausdorff and for every $a \in E$, there is some compact neighborhood K of a. See Figure 37.33.

From Proposition 37.30, every compact space is locally compact but the converse is false. For example, $\mathbb{R}$ is locally compact but not compact. In fact it can be shown that a normed vector space of finite dimension is locally compact.

Proposition 37.34. *Given a locally compact topological space, E, for every $a \in E$, for every neighborhood, N, of a, there exists a compact neighborhood, U, of a, such that $U \subseteq N$.*

Proof. For any $a \in E$, there is some compact neighborhood, V, of a. By Proposition 37.30, every neigborhood of a relative to V contains some compact neighborhood U of a relative to V. But every neighborhood of a relative to V is a neighborhood of a relative to E and every neighborhood N of a in E yields a neighborhood, $V \cap N$, of a in V and thus, for every neighborhood, N, of a, there exists a compact neighborhood, U, of a such that $U \subseteq N$. $\square$

When E is a metric space, the subsets $V_r(A)$ defined in Definition 37.6 have the following property.

Proposition 37.35. *Let (E, d) be a metric space. If E is locally compact, then for any nonempty compact subset A of E, there is some $r > 0$ such that $\overline{V_r(A)}$ is compact.*

Proof. Since E is locally compact, for every $x \in A$, there is some compact subset V_x whose interior $\overset{\circ}{V}_x$ contains x. The family of open subsets $\overset{\circ}{V}_x$ is an open cover A, and since A is compact, it has a finite subcover $\{\overset{\circ}{V}_{x_1}, \ldots, \overset{\circ}{V}_{x_n}\}$. Then $U = V_{x_1} \cup \cdots \cup V_{x_n}$ is compact (as a finite union of compact subsets), and it contains an open subset containing A (the union of the $\overset{\circ}{V}_{x_i}$). By Proposition 37.33, there is some $r > 0$ such that $V_r(A) \subseteq \overset{\circ}{U}$, and thus $\overline{V_r(A)} \subseteq U$. Since U is compact and $\overline{V_r(A)}$ is closed, $\overline{V_r(A)}$ is compact. $\square$

It is much harder to deal with noncompact manifolds than it is to deal with compact manifolds. However, manifolds are locally compact and it turns out that there are various ways of embedding a locally compact Hausdorff space into a compact Hausdorff space. The most economical construction consists in adding just one point. This construction, known as the *Alexandroff compactification*, is technically useful, and we now describe it and sketch the proof that it achieves its goal.

To help the reader's intuition, let us consider the case of the plane, $\mathbb{R}^2$. If we view the plane, $\mathbb{R}^2$, as embedded in 3-space, $\mathbb{R}^3$, say as the xy plane of equation $z = 0$, we can consider the sphere, Σ, of radius 1 centered on the z-axis at the point $(0, 0, 1)$ and tangent to the xOy plane at the origin (sphere of equation $x^2 + y^2 + (z - 1)^2 = 1$). If N denotes the north pole on the sphere, i.e., the point of coordinates $(0, 0, 2)$, then any line, D, passing through the north pole and not tangent to the sphere (i.e., not parallel to the xOy plane) intersects the xOy plane in a unique point, M, and the sphere in a unique point, P, other than the north pole, N. This, way, we obtain a bijection between the xOy plane and the punctured sphere Σ, i.e., the sphere with the north pole N deleted. This bijection is called a *stereographic projection*. See Figure 37.35.

The Alexandroff compactification of the plane puts the north pole back on the sphere, which amounts to adding a single point at infinity ∞ to the plane. Intuitively, as we travel away from the origin O towards infinity (in any direction!), we tend towards an ideal point at infinity ∞. Imagine that we "bend" the plane so that it gets wrapped around the sphere,

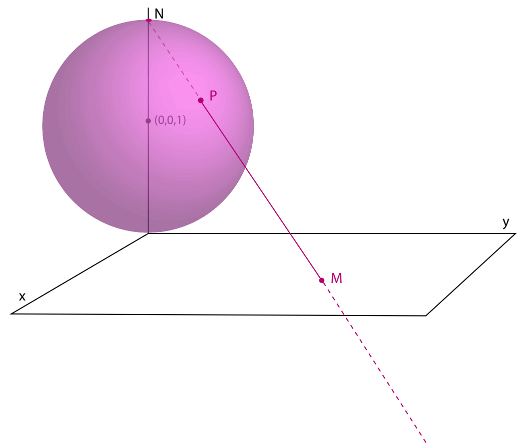


Figure 37.35: The stereographic projections of $x^2 + y^2 + (z - 1)^2 = 1$ onto the xy-plane.

according to stereographic projection. See Figure 37.36. A simpler example takes a line and gets a circle as its compactification. The Alexandroff compactification is a generalization of these simple constructions.

Definition 37.32. Let $(E, \mathcal{O})$ be a locally compact space. Let ω be any point not in E, and let $E_\omega = E \cup \{\omega\}$. Define the family, $\mathcal{O}_\omega$, as follows:

$$\mathcal{O}_\omega = \mathcal{O} \cup \{(E - K) \cup \{\omega\} \mid K \text{ compact in } E\}.$$

The pair, $(E_\omega, \mathcal{O}_\omega)$, is called the *Alexandroff compactification (or one point compactification) of* $(E, \mathcal{O})$. See Figure 37.37.

The following theorem shows that $(E_\omega, \mathcal{O}_\omega)$ is indeed a topological space, and that it is compact.

Theorem 37.36. *Let E be a locally compact topological space. The Alexandroff compactification, E_ω, of E is a compact space such that E is a subspace of E_ω and if E is not compact, then $\overline{E} = E_\omega$.*

Proof. The verification that $\mathcal{O}_\omega$ is a family of open sets is not difficult but a bit tedious. Details can be found in Munkres [131] or Schwartz [150]. Let us show that E_ω is compact. For every open cover, $(U_i)_{i \in I}$, of E_ω, since ω must be covered, there is some U_{i_0} of the form

$$U_{i_0} = (E - K_0) \cup \{\omega\}$$

where K_0 is compact in E. Consider the family, $(V_i)_{i \in I}$, defined as follows:

$$\begin{aligned} V_i &= U_i \quad \text{if} \quad U_i \in \mathcal{O}, \\ V_i &= E - K \quad \text{if} \quad U_i = (E - K) \cup \{\omega\}, \end{aligned}$$

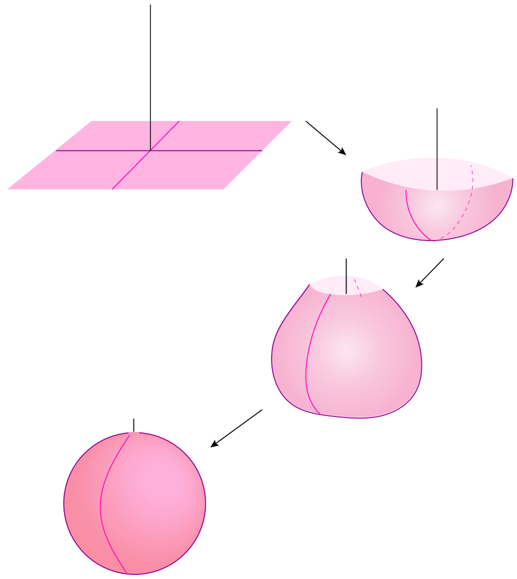


Figure 37.36: A four stage illustration of how the xy-plane is wrapped around the unit sphere centered at $(0, 0, 1)$. When finished all of the sphere is covered except the point $(0, 0, 2)$.

where K is compact in E. Then, because each K is compact and thus closed in E (since E is Hausdorff), $E - K$ is open, and every V_i is an open subset of E. Furthermore, the family, $(V_i)_{i\in(I-\{i_0\})}$, is an open cover of K_0. Since K_0 is compact, there is a finite open subcover, $(V_j)_{j\in J}$, of K_0, and thus, $(U_j)_{j\in J\cup\{i_0\}}$ is a finite open cover of E_ω.

Let us show that E_ω is Hausdorff. Given any two points, $a, b \in E_\omega$, if both $a, b \in E$, since E is Hausdorff and every open set in $\mathcal{O}$ is an open set in $\mathcal{O}_\omega$, there exist disjoint open sets, U, V (in $\mathcal{O}$), such that $a \in U$ and $b \in V$. If $b = \omega$, since E is locally compact, there is some compact set, K, containing an open set, U, containing a and then, U and $V = (E-K)\cup\{\omega\}$ are disjoint open sets (in $\mathcal{O}_\omega$) such that $a \in U$ and $b \in V$.

The space E is a subspace of E_ω because for every open set, U, in $\mathcal{O}_\omega$, either $U \in \mathcal{O}$ and $E \cap U = U$ is open in E, or $U = (E - K) \cup \{\omega\}$, where K is compact in E, and thus, $U \cap E = E - K$, which is open in E, since K is compact in E and thus, closed (since E is Hausdorff). Finally, if E is not compact, for every compact subset, K, of E, $E - K$ is nonempty and thus, for every open set, $U = (E-K)\cup\{\omega\}$, containing ω, we have $U\cap E \neq \emptyset$, which shows that $\omega \in \overline{E}$ and thus, that $\overline{E} = E_\omega$. $\square$

37.6 Second-Countable and Separable Spaces

In studying surfaces and manifolds, an important property is the existence of a countable basis for the topology. Indeed this property, among other things, guarantees the existence of triangulations of manifolds, and the fact that a manifold is metrizable.

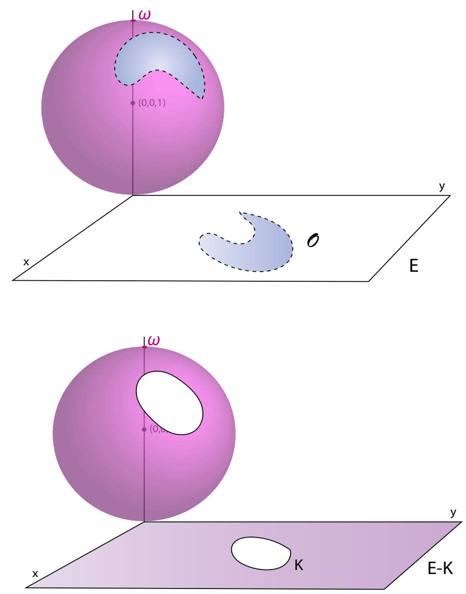


Figure 37.37: The two types of open sets associated with the Alexandroff compactification of the xy-plane. The first type of open set does not include ω, i.e. the north pole, while the second type of open set contains ω.

Definition 37.33. A topological space E is called *second-countable* if there is a countable basis for its topology, i.e., if there is a countable family, $(U_i)_{i\geq 0}$, of open sets such that every open set of E is a union of open sets U_i.

It is easily seen that $\mathbb{R}^n$ is second-countable and more generally, that every normed vector space of finite dimension is second-countable. More generally, a metric space is second-countable if and only if it is separable, a very useful property that holds for all of the spaces that we will consider in practice.

Definition 37.34. A topological space E is *separable* if it contains some countable subset S which is dense in X, that is, $\overline{S} = E$.

Observe that by Proposition 37.4, a subset S of E is dense in E if and only if every nonempty open subset of E contains some element of S.

The (metric) space $\mathbb{R}$ is separable because $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$. Similarly, $\mathbb{C}$ is separable. In general, $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$, so $\mathbb{R}^n$ is separable, and similarly, every finite-dimensional normed vector space over $\mathbb{R}$ (or $\mathbb{C}$) is separable. For metric spaces, we have the following useful result.

Proposition 37.37. *If E is a metric space, then E is second-countable iff E is separable.*

Proof. If $\mathcal{B} = (B_n)$ is a countable basis for the topology of E, then for any set S obtained by picking some point s_n in B_n, since every nonempty open subset U of E is the union of some of the B_n, the intersection $U \cap S$ is nonempty, and so S is dense in E.

Conversely, assume that there is a countable subset $S = (s_n)$ of E which is dense in E. We claim that the countable family $\mathcal{B}$ of open balls $B_0(s_n, 1/m)$ $(m \in \mathbb{N}, m > 0)$ is a basis for the topology of E. For every $x \in E$ and every $r > 0$, there is some $m > 0$ such that $1/m < r/2$, and some n such that $s_n \in B_0(x, 1/m)$. It follows that $x \in B_0(s_n, 1/m)$. For all $y \in B_0(s_n, 1/m)$, we have

$$d(x, y) \leq d(x, s_n) + d(s_n, y) \leq 2/m < r,$$

thus $B_0(s_n, 1/m) \subseteq B_0(x, r)$, which by Proposition 37.8(a) implies that $\mathcal{B}$ is a basis for the topology of E. $\square$

Proposition 37.38. *If E is a compact metric space, then E is separable.*

Proof. For every $n > 0$, the family of open balls of radius $1/n$ forms an open cover of E, and since E is compact, there is a finite subset A_n of E such that $E = \bigcup_{a_i \in A_n} B_0(a_i, 1/n)$. It is easy to see that this is equivalent to the condition $d(x, A_n) < 1/n$ for all $x \in E$. Let $A = \bigcup_{n \geq 1} A_n$. Then A is countable, and for evey $x \in E$, we have

$$d(x, A) \leq d(x, A_n) < \frac{1}{n}, \quad \text{for all } n \geq 1,$$

which implies that $d(x, A) = 0$; that is, A is dense in E. $\square$

The following theorem due to Uryshon gives a very useful sufficient condition for a topological space to be metrizable.

Theorem 37.39. *(Urysohn metrization theorem) If a topological space E is regular and second-countable, then it is metrizable.*

The proof of Theorem 37.39 can be found in Munkres [131] (Chapter 4, Theorem 34.1). As a corollary of Theorem 37.39, every (second-countable) manifold, and thus every Lie group, is metrizable.

The following technical result shows that a locally compact metrizable space which is also separable can be expressed as the union of a countable monotonic sequence of compact subsets. This gives us a method for generalizing various properties of compact metric spaces to locally compact metric spaces of the above kind.

Proposition 37.40. *Let E be a locally compact metric space. The following properties are equivalent:*

(1) There is a sequence $(U_n)_{n \geq 0}$ of open subsets such that for all $n \in \mathbb{N}$, $U_n \subseteq U_{n+1}$, $\overline{U_n}$ is compact, $\overline{U_n} \subseteq U_{n+1}$, and $E = \bigcup_{n \geq 0} U_n = \bigcup_{n \geq 0} \overline{U_n}$.

(2) The space E is the union of a countable family of compact subsets of E.

(3) The space E is separable.

Proof. We show (1) implies (2), (2) implies (3), and (3) implies (1). Obviously, (1) implies (2) since the $\overline{U_n}$ are compact.

If (2) holds, then $E = \bigcup_{n\geq 0} K_n$, for some compact subsets K_n. By Proposition 37.38, each compact subset K_n is separable, so let S_n be a countable dense subset of K_n, Then $S = \bigcup_{n\geq 0} S_n$ is a countable dense subset of E, since

$$E = \bigcup_{n\geq 0} K_n \subseteq \bigcup_{n\geq 0} \overline{S_n} \subseteq \overline{S} \subseteq E.$$

Consequently (3) holds.

If (3) holds, let $S = \{s_n\}$ be a countable dense subset of E. By Proposition 37.37, the space E has a countable basis $\mathcal{B}$ of open sets O_n. Since E is locally compact, for every $x \in E$, there is some compact neighborhood W_x containing x, and by Proposition 37.8, there some index $n(x)$ such that $x \in O_{n(x)} \subseteq W_x$. Since W_x is a compact neighborhood, we deduce that $\overline{O_{n(x)}}$ is compact. Consequently, there is a subfamily of $\mathcal{B}$ consisting of open subsets O_i such that $\overline{O_i}$ is compact, which is a countable basis for the topology of E, so we may assume that we restrict our attention to this basis. We define the sequence $(U_n)_{n\geq 1}$ of open subsets of E by induction as follows: Set $U_1 = O_1$, and let

$$U_{n+1} = O_{n+1} \cup V_r(\overline{U_n}),$$

where $r > 0$ is chosen so that $\overline{V_r(\overline{U_n})}$ is compact, which is possible by Proposition 37.35. We immediately check that the U_n satisfy (1) of Proposition 37.40. ☐

It can also be shown that if E is a locally compact space that has a countable basis, then E_ω also has a countable basis (and in fact, is metrizable).

We also have the following property.

Proposition 37.41. *Given a second-countable topological space E, every open cover $(U_i)_{i\in I}$, of E contains some countable subcover.*

Proof. Let $(O_n)_{n\geq 0}$ be a countable basis for the topology. Then all sets O_n contained in some U_i can be arranged into a countable subsequence, $(\Omega_m)_{m\geq 0}$, of $(O_n)_{n\geq 0}$ and for every Ω_m, there is some U_{i_m} such that $\Omega_m \subseteq U_{i_m}$. Furthermore, every U_i is some union of sets Ω_j, and thus, every $a \in E$ belongs to some Ω_j, which shows that $(\Omega_m)_{m\geq 0}$ is a countable open subcover of $(U_i)_{i\in I}$. ☐

As an immediate corollary of Proposition 37.41, a locally connected second-countable space has countably many connected components.

37.7 Sequential Compactness

For a general topological Hausdorff space E, the definition of compactness relies on the existence of finite cover. However, when E has a countable basis or is a metric space, we may define the notion of compactness in terms of sequences. To understand how this is done, we need to first define accumulation points.

Definition 37.35. Given a topological Hausdorff space, E, given any sequence, (x_n), of points in E, a point, $l \in E$, is an *accumulation point (or cluster point)* of the sequence (x_n) if every open set, U, containing l contains x_n for infinitely many n. See Figure 37.38.

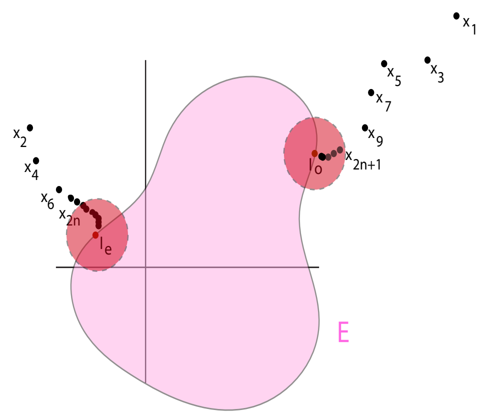


Figure 37.38: The space E is the closed, bounded pink subset of $\mathbb{R}^2$. The sequence (x_n) has two accumulation points, one for the subsequence (x_{2n+1}) and one for (x_{2n}).

Clearly, if l is a limit of the sequence, (x_n), then it is an accumulation point, since every open set, U, containing a contains all x_n except for finitely many n.

For second-countable spaces we are able to give another characterization of accumulation points.

Proposition 37.42. *Given a second-countable topological Hausdorff space, E, a point, l, is an accumulation point of the sequence, (x_n), iff l is the limit of some subsequence, (x_{n_k}), of (x_n).*

Proof. Clearly, if l is the limit of some subsequence (x_{n_k}) of (x_n), it is an accumulation point of (x_n).

Conversely, let $(U_k)_{k\geq 0}$ be the sequence of open sets containing l, where each U_k belongs to a countable basis of E, and let $V_k = U_1 \cap \cdots \cap U_k$. For every $k \geq 1$, we can find some $n_k > n_{k-1}$ such that $x_{n_k} \in V_k$, since l is an accumulation point of (x_n). Now, since every open set containing l contains some U_{k_0} and since $x_{n_k} \in U_{k_0}$ for all $k \geq 0$, the sequence (x_{n_k}) has limit l. $\square$

Remark: Proposition 37.42 also holds for metric spaces.

As an illustration of Proposition 37.42 let (x_n) be the sequence $(1, -1, 1, -1, \ldots)$. This sequence has two accumulation points, namely 1 and -1 since $(x_{2n+1}) = (1)$ and $(x_{2n}) = (-1)$.

In second-countable Hausdorff spaces, compactness can be characterized in terms of accumulation points (this is also true for metric spaces).

Proposition 37.43. *A second-countable topological Hausdorff space, E, is compact iff every sequence, (x_n), of E has some accumulation point in E.*

Proof. Assume that every sequence, (x_n), has some accumulation point. Let $(U_i)_{i \in I}$ be some open cover of E. By Proposition 37.41, there is a countable open subcover, $(O_n)_{n \geq 0}$, for E. Now, if E is not covered by any finite subcover of $(O_n)_{n \geq 0}$, we can define a sequence, (x_m), by induction as follows:

Let x_0 be arbitrary and for every $m \geq 1$, let x_m be some point in E not in $O_1 \cup \cdots \cup O_m$, which exists, since $O_1 \cup \cdots \cup O_m$ is not an open cover of E. We claim that the sequence, (x_m), does not have any accumulation point. Indeed, for every $l \in E$, since $(O_n)_{n \geq 0}$ is an open cover of E, there is some O_m such that $l \in O_m$, and by construction, every x_n with $n \geq m + 1$ does not belong to O_m, which means that $x_n \in O_m$ for only finitely many n and l is not an accumulation point. See Figure 37.39.

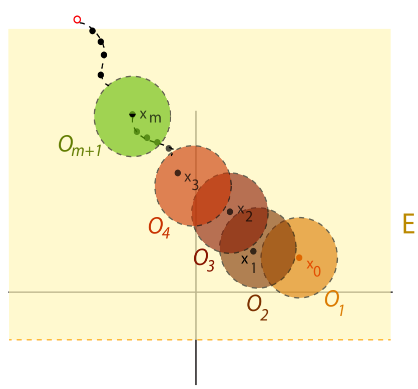


Figure 37.39: The space E is the open half plane above the line $y = -1$. S ince E is not compact, we inductively build a sequence, (x_n) that will have no accumulation point in E. Note the y coordinate of x_n approaches infinity.

Conversely, assume that E is compact, and let (x_n) be any sequence . If $l \in E$ is not an accumulation point of the sequence, then there is some open set, U_l, such that $l \in U_l$

and $x_n \in U_l$ for only finitely many n. Thus, if (x_n) does not have any accumulation point, the family, $(U_l)_{l \in E}$, is an open cover of E and since E is compact, it has some finite open subcover, $(U_l)_{l \in J}$, where J is a finite subset of E. But every U_l with $l \in J$ is such that $x_n \in U_l$ for only finitely many n, and since J is finite, $x_n \in \bigcup_{l \in J} U_l$ for only finitely many n, which contradicts the fact that $(U_l)_{l \in J}$ is an open cover of E, and thus contains all the x_n. Thus, (x_n) has some accumulation point. See Figure 37.40. $\square$

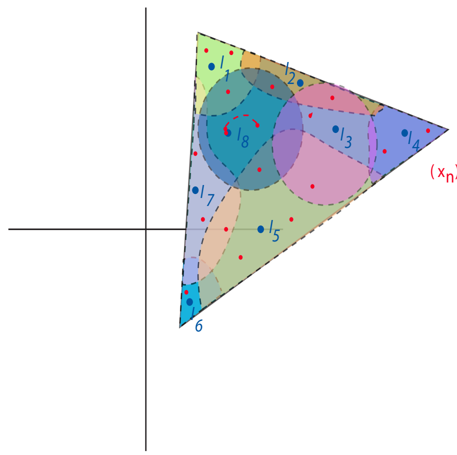


Figure 37.40: The space E the closed triangular region of $\mathbb{R}^2$. Given a sequence (x_n) of red points in E, if the sequence has no accumulation points, then each l_i for $1 \leq i \leq 8$, is not an accumulation point. But as implied by the illustration, l_8 actually is an accumulation point of (x_n).

Remarks:

1. By combining Propositions 37.42 and 37.43, we have observe that a second-countable Hausdorff space E is compact iff every sequence (x_n) has a convergent subsequence (x_{n_k}). In other words, we say a second-countable Hausdorff space E is compact iff it is *sequentially compact*.

2. It should be noted that the proof showing that if E is compact, then every sequence has some accumulation point, holds for any arbitrary compact space (the proof does not use a countable basis for the topology). The converse also holds for metric spaces. We will prove this converse since it is a major property of metric spaces.

Given a metric space in which every sequence has some accumulation point, we first prove the existence of a *Lebesgue number*.

Lemma 37.44. *Given a metric space, E, if every sequence, (x_n), has an accumulation point, for every open cover, $(U_i)_{i\in I}$, of E, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i\in I}$) such that, for every open ball, $B_0(a, \epsilon)$, of radius $\epsilon \leq \delta$, there is some open subset, U_i, such that $B_0(a, \epsilon) \subseteq U_i$. See Figure 37.41*

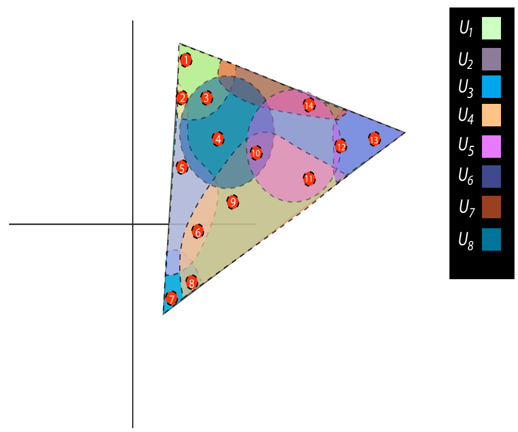


Figure 37.41: The space E the closed triangular region of $\mathbb{R}^2$. It's open cover is $(U_i)_{i=1}^{8}$. The Lebesque number is the radius of the small orange balls labelled 1 through 14. Each open ball of this radius entirely contained within at least one U_i. For example, Ball 2 is contained in both U_1 and U_2.

Proof. If there was no δ with the above property, then, for every natural number, n, there would be some open ball, $B_0(a_n, 1/n)$, which is not contained in any open set, U_i, of the open cover, $(U_i)_{i\in I}$. However, the sequence, (a_n), has some accumulation point, a, and since $(U_i)_{i\in I}$ is an open cover of E, there is some U_i such that $a \in U_i$. Since U_i is open, there is some open ball of center a and radius ϵ contained in U_i. Now, since a is an accumulation point of the sequence, (a_n), every open set containing a contains a_n for infinitely many n and thus, there is some n large enough so that

$$1/n \leq \epsilon/2 \quad \text{and} \quad a_n \in B_0(a, \epsilon/2),$$

which implies that

$$B_0(a_n, 1/n) \subseteq B_0(a, \epsilon) \subseteq U_i,$$

a contradiction. $\square$

By a previous remark, since the proof of Proposition 37.43 implies that in a compact topological space, every sequence has some accumulation point, by Lemma 37.44, in a compact metric space, every open cover has a Lebesgue number. This fact can be used to prove another important property of compact metric spaces, the uniform continuity theorem.

Definition 37.36. Given two metric spaces, (E, d_E) and (F, d_F), a function, $f\colon E \to F$, is *uniformly continuous* if for every $\epsilon > 0$, there is some $\eta > 0$, such that, for all $a, b \in E$,

$$\text{if} \quad d_E(a, b) \leq \eta \quad \text{then} \quad d_F(f(a), f(b)) \leq \epsilon.$$

See Figures 37.42 and 37.43.

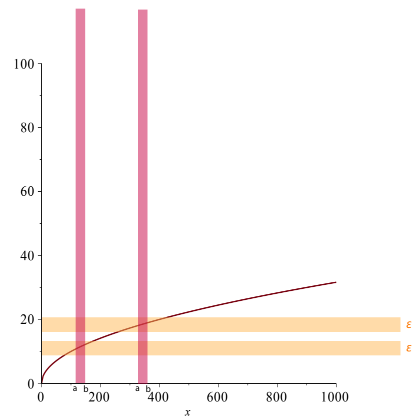


Figure 37.42: The real valued function $f(x) = \sqrt{x}$ is uniformly continuous over $(0, \infty)$. Fix ϵ. If the x values lie within the rose colored η strip, the y values always lie within the peach ϵ strip.

As we saw earlier, the metric on a metric space is uniformly continuous, and the norm on a normed metric space is uniformly continuous.

The *uniform continuity theorem* can be stated as follows:

Theorem 37.45. *Given two metric spaces, (E, d_E) and (F, d_F), if E is compact and if $f\colon E \to F$ is a continuous function, then f is uniformly continuous.*

Proof. Consider any $\epsilon > 0$ and let $(B_0(y, \epsilon/2))_{y \in F}$ be the open cover of F consisting of open balls of radius $\epsilon/2$. Since f is continuous, the family,

$$(f^{-1}(B_0(y, \epsilon/2)))_{y \in F},$$

is an open cover of E. Since, E is compact, by Lemma 37.44, there is a Lebesgue number, δ, such that for every open ball, $B_0(a, \eta)$, of radius $\eta \leq \delta$, then $B_0(a, \eta) \subseteq f^{-1}(B_0(y, \epsilon/2))$, for some $y \in F$. In particular, for any $a, b \in E$ such that $d_E(a, b) \leq \eta = \delta/2$, we have $a, b \in B_0(a, \delta)$ and thus, $a, b \in f^{-1}(B_0(y, \epsilon/2))$, which implies that $f(a), f(b) \in B_0(y, \epsilon/2)$. But then, $d_F(f(a), f(b)) \leq \epsilon$, as desired. $\square$

We now prove another lemma needed to obtain the characterization of compactness in metric spaces in terms of accumulation points.

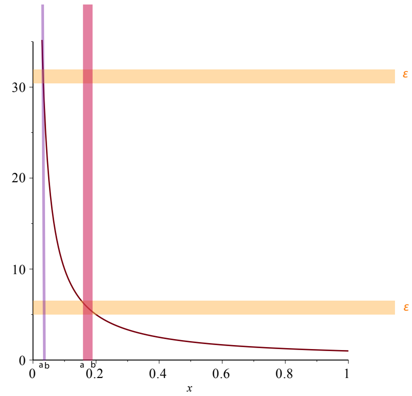


Figure 37.43: The real valued function $f(x) = 1/x$ is not uniformly continuous over $(0, \infty)$. Fix ϵ. In order for the y values to lie within the peach epsilon strip, the widths of the eta strips decrease as $x \to 0$.

Lemma 37.46. *Given a metric space, E, if every sequence, (x_n), has an accumulation point, then for every $\epsilon > 0$, there is a finite open cover, $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, of E by open balls of radius ϵ.*

Proof. Let a_0 be any point in E. If $B_0(a_0, \epsilon) = E$, then the lemma is proved. Otherwise, assume that a sequence, $(a_0, a_1, \ldots, a_n)$, has been defined, such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ does not cover E. Then, there is some a_{n+1} not in $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ and either

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon) = E,$$

in which case the lemma is proved, or we obtain a sequence, $(a_0, a_1, \ldots, a_{n+1})$, such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon)$ does not cover E. If this process goes on forever, we obtain an infinite sequence, (a_n), such that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. Since every sequence in E has some accumulation point, the sequence, (a_n), has some accumulation point, a. Then, for infinitely many n, we must have $d(a_n, a) \leq \epsilon/3$ and thus, for at least two distinct natural numbers, p, q, we must have $d(a_p, a) \leq \epsilon/3$ and $d(a_q, a) \leq \epsilon/3$, which implies $d(a_p, a_q) \leq d(a_p, a) + d(a_q, a) \leq 2\epsilon/3$, contradicting the fact that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. See Figure 37.44. Thus, there must be some n such that

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon) = E. \qquad \square$$

Definition 37.37. A metric space E is said to be *precompact* (or *totally bounded*) if for every $\epsilon > 0$, there is a finite open cover, $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, of E by open balls of radius ϵ.

We now obtain the *Weierstrass–Bolzano* property.

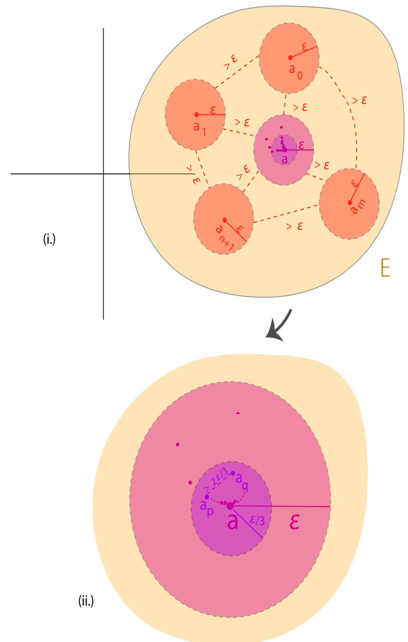


Figure 37.44: Let E be the peach region of $\mathbb{R}^2$. If E is not covered by a finite collection of orange balls with radius ϵ, the points of the sequence (a_n) are separated by a distance of at least ϵ. This contradicts the fact that a is the accumulation point of a, as evidenced by the enlargement of the plum disk in Figure (ii).

Theorem 37.47. *A metric space, E, is compact iff every sequence, (x_n), has an accumulation point.*

Proof. We already observed that the proof of Proposition 37.43 shows that for any compact space (not necessarily metric), every sequence, (x_n), has an accumulation point. Conversely, let E be a metric space, and assume that every sequence, (x_n), has an accumulation point. Given any open cover, $(U_i)_{i\in I}$ for E, we must find a finite open subcover of E. By Lemma 37.44, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i\in I}$) such that, for every open ball, $B_0(a, \epsilon)$, of radius $\epsilon \leq \delta$, there is some open subset, U_j, such that $B_0(a, \epsilon) \subseteq U_j$. By Lemma 37.46, for every $\delta > 0$, there is a finite open cover, $B_0(a_0, \delta) \cup \cdots \cup B_0(a_n, \delta)$, of E by open balls of radius δ. But from the previous statement, every open ball, $B_0(a_i, \delta)$, is contained in some open set, U_{j_i}, and thus, $\{U_{j_1}, \ldots, U_{j_n}\}$ is an open cover of E. $\square$

37.8 Complete Metric Spaces and Compactness

Another very useful characterization of compact metric spaces is obtained in terms of Cauchy sequences. Such a characterization is quite useful in fractal geometry (and elsewhere). First recall the definition of a Cauchy sequence and of a complete metric space.

Definition 37.38. Given a metric space, (E, d), a sequence, $(x_n)_{n\in\mathbb{N}}$, in E is a *Cauchy sequence* if the following condition holds: for every $\epsilon > 0$, there is some $p \geq 0$, such that, for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon$.

If every Cauchy sequence in (E, d) converges we say that (E, d) is a *complete metric space*.

First let us show the following proposition:

Proposition 37.48. *Given a metric space, E, if a Cauchy sequence, (x_n), has some accumulation point, a, then a is the limit of the sequence, (x_n).*

Proof. Since (x_n) is a Cauchy sequence, for every $\epsilon > 0$, there is some $p \geq 0$, such that, for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon/2$. Since a is an accumulation point for (x_n), for infinitely many n, we have $d(x_n, a) \leq \epsilon/2$, and thus, for at least some $n \geq p$, we have $d(x_n, a) \leq \epsilon/2$. Then, for all $m \geq p$,

$$d(x_m, a) \leq d(x_m, x_n) + d(x_n, a) \leq \epsilon,$$

which shows that a is the limit of the sequence (x_n). ☐

We can now prove the following theorem.

Theorem 37.49. *A metric space, E, is compact iff it is precompact and complete.*

Proof. Let E be compact. For every $\epsilon > 0$, the family of all open balls of radius ϵ is an open cover for E and since E is compact, there is a finite subcover, $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, of E by open balls of radius ϵ. Thus E is precompact. Since E is compact, by Theorem 37.47, every sequence, (x_n), has some accumulation point. Thus every Cauchy sequence, (x_n), has some accumulation point, a, and, by Proposition 37.48, a is the limit of (x_n). Thus, E is complete.

Now assume that E is precompact and complete. We prove that every sequence, (x_n), has an accumulation point. By the other direction of Theorem 37.47, this shows that E is compact. Given any sequence, (x_n), we construct a Cauchy subsequence, (y_n), of (x_n) as follows: Since E is precompact, letting $\epsilon = 1$, there exists a finite cover, $\mathcal{U}_1$, of E by open balls of radius 1. Thus some open ball, B_o^0, in the cover, $\mathcal{U}_1$, contains infinitely many elements from the sequence (x_n). Let y_0 be any element of (x_n) in B_o^0. By induction, assume that a sequence of open balls, $(B_o^i)_{1\leq i\leq m}$, has been defined, such that every ball, B_o^i, has radius $\frac{1}{2^i}$, contains infinitely many elements from the sequence (x_n) and contains some y_i from (x_n) such that

$$d(y_i, y_{i+1}) \leq \frac{1}{2^i},$$

for all i, $0 \leq i \leq m-1$. See Figure 37.45. Then letting $\epsilon = \frac{1}{2^{m+1}}$, because E is precompact, there is some finite cover, $\mathcal{U}_{m+1}$, of E by open balls of radius ϵ and thus, of the open ball B_o^m.

Thus, some open ball, B_o^{m+1}, in the cover, $\mathcal{U}_{m+1}$, contains infinitely many elements from the sequence, (x_n), and we let y_{m+1} be any element of (x_n) in B_o^{m+1}. Thus, we have defined by induction a sequence, (y_n), which is a subsequence of, (x_n), and such that

$$d(y_i, y_{i+1}) \leq \frac{1}{2^i},$$

for all i. However, for all $m, n \geq 1$, we have

$$d(y_m, y_n) \leq d(y_m, y_{m+1}) + \cdots + d(y_{n-1}, y_n) \leq \sum_{i=m}^{n} \frac{1}{2^i} \leq \frac{1}{2^{m-1}},$$

and thus, (y_n) is a Cauchy sequence Since E is complete, the sequence, (y_n), has a limit, and since it is a subsequence of (x_n), the sequence, (x_n), has some accumulation point. $\square$

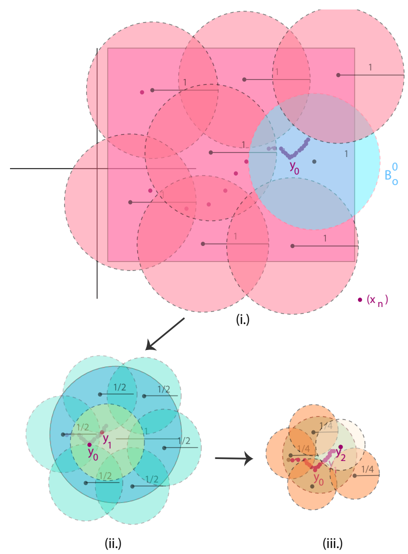


Figure 37.45: The first three stages of the construction of the Cauchy sequence (y_n), where E is the pink square region of $\mathbb{R}^2$. The original sequence (x_n) is illustrated with plum colored dots. Figure $(i.)$ covers E with ball of radius 1 and shows the selection of B_o^0 and y_0. Figure $(ii.)$ covers B_o^0 with balls of radius $1/2$ and selects the yellow ball as B_o^1 with point y_1. Figure $(iii.)$ covers B_o^1 with balls of radius $1/4$ and selects the pale peach ball as B_o^2 with point y_2.

Another useful property of a complete metric space is that a subset is closed iff it is complete. This is shown in the following two propositions.

Proposition 37.50. *Let (E, d) be a metric space, and let A be a subset of E. If A is complete (which means that every Cauchy sequence of elements in A converges to some point of A), then A is closed in E.*

Proof. Assume $x \in \overline{A}$. By Proposition 37.13, there is some sequence (a_n) of points $a_n \in A$ which converges to x. Consequently (a_n) is a Cauchy sequence in E, and thus a Cauchy sequence in A (since $a_n \in A$ for all n). Since A is complete, the sequence (a_n) has a limit $a \in A$, but since E is a metric space it is Hausdorff, so $a = x$, which shows that $x \in A$; that is, A is closed. $\square$

Proposition 37.51. *Let (E, d) be a metric space, and let A be a subset of E. If E is complete and if A is closed in E, then A is complete.*

Proof. Let (a_n) be a Cauchy sequence in A. The sequence (a_n) is also a Cauchy sequence in E, and since E is complete, it has a limit $x \in E$. But $a_n \in A$ for all n, so by Proposition 37.13 we must have $x \in \overline{A}$. Since A is closed, actually $x \in A$, which proves that A is complete. $\square$

An arbitrary metric space (E, d) is not necessarily complete, but there is a construction of a metric space $(\widehat{E}, \widehat{d})$ such that $\widehat{E}$ is complete, and there is a continuous (injective) distance-preserving map $\varphi\colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$. This is a generalization of the construction of the set $\mathbb{R}$ of real numbers from the set $\mathbb{Q}$ of rational numbers in terms of Cauchy sequences. This construction can be immediately adapted to a normed vector space $(E, \| \|)$ to embed $(E, \| \|)$ into a complete normed vector space $(\widehat{E}, \| \|_{\widehat{E}})$ (a Banach space). This construction is used heavily in integration theory, where E is a set of functions.

37.9 Completion of a Metric Space

In order to prove a kind of uniqueness result for the completion $(\widehat{E}, \widehat{d})$ of a metric space (E, d), we need the following result about extending a uniformly continuous function.

Recall that E_0 is dense in E iff $\overline{E_0} = E$. Since E is a metric space, by Proposition 37.13, this means that for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$.

Theorem 37.52. *Let E and F be two metric spaces, let E_0 be a dense subspace of E, and let $f_0\colon E_0 \to F$ be a continuous function. If f_0 is uniformly continuous and if F is complete, then there is a unique uniformly continuous function $f\colon E \to F$ extending f_0.*

Proof. We follow Schwartz's proof; see Schwartz [149] (Chapter XI, Section 3, Theorem 1).

Step 1. We begin by constructing a function $f\colon E \to F$ extending f_0. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Then the sequence (x_n) is a Cauchy sequence in E. We claim that $(f_0(x_n))$ is a Cauchy sequence in F.

Proof of the claim. For every $\epsilon > 0$, since f_0 is uniformly continuous, there is some $\eta > 0$ such that for all $(y, z) \in E_0$, if $d(y, z) \leq \eta$, then $d(f_0(y), f_0(z)) \leq \epsilon$. Since (x_n) is a Cauchy sequence with $x_n \in E_0$, there is some integer $p > 0$ such that if $m, n \geq p$, then $d(x_m, x_n) \leq \eta$, thus $d(f_0(x_m), f_0(x_n)) \leq \epsilon$, which proves that $(f_0(x_n))$ is a Cauchy sequence in F. $\square$

Since F is complete and $(f_0(x_n))$ is a Cauchy sequence in F, the sequence $(f_0(x_n))$ converges to some element of F; denote this element by $f(x)$.

Step 2. Let us now show that $f(x)$ does not depend on the sequence (x_n) converging to x. Suppose that (x'_n) and (x''_n) are two sequences of elements in E_0 converging to x. Then the mixed sequence

$$x'_0, x''_0, x'_1, x''_1, \ldots, x'_n, x''_n, \ldots,$$

also converges to x. It follows that the sequence

$$f_0(x'_0), f_0(x''_0), f_0(x'_1), f_0(x''_1), \ldots, f_0(x'_n), f_0(x''_n), \ldots,$$

is a Cauchy sequence in F, and since F is complete, it converges to some element of F, which implies that the sequences $(f_0(x'_n))$ and $(f_0(x''_n))$ converge to the same limit.

As a summary, we have defined a function $f\colon E \to F$ by

$$f(x) = \lim_{n \mapsto \infty} f_0(x_n).$$

for any sequence (x_n) converging to x, with $x_n \in E_0$.

Step 3. The function f extends f_0. Since every element $x \in E_0$ is the limit of the constant sequence (x_n) with $x_n = x$ for all $n \geq 0$, by definition $f(x)$ is the limit of the sequence $(f_0(x_n))$, which is the constant sequence with value $f_0(x)$, so $f(x) = f_0(x)$; that is, f extends f_0.

Step 4. We now prove that f is uniformly continuous. Since f_0 is uniformly continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that if $a, b \in E_0$ and $d(a, b) \leq \eta$, then $d(f_0(a), f_0(b)) \leq \epsilon$. Consider any two points $x, y \in E$ such that $d(x, y) \leq \eta/2$. We claim that $d(f(x), f(y)) \leq \epsilon$, which shows that f is uniformly continuous.

Let (x_n) be a sequence of points in E_0 converging to x, and let (y_n) be a sequence of points in E_0 converging to y. By the triangle inequality,

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n) = d(x, y) + d(x_n, x) + d(y_n, y),$$

and since (x_n) converges to x and (y_n) converges to y, there is some integer $p > 0$ such that for all $n \geq p$, we have $d(x_n, x) \leq \eta/4$ and $d(y_n, y) \leq \eta/4$, and thus

$$d(x_n, y_n) \leq d(x, y) + \frac{\eta}{2}.$$

Since we assumed that $d(x, y) \leq \eta/2$, we get $d(x_n, y_n) \leq \eta$ for all $n \geq p$, and by uniform continuity of f_0, we get

$$d(f_0(x_n), f_0(y_n)) \leq \epsilon$$

for all $n \geq p$. Since the distance function on F is also continuous, and since $(f_0(x_n))$ converges to $f(x)$ and $(f_0(y_n))$ converges to $f(y)$, we deduce that the sequence $(d(f_0(x_n), f_0(y_n)))$ converges to $d(f(x), f(y))$. This implies that $d(f(x), f(y)) \leq \epsilon$, as desired.

Step 5. It remains to prove that f is unique. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Since f extends f_0 and since f is continuous, we get

$$f(x) = \lim_{n \mapsto \infty} f_0(x_n),$$

which only depends on f_0 and x, and shows that f is unique. $\square$

Remark: It can be shown that the theorem no longer holds if we either omit the hypothesis that F is complete or omit that f_0 is uniformly continuous.

For example, if $E_0 \neq E$ and if we let $F = E_0$ and f_0 be the identity function, it is easy to see that f_0 cannot be extended to a continuous function from E to E_0 (for any $x \in E - E_0$, any continous extension f of f_0 would satisfy $f(x) = x$, which is absurd since $x \notin E_0$).

If f_0 is continuous but not uniformly continuous, a counter-example can be given by using $E = \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ made into a metric space, $E_0 = \mathbb{R}$, $F = \mathbb{R}$, and f_0 the identity function; for details, see Schwartz [149] (Chapter XI, Section 3, page 134).

Definition 37.39. If (E, d_E) and (F, d_F) are two metric spaces, then a function $f \colon E \to F$ is *distance-preserving*, or an *isometry*, if

$$d_F(f(x), f(y)) = d_E(x, y), \quad \text{for all for all } x, y \in E.$$

Observe that an isometry must be injective, because if $f(x) = f(y)$, then $d_F(f(x), f(y)) = 0$, and since $d_F(f(x), f(y)) = d_E(x, y)$, we get $d_E(x, y) = 0$, but $d_E(x, y) = 0$ implies that $x = y$. Also, an isometry is uniformly continuous (since we can pick $\eta = \epsilon$ to satisfy the condition of uniform continuity). However, an isometry is not necessarily surjective.

We now give a construction of the completion of a metric space. This construction is just a generalization of the classical construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences.

Theorem 37.53. *Let (E, d) be any metric space. There is a complete metric space $(\widehat{E}, \widehat{d})$ called a completion of (E, d), and a distance-preserving (uniformly continuous) map $\varphi \colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$, and the following extension property holds: for every complete metric space F and for every uniformly continuous function $f \colon E \to F$, there is a unique uniformly continuous function $\widehat{f} \colon \widehat{E} \to F$ such that*

$$f = \widehat{f} \circ \varphi,$$

as illustrated in the following diagram.

$$\begin{tikzcd} E \arrow[r, "\varphi"] \arrow[dr, "f"'] & \widehat{E} \arrow[d, "\widehat{f}"] \\ & F. \end{tikzcd}$$

As a consequence, for any two completions $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ of (E, d), there is a unique bijective isometry betwen $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$.

Proof. Consider the set $\mathcal{E}$ of all Cauchy sequences (x_n) in E, and define the relation $\sim$ on $\mathcal{E}$ as follows:

$$(x_n) \sim (y_n) \quad \text{iff} \quad \lim_{n\mapsto\infty} d(x_n, y_n) = 0.$$

It is easy to check that $\sim$ is an equivalence relation on $\mathcal{E}$, and let $\widehat{E} = \mathcal{E}/\sim$ be the quotient set, that is, the set of equivalence classes modulo $\sim$. Our goal is to show that we can endow $\widehat{E}$ with a distance that makes it into a complete metric space satisfying the conditions of the theorem. We proceed in several steps.

Step 1. First, let us construct the function $\varphi\colon E \to \widehat{E}$. For every $a \in E$, we have the constant sequence (a_n) such that $a_n = a$ for all $n \geq 0$, which is obviously a Cauchy sequence. Let $\varphi(a) \in \widehat{E}$ be the equivalence class $[(a_n)]$ of the constant sequence (a_n) with $a_n = a$ for all n. By definition of $\sim$, the equivalence class $\varphi(a)$ is also the equivalence class of all sequences converging to a. The map $a \mapsto \varphi(a)$ is injective because a metric space is Hausdorff, so if $a \neq b$, then a sequence converging to a does not converge to b. After having defined a distance on $\widehat{E}$, we will check that φ is an isometry.

Step 2. Let us now define a distance on $\widehat{E}$. Let $\alpha = [(a_n)]$ and $\beta = [(b_n)]$ be two equivalence classes of Cauchy sequences in E. The triangle inequality implies that

$$d(a_m, b_m) \leq d(a_m, a_n) + d(a_n, b_n) + d(b_n, b_m) = d(a_n, b_n) + d(a_m, a_n) + d(b_m, b_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a_m) + d(a_m, b_m) + d(b_m, b_n) = d(a_m, b_m) + d(a_m, a_n) + d(b_m, b_n),$$

which implies that

$$|d(a_m, b_m) - d(a_n, b_n)| \leq d(a_m, a_n) + d(b_m, b_n).$$

Since (a_n) and (b_n) are Cauchy sequences, it follows that $(d(a_n, b_n))$ is a Cauchy sequence of nonnegative reals. Since $\mathbb{R}$ is complete, the sequence $(d(a_n, b_n))$ has a limit, which we denote by $\widehat{d}(\alpha, \beta)$; that is, we set

$$\widehat{d}(\alpha, \beta) = \lim_{n\mapsto\infty} d(a_n, b_n), \quad \alpha = [(a_n)],\ \beta = [(b_n)].$$

Step 3. Let us check that $\widehat{d}(\alpha, \beta)$ does not depend on the Cauchy sequences (a_n) and (b_n) chosen in the equivalence classes α and β.

If $(a_n) \sim (a'_n)$ and $(b_n) \sim (b'_n)$, then $\lim_{n\mapsto\infty} d(a_n, a'_n) = 0$ and $\lim_{n\mapsto\infty} d(b_n, b'_n) = 0$, and since

$$d(a'_n, b'_n) \leq d(a'_n, a_n) + d(a_n, b_n) + d(b_n, b'_n) = d(a_n, b_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a'_n) + d(a'_n, b'_n) + d(b'_n, b_n) = d(a'_n, b'_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

we have

$$|d(a_n, b_n) - d(a'_n, b'_n)| \leq d(a_n, a'_n) + d(b_n, b'_n),$$

so we have $\lim_{n\mapsto\infty} d(a'_n, b'_n) = \lim_{n\mapsto\infty} d(a_n, b_n) = \widehat{d}(\alpha, \beta)$. Therefore, $\widehat{d}(\alpha, \beta)$ is indeed well defined.

Step 4. Let us check that φ is indeed an isometry.

Given any two elements $\varphi(a)$ and $\varphi(b)$ in $\widehat{E}$, since they are the equivalence classes of the constant sequences (a_n) and (b_n) such that $a_n = a$ and $b_n = b$ for all n, the constant sequence $(d(a_n, b_n))$ with $d(a_n, b_n) = d(a, b)$ for all n converges to $d(a, b)$, so by definition $\widehat{d}(\varphi(a), \varphi(b)) = \lim_{n\mapsto\infty} d(a_n, b_n) = d(a, b)$, which shows that φ is an isometry.

Step 5. Let us verify that $\widehat{d}$ is a metric on $\widehat{E}$. By definition it is obvious that $\widehat{d}(\alpha, \beta) = \widehat{d}(\beta, \alpha)$. If α and β are two distinct equivalence classes, then for any Cauchy sequence (a_n) in the equivalence class α and for any Cauchy sequence (b_n) in the equivalence class β, the sequences (a_n) and (b_n) are inequivalent, which means that $\lim_{n\mapsto\infty} d(a_n, b_n) \neq 0$, that is, $\widehat{d}(\alpha, \beta) \neq 0$. Obviously, $\widehat{d}(\alpha, \alpha) = 0$.

For any equivalence classes $\alpha = [(a_n)]$, $\beta = [(b_n)]$, and $\gamma = [(c_n)]$, we have the triangle inequality

$$d(a_n, c_n) \leq d(a_n, b_n) + d(b_n, c_n),$$

so by continuity of the distance function, by passing to the limit, we obtain

$$\widehat{d}(\alpha, \gamma) \leq \widehat{d}(\alpha, \beta) + \widehat{d}(\beta, \gamma),$$

which is the triangle inequality for $\widehat{d}$. Therefore, $\widehat{d}$ is a distance on $\widehat{E}$.

Step 6. Let us prove that $\varphi(E)$ is dense in $\widehat{E}$. For any $\alpha = [(a_n)]$, let (x_n) be the constant sequence such that $x_k = a_n$ for all $k \geq 0$, so that $\varphi(a_n) = [(x_n)]$. Then we have

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m\mapsto\infty} d(a_m, a_n) \leq \sup_{p,q\geq n} d(a_p, a_q).$$

Since (a_n) is a Cauchy sequence, $\sup_{p,q\geq n} d(a_p, a_q)$ tends to 0 as n goes to infinity, so

$$\lim_{n\mapsto\infty} d(\alpha, \varphi(a_n)) = 0,$$

which means that the sequence $(\varphi(a_n))$ converge to α, and $\varphi(E)$ is indeed dense in $\widehat{E}$.

Step 7. Finally, let us prove that the metric space $\widehat{E}$ is complete.

Let (α_n) be a Cauchy sequence in $\widehat{E}$. Since $\varphi(E)$ is dense in $\widehat{E}$, for every $n > 0$, there some $a_n \in E$ such that

$$\widehat{d}(\alpha_n, \varphi(a_n)) \leq \frac{1}{n}.$$

Since

$$\widehat{d}(\varphi(a_m), \varphi(a_n)) \leq \widehat{d}(\varphi(a_m), \alpha_m) + \widehat{d}(\alpha_m, \alpha_n) + \widehat{d}(\alpha_n, \varphi(a_n)) \leq \widehat{d}(\alpha_m, \alpha_n) + \frac{1}{m} + \frac{1}{n},$$

and since (α_m) is a Cauchy sequence, so is $(\varphi(a_n))$, and as φ is an isometry, the sequence (a_n) is a Cauchy sequence in E. Let $\alpha \in \widehat{E}$ be the equivalence class of (a_n). Since

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m \mapsto \infty} d(a_m, a_n)$$

and (a_n) is a Cauchy sequence, we deduce that the sequence $(\varphi(a_n))$ converges to α, and since $d(\alpha_n, \varphi(a_n)) \leq 1/n$ for all $n > 0$, the sequence (α_n) also converges to α.

Step 8. Let us prove the extension property. Let F be any complete metric space and let $f \colon E \to F$ be any uniformly continuous function. The function $\varphi \colon E \to \widehat{E}$ is an isometry and a bijection between E and its image $\varphi(E)$, so its inverse $\varphi^{-1} \colon \varphi(E) \to E$ is also an isometry, and thus is uniformly continuous. If we let $g = f \circ \varphi^{-1}$, then $g \colon \varphi(E) \to F$ is a uniformly continuous function, and $\varphi(E)$ is dense in $\widehat{E}$, so by Theorem 37.52 there is a unique uniformly continuous function $\widehat{f} \colon \widehat{E} \to F$ extending $g = f \circ \varphi^{-1}$; see the diagram below:

$$\begin{tikzcd}
E \arrow[drr, "f"'] & \varphi(E) \arrow[l, "\varphi^{-1}"'] \arrow[dr, "g"] & \subseteq & \widehat{E} \arrow[dl, "\widehat{f}"] \\
 & & F &
\end{tikzcd}$$

This means that

$$\widehat{f}|\varphi(E) = f \circ \varphi^{-1},$$

which implies that

$$(\widehat{f}|\varphi(E)) \circ \varphi = f,$$

that is, $f = \widehat{f} \circ \varphi$, as illustrated in the diagram below:

$$\begin{tikzcd}
E \arrow[r, "\varphi"] \arrow[dr, "f"'] & \widehat{E} \arrow[d, "\widehat{f}"] \\
 & F
\end{tikzcd}$$

If $h \colon \widehat{E} \to F$ is any other uniformly continuous function such that $f = h \circ \varphi$, then $g = f \circ \varphi^{-1} = h|\varphi(E)$, so h is a uniformly continuous function extending g, and by Theorem 37.52, we have have $h = \widehat{f}$, so $\widehat{f}$ is indeed unique.

Step 9. Uniqueness of the completion $(\widehat{E}, \widehat{d})$ up to a bijective isometry.

Let $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ be any two completions of (E, d). Then we have two uniformly continuous isometries $\varphi_1 \colon E \to \widehat{E}_1$ and $\varphi_2 \colon E \to \widehat{E}_2$, so by the unique extension property, there exist unique uniformly continuous maps $\widehat{\varphi_2} \colon \widehat{E}_1 \to \widehat{E}_2$ and $\widehat{\varphi_1} \colon \widehat{E}_2 \to \widehat{E}_1$ such that the following diagrams commute:

$$\begin{tikzcd}
E \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_2"'] & \widehat{E}_1 \arrow[d, "\widehat{\varphi_2}"] \\
 & \widehat{E}_2
\end{tikzcd}
\qquad
\begin{tikzcd}
E \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_1"'] & \widehat{E}_2 \arrow[d, "\widehat{\varphi_1}"] \\
 & \widehat{E}_1.
\end{tikzcd}$$

Consequently we have the following commutative diagrams:

$$
\begin{tikzcd}
 & \widehat{E}_2 \arrow[d, "\widehat{\varphi_1}"] \\
E \arrow[ur, "\varphi_2"] \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_2"'] & \widehat{E}_1 \arrow[d, "\widehat{\varphi_2}"] \\
 & \widehat{E}_2
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \widehat{E}_1 \arrow[d, "\widehat{\varphi_2}"] \\
E \arrow[ur, "\varphi_1"] \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_1"'] & \widehat{E}_2 \arrow[d, "\widehat{\varphi_1}"] \\
 & \widehat{E}_1.
\end{tikzcd}
$$

However, $\mathrm{id}_{\widehat{E}_1}$ and $\mathrm{id}_{\widehat{E}_2}$ are uniformly continuous functions making the following diagrams commute

$$
\begin{tikzcd}
E \arrow[r, "\varphi_1"] \arrow[dr, "\varphi_1"'] & \widehat{E}_1 \arrow[d, "\mathrm{id}_{\widehat{E}_1}"] \\
 & \widehat{E}_1
\end{tikzcd}
\qquad
\begin{tikzcd}
E \arrow[r, "\varphi_2"] \arrow[dr, "\varphi_2"'] & \widehat{E}_2 \arrow[d, "\mathrm{id}_{\widehat{E}_2}"] \\
 & \widehat{E}_2,
\end{tikzcd}
$$

so by the uniqueness of extensions we must have

$$\widehat{\varphi_1} \circ \widehat{\varphi_2} = \mathrm{id}_{\widehat{E}_1} \quad \text{and} \quad \widehat{\varphi_2} \circ \widehat{\varphi_1} = \mathrm{id}_{\widehat{E}_2}.$$

This proves that $\widehat{\varphi_1}$ and $\widehat{\varphi_2}$ are mutual inverses. Now, since $\varphi_2 = \widehat{\varphi_2} \circ \varphi_1$, we have

$$\widehat{\varphi_2} | \varphi_1(E) = \varphi_2 \circ \varphi_1^{-1},$$

and since φ_1^{-1} and φ_2 are isometries, so is $\widehat{\varphi_2}|\varphi_1(E)$. But we saw earlier that $\widehat{\varphi_2}$ is the uniform continuous extension of $\widehat{\varphi_2}|\varphi_1(E)$ and $\varphi_1(E)$ is dense in $\widehat{E}_1$, so for any two elements $\alpha, \beta \in \widehat{E}_1$, if (a_n) and (b_n) are sequences in $\varphi_1(E)$ converging to α and β, we have

$$\widehat{d}_2((\widehat{\varphi_2}|\varphi_1(E))(a_n), ((\widehat{\varphi_2}|\varphi_1(E))(b_n)) = \widehat{d}_1(a_n, b_n),$$

and by passing to the limit we get

$$\widehat{d}_2(\widehat{\varphi_2}(\alpha), \widehat{\varphi_2}(\beta)) = \widehat{d}_1(\alpha, \beta),$$

which shows that $\widehat{\varphi_2}$ is an isometry (similarly, $\widehat{\varphi_1}$ is an isometry). $\square$

Remarks:

1. Except for Step 8 and Step 9, the proof of Theorem 37.53 is the proof given in Schwartz [149] (Chapter XI, Section 4, Theorem 1), and Kormogorov and Fomin [105] (Chapter 2, Section 7, Theorem 4).

2. The construction of $\widehat{E}$ relies on the completeness of $\mathbb{R}$, and so it cannot be used to construct $\mathbb{R}$ from $\mathbb{Q}$. However, this construction can be modified to yield a construction of $\mathbb{R}$ from $\mathbb{Q}$.

We show in Section 37.12 that Theorem 37.53 yields a construction of the completion of a normed vector space.

37.10 The Contraction Mapping Theorem

If (E, d) is a nonempty complete metric space, every map, $f\colon E \to E$, for which there is some k such that $0 \leq k < 1$ and

$$d(f(x), f(y)) \leq kd(x, y)$$

for all $x, y \in E$, has the very important property that it has a unique fixed point, that is, there is a unique, $a \in E$, such that $f(a) = a$. A map as above is called a *contraction mapping*. Furthermore, the fixed point of a contraction mapping can be computed as the limit of a fast converging sequence.

The fixed point property of contraction mappings is used to show some important theorems of analysis, such as the implicit function theorem and the existence of solutions to certain differential equations. It can also be used to show the existence of fractal sets defined in terms of iterated function systems. Since the proof is quite simple, we prove the fixed point property of contraction mappings. First, observe that a contraction mapping is (uniformly) continuous.

Proposition 37.54. *If (E, d) is a nonempty complete metric space, every contraction mapping, $f\colon E \to E$, has a unique fixed point. Furthermore, for every $x_0 \in E$, defining the sequence, (x_n), such that $x_{n+1} = f(x_n)$, the sequence, (x_n), converges to the unique fixed point of f.*

Proof. First we prove that f has at most one fixed point. Indeed, if $f(a) = a$ and $f(b) = b$, since

$$d(a, b) = d(f(a), f(b)) \leq kd(a, b)$$

and $0 \leq k < 1$, we must have $d(a, b) = 0$, that is, $a = b$.

Next, we prove that (x_n) is a Cauchy sequence. Observe that

$$\begin{aligned}
d(x_2, x_1) &\leq kd(x_1, x_0), \\
d(x_3, x_2) &\leq kd(x_2, x_1) \leq k^2 d(x_1, x_0), \\
\vdots \quad & \qquad \vdots \\
d(x_{n+1}, x_n) &\leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n) \\
&\leq (k^{p-1} + k^{p-2} + \cdots + k + 1)k^n d(x_1, x_0) \\
&\leq \frac{k^n}{1-k}\, d(x_1, x_0).
\end{aligned}$$

We conclude that $d(x_{n+p}, x_n)$ converges to 0 when n goes to infinity, which shows that (x_n) is a Cauchy sequence. Since E is complete, the sequence (x_n) has a limit, a. Since f is continuous, the sequence $(f(x_n))$ converges to $f(a)$. But $x_{n+1} = f(x_n)$ converges to a and so $f(a) = a$, the unique fixed point of f. $\square$

Note that no matter how the starting point x_0 of the sequence (x_n) is chosen, (x_n) converges to the unique fixed point of f. Also, the convergence is fast, since

$$d(x_n, a) \leq \frac{k^n}{1-k}\, d(x_1, x_0).$$

The Hausdorff distance between compact subsets of a metric space provides a very nice illustration of some of the theorems on complete and compact metric spaces just presented.

Definition 37.40. Given a metric space, (X, d), for any subset, $A \subseteq X$, for any, $\epsilon \geq 0$, define the *ϵ-hull* of A as the set

$$V_\epsilon(A) = \{x \in X,\ \exists a \in A \mid d(a, x) \leq \epsilon\}.$$

See Figure 37.46. Given any two nonempty bounded subsets, A, B of X, define $D(A, B)$, *the Hausdorff distance between A and B*, by

$$D(A, B) = \inf\{\epsilon \geq 0 \mid A \subseteq V_\epsilon(B) \text{ and } B \subseteq V_\epsilon(A)\}.$$

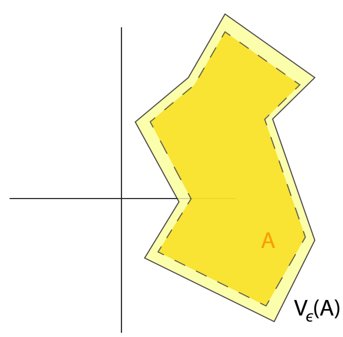


Figure 37.46: The ϵ-hull of a polygonal region A of $\mathbb{R}^2$

Note that since we are considering nonempty bounded subsets, $D(A, B)$ is well defined (i.e., not infinite). However, D is not necessarily a distance function. It is a distance function if we restrict our attention to nonempty compact subsets of X (actually, it is also a metric on closed and bounded subsets). We let $\mathcal{K}(X)$ denote the set of all nonempty compact subsets of X. The remarkable fact is that D is a distance on $\mathcal{K}(X)$ and that if X is complete or compact, then so is $\mathcal{K}(X)$. The following theorem is taken from Edgar [55].

Theorem 37.55. *If (X, d) is a metric space, then the Hausdorff distance, D, on the set, $\mathcal{K}(X)$, of nonempty compact subsets of X is a distance. If (X, d) is complete, then $(\mathcal{K}(X), D)$ is complete and if (X, d) is compact, then $(\mathcal{K}(X), D)$ is compact.*

Proof. Since (nonempty) compact sets are bounded, $D(A, B)$ is well defined. Clearly D is symmetric. Assume that $D(A, B) = 0$. Then for every $\epsilon > 0$, $A \subseteq V_\epsilon(B)$, which means that for every $a \in A$, there is some $b \in B$ such that $d(a, b) \leq \epsilon$, and thus, that $A \subseteq \overline{B}$. Since Proposition 37.26 implies that B is closed, $\overline{B} = B$, and we have $A \subseteq B$. Similarly, $B \subseteq A$, and thus, $A = B$. Clearly, if $A = B$, we have $D(A, B) = 0$. It remains to prove the triangle inequality. Assume that $D(A, B) \leq \epsilon_1$ and that $D(B, C) \leq \epsilon_2$. We must show that $D(A, C) \leq \epsilon_1 + \epsilon_2$. This will be accomplished if we can show that $C \subseteq V_{\epsilon_1+\epsilon_2}(A)$ and $A \subseteq V_{\epsilon_1+\epsilon_2}(C)$. By assumption and definition of D, $B \subseteq V_{\epsilon_1}(A)$ and $C \subseteq V_{\epsilon_2}(B)$. Then

$$V_{\epsilon_2}(B) \subseteq V_{\epsilon_2}(V_{\epsilon_1}(A)),$$

and since a basic application of the triangle inequality implies that

$$V_{\epsilon_2}(V_{\epsilon_1}(A)) \subseteq V_{\epsilon_1+\epsilon_2}(A),$$

we get

$$C \subseteq V_{\epsilon_2}(B) \subseteq V_{\epsilon_1+\epsilon_2}(A).$$

See Figure 37.47.

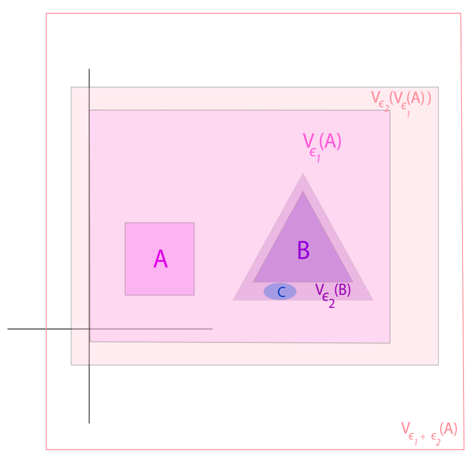


Figure 37.47: Let A be the small pink square and B be the small purple triangle in $\mathbb{R}^2$. The periwinkle oval C is contained in $V_{\epsilon_1+\epsilon_2}(A)$.

Similarly, the conditions $(A, B) \leq \epsilon_1$ and $D(B, C) \leq \epsilon_2$ imply that

$$A \subseteq V_{\epsilon_1}(B), \qquad B \subseteq V_{\epsilon_2}(C).$$

Hence

$$A \subseteq V_{\epsilon_1}(B) \subseteq V_{\epsilon_1}(V_{\epsilon_2}(C)) \subseteq V_{\epsilon_1+\epsilon_2}(C),$$

and thus the triangle inequality follows.

Next we need to prove that if (X, d) is complete, then $(\mathcal{K}(X), D)$ is also complete. First we show that if (A_n) is a sequence of nonempty compact sets converging to a nonempty compact set A in the Hausdorff metric, then

$$A = \{x \in X \mid \text{there is a sequence, } (x_n), \text{ with } x_n \in A_n \text{ converging to } x\}.$$

Indeed, if (x_n) is a sequence with $x_n \in A_n$ converging to x and (A_n) converges to A then, for every $\epsilon > 0$, there is some x_n such that $d(x_n, x) \leq \epsilon/2$ and there is some $a_n \in A$ such that $d(a_n, x_n) \leq \epsilon/2$ and thus, $d(a_n, x) \leq \epsilon$, which shows that $x \in \overline{A}$. Since A is compact, it is closed, and $x \in A$. See Figure 37.48.

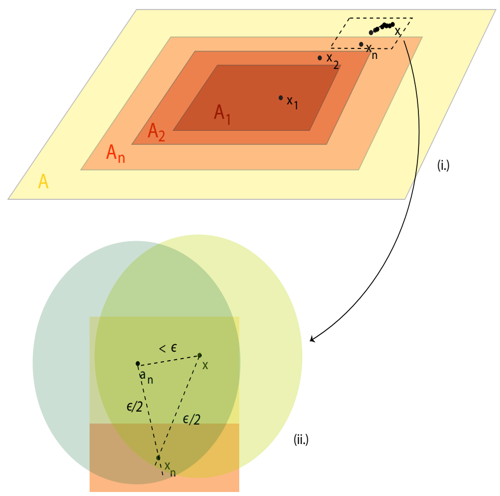


Figure 37.48: Let (A_n) be the sequence of parallelograms converging to A, the large pale yellow parallelogram. Figure $(ii.)$ expands the dashed region and shows why $d(a_n, x) < \epsilon$.

Conversely, since (A_n) converges to A, for every $x \in A$, for every $n \geq 1$, there is some $x_n \in A_n$ such that $d(x_n, x) \leq 1/n$ and the sequence (x_n) converges to x.

Now let (A_n) be a Cauchy sequence in $\mathcal{K}(X)$. It can be proven that (A_n) converges to the set

$$A = \{x \in X \mid \text{there is a sequence, } (x_n), \text{ with } x_n \in A_n \text{ converging to } x\},$$

and that A is nonempty and compact. To prove that A is compact, one proves that it is totally bounded and complete. Details are given in Edgar [55].

Finally we need to prove that if (X, d) is compact, then $(\mathcal{K}(X), D)$ is compact. Since we already know that $(\mathcal{K}(X), D)$ is complete if (X, d) is, it is enough to prove that $(\mathcal{K}(X), D)$ is totally bounded if (X, d) is, which is not hard. $\square$

In view of Theorem 37.55 and Theorem 37.54, it is possible to define some nonempty compact subsets of X in terms of fixed points of contraction maps. This can be done in terms of iterated function systems, yielding a large class of fractals. However, we will omit this topic and instead refer the reader to Edgar [55].

In Chapter 38 we show how certain fractals can be defined by iterated function systems, using Theorem 37.55 and Theorem 37.54.

Before considering differentials, we need to look at the continuity of linear maps.

37.11 Continuous Linear and Multilinear Maps

If E and F are normed vector spaces, we first characterize when a linear map $f\colon E \to F$ is continuous.

Proposition 37.56. *Given two normed vector spaces E and F, for any linear map $f\colon E \to F$, the following conditions are equivalent:*

(1) The function f is continuous at 0.

(2) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k, \;\; \textit{for every } u \in E \textit{ such that } \|u\| \leq 1.$$

(3) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k\|u\|, \;\; \textit{for every } u \in E.$$

(4) The function f is continuous at every point of E.

Proof. Assume (1). Then for every $\epsilon > 0$, there is some $\eta > 0$ such that, for every $u \in E$, if $\|u\| \leq \eta$, then $\|f(u)\| \leq \epsilon$. Pick $\epsilon = 1$, so that there is some $\eta > 0$ such that, if $\|u\| \leq \eta$, then $\|f(u)\| \leq 1$. If $\|u\| \leq 1$, then $\|\eta u\| \leq \eta\|u\| \leq \eta$, and so, $\|f(\eta u)\| \leq 1$, that is, $\eta\|f(u)\| \leq 1$, which implies $\|f(u)\| \leq \eta^{-1}$. Thus, (2) holds with $k = \eta^{-1}$.

Assume that (2) holds. If $u = 0$, then by linearity, $f(0) = 0$, and thus $\|f(0)\| \leq k\|0\|$ holds trivially for all $k \geq 0$. If $u \neq 0$, then $\|u\| > 0$, and since

$$\left\|\frac{u}{\|u\|}\right\| = 1,$$

we have

$$\left\| f\left(\frac{u}{\|u\|}\right)\right\| \leq k,$$

which implies that

$$\|f(u)\| \leq k\|u\|.$$

Thus, (3) holds.

If (3) holds, then for all $u, v \in E$, we have

$$\|f(v) - f(u)\| = \|f(v-u)\| \leq k\|v-u\|.$$

If $k = 0$, then f is the zero function, and continuity is obvious. Otherwise, if $k > 0$, for every $\epsilon > 0$, if $\|v - u\| \leq \frac{\epsilon}{k}$, then $\|f(v-u)\| \leq \epsilon$, which shows continuity at every $u \in E$. Finally, it is obvious that (4) implies (1). $\square$

Among other things, Proposition 37.56 shows that a linear map is continuous iff the image of the unit (closed) ball is bounded. Since a continuous linear map satisfies the condition $\|f(u)\| \leq k\|u\|$ (for some $k \geq 0$), it is also uniformly continuous.

If E and F are normed vector spaces, the set of all continuous linear maps $f\colon E \to F$ is denoted by $\mathcal{L}(E; F)$.

Using Proposition 37.56, we can define a norm on $\mathcal{L}(E; F)$ which makes it into a normed vector space. This definition has already been given in Chapter 9 (Definition 9.7) but for the reader's convenience, we repeat it here.

Definition 37.41. Given two normed vector spaces E and F, for every continuous linear map $f\colon E \to F$, we define the *operator norm* $\|f\|$ *of* f as

$$\|f\| = \inf\left\{k \geq 0 \mid \|f(x)\| \leq k\|x\|, \text{ for all } x \in E\right\} = \sup\left\{\|f(x)\| \mid \|x\| \leq 1\right\}.$$

From Definition 37.41, for every continuous linear map $f \in \mathcal{L}(E; F)$, we have

$$\|f(x)\| \leq \|f\|\|x\|,$$

for every $x \in E$. It is easy to verify that $\mathcal{L}(E; F)$ is a normed vector space under the norm of Definition 37.41. Furthermore, if E, F, G, are normed vector spaces, and $f\colon E \to F$ and $g\colon F \to G$ are continuous linear maps, we have

$$\|g \circ f\| \leq \|g\|\|f\|.$$

We can now show that when $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \,\|_1$, $\| \,\|_2$, or $\| \,\|_\infty$, then every linear map $f\colon E \to F$ is continuous.

Proposition 37.57. *If $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \,\|_1$, $\| \,\|_2$, or $\| \,\|_\infty$, and F is any normed vector space, then every linear map $f\colon E \to F$ is continuous.*

Proof. Let $(e_1, \ldots, e_n)$ be the standard basis of $\mathbb{R}^n$ (a similar proof applies to $\mathbb{C}^n$). In view of Proposition 9.3, it is enough to prove the proposition for the norm

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

We have,

$$\|f(v) - f(u)\| = \|f(v-u)\| = \left\| f\Big(\sum_{1\leq i\leq n} (v_i - u_i)e_i\Big) \right\| = \left\| \sum_{1\leq i\leq n} (v_i - u_i) f(e_i) \right\|,$$

and so,

$$\|f(v) - f(u)\| \leq \Big(\sum_{1\leq i\leq n} \|f(e_i)\|\Big) \max_{1\leq i\leq n} |v_i - u_i| = \Big(\sum_{1\leq i\leq n} \|f(e_i)\|\Big) \|v - u\|_\infty.$$

By the argument used in Proposition 37.56 to prove that (3) implies (4), f is continuous. ☐

Actually, we proved in Theorem 9.5 that if E is a vector space of finite dimension, then any two norms are equivalent, so that they define the same topology. This fact together with Proposition 37.57 prove the following:

Theorem 37.58. *If E is a vector space of finite dimension (over $\mathbb{R}$ or $\mathbb{C}$), then all norms are equivalent (define the same topology). Furthermore, for any normed vector space F, every linear map $f\colon E \to F$ is continuous.*

If E is a normed vector space of infinite dimension, a linear map $f\colon E \to F$ may not be continuous. As an example, let E be the infinite vector space of all polynomials over $\mathbb{R}$. Let

$$\|P(X)\| = \max_{0\leq x\leq 1} |P(x)|.$$

We leave as an exercise to show that this is indeed a norm. Let $F = \mathbb{R}$, and let $f\colon E \to F$ be the map defined such that, $f(P(X)) = P(3)$. It is clear that f is linear. Consider the sequence of polynomials

$$P_n(X) = \left(\frac{X}{2}\right)^n.$$

It is clear that $\|P_n\| = \left(\frac{1}{2}\right)^n$, and thus, the sequence P_n has the null polynomial as a limit. However, we have

$$f(P_n(X)) = P_n(3) = \left(\frac{3}{2}\right)^n,$$

and the sequence $f(P_n(X))$ diverges to $+\infty$. Consequently, in view of Proposition 37.15 (1), f is not continuous.

We now consider the continuity of multilinear maps. We treat explicitly bilinear maps, the general case being a straightforward extension.

Proposition 37.59. *Given normed vector spaces E, F and G, for any bilinear map $f\colon E \times F \to G$, the following conditions are equivalent:*

(1) The function f is continuous at $\langle 0, 0\rangle$.

2) There is a constant $k \geq 0$ such that,

$$\|f(u,v)\| \leq k, \ \textit{for all } u \in E, v \in F \textit{ such that } \|u\|, \|v\| \leq 1.$$

3) There is a constant $k \geq 0$ such that,

$$\|f(u,v)\| \leq k\|u\|\|v\|, \ \textit{for all } u \in E, v \in F.$$

4) The function f is continuous at every point of $E \times F$.

Proof. It is similar to that of Proposition 37.56, with a small subtlety in proving that (3) implies (4), namely that two different η's that are not independent are needed. $\square$

In contrast to continuous linear maps, which must be uniformly continuous, nonzero continuous bilinear maps are **not** uniformly continuous. Let $f\colon E \times F \to G$ be a continuous bilinear map such that $f(a,b) \neq 0$ for some $a \in E$ and some $b \in F$. Consider the sequences (u_n) and (v_n) (with $n \geq 1$) given by

$$\begin{aligned} u_n &= (x_n, y_n) = (na, nb) \\ v_n &= (x'_n, y'_n) = \left(\left(n + \frac{1}{n}\right) a, \left(n + \frac{1}{n}\right) b\right). \end{aligned}$$

Obviously

$$\|v_n - u_n\| \leq \frac{1}{n}(\|a\| + \|b\|),$$

so $\lim_{n\mapsto\infty} \|v_n - u_n\| = 0$. On the other hand

$$f(x'_n, y'_n) - f(x_n, y_n) = \left(2 + \frac{1}{n^2}\right) f(a,b),$$

and thus $\lim_{n\mapsto\infty} \|f(x'_n, y'_n) - f(x_n, y_n)\| = 2\,\|f(a,b)\| \neq 0$, which shows that f is not uniformly continuous, because if this was the case, this limit would be zero.

If E, F, and G, are normed vector spaces, we denote the set of all continuous bilinear maps $f\colon E \times F \to G$ by $\mathcal{L}_2(E, F; G)$. Using Proposition 37.59, we can define a norm on $\mathcal{L}_2(E, F; G)$ which makes it into a normed vector space.

Definition 37.42. Given normed vector spaces E, F, and G, for every continuous bilinear map $f\colon E \times F \to G$, we define the *norm* $\|f\|$ *of* f as

$$\begin{aligned} \|f\| &= \inf \{k \geq 0 \mid \|f(x,y)\| \leq k\|x\|\|y\|, \text{ for all } x \in E, y \in F\} \\ &= \sup \{\|f(x,y)\| \mid \|x\|, \|y\| \leq 1\}. \end{aligned}$$

From Definition 37.41, for every continuous bilinear map $f \in \mathcal{L}_2(E,F;G)$, we have

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

for all $x, y \in E$. It is easy to verify that $\mathcal{L}_2(E,F;G)$ is a normed vector space under the norm of Definition 37.42.

Given a bilinear map $f\colon E \times F \to G$, for every $u \in E$, we obtain a linear map denoted $fu\colon F \to G$, defined such that, $fu(v) = f(u,v)$. Furthermore, since

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

it is clear that fu is continuous. We can then consider the map $\varphi\colon E \to \mathcal{L}(F;G)$, defined such that, $\varphi(u) = fu$, for any $u \in E$, or equivalently, such that,

$$\varphi(u)(v) = f(u,v).$$

Actually, it is easy to show that φ is linear and continuous, and that $\|\varphi\| = \|f\|$. Thus, $f \mapsto \varphi$ defines a map from $\mathcal{L}_2(E,F;G)$ to $\mathcal{L}(E;\mathcal{L}(F;G))$. We can also go back from $\mathcal{L}(E;\mathcal{L}(F;G))$ to $\mathcal{L}_2(E,F;G)$. We summarize all this in the following proposition.

Proposition 37.60. *Let E, F, G be three normed vector spaces. The map $f \mapsto \varphi$, from $\mathcal{L}_2(E,F;G)$ to $\mathcal{L}(E;\mathcal{L}(F;G))$, defined such that, for every $f \in \mathcal{L}_2(E,F;G)$,*

$$\varphi(u)(v) = f(u,v),$$

is an isomorphism of vector spaces, and furthermore, $\|\varphi\| = \|f\|$.

As a corollary of Proposition 37.60, we get the following proposition which will be useful when we define second-order derivatives.

Proposition 37.61. *Let E, F be normed vector spaces. The map app from $\mathcal{L}(E;F) \times E$ to F, defined such that, for every $f \in \mathcal{L}(E;F)$, for every $u \in E$,*

$$app(f,u) = f(u),$$

is a continuous bilinear map.

Remark: If E and F are nontrivial, it can be shown that $\|\mathrm{app}\| = 1$. It can also be shown that composition

$$\circ\colon \mathcal{L}(E;F) \times \mathcal{L}(F;G) \to \mathcal{L}(E;G),$$

is bilinear and continuous.

The above propositions and definition generalize to arbitrary n-multilinear maps, with $n \geq 2$. Proposition 37.59 extends in the obvious way to any n-multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, but condition (3) becomes:

There is a constant $k \geq 0$ such that,

$$\|f(u_1, \ldots, u_n)\| \leq k\|u_1\| \cdots \|u_n\|, \text{ for all } u_1 \in E_1, \ldots, u_n \in E_n.$$

Definition 37.42 also extends easily to

$$\begin{aligned}\|f\| &= \inf\left\{k \geq 0 \mid \|f(x_1, \ldots, x_n)\| \leq k\|x_1\| \cdots \|x_n\|, \text{ for all } x_i \in E_i, 1 \leq i \leq n\right\}\\ &= \sup\left\{\|f(x_1, \ldots, x_n)\| \mid \|x_n\|, \ldots, \|x_n\| \leq 1\right\}.\end{aligned}$$

Proposition 37.60 is also easily extended, and we get an isomorphism between continuous n-multilinear maps in $\mathcal{L}_n(E_1, \ldots, E_n; F)$, and continuous linear maps in

$$\mathcal{L}(E_1; \mathcal{L}(E_2; \ldots; \mathcal{L}(E_n; F)))$$

An obvious extension of Proposition 37.61 also holds.

Definition 37.43. A normed vector space $(E, \| \, \|)$ over $\mathbb{R}$ (or $\mathbb{C}$) which is a complete metric space for the distance $d(u, v) = \|v - u\|$, is called a *Banach space*.

The following theorem is a key result of the theory of Banach spaces worth proving.

Theorem 37.62. *If E and F are normed vector spaces, and if F is a Banach space, then $\mathcal{L}(E;F)$ is a Banach space (with the operator norm).*

Proof. Let $(f)_{n\geq 1}$ be a Cauchy sequence of continuous linear maps $f_n\colon E \to F$. We proceed in several steps.

Step 1. Define the pointwise limit $f\colon E \to F$ of the sequence $(f_n)_{n\geq 1}$.

Since $(f)_{n\geq 1}$ is a Cauchy sequence, for every $\epsilon > 0$, there is some $N > 0$ such that $\|f_m - f_n\| < \epsilon$ for all $m, n \geq N$. Since $\| \, \|$ is the operator norm, we deduce that for any $u \in E$, we have

$$\|f_m(u) - f_n(u)\| = \|(f_m - f_n)(u)\| \leq \|f_m - f_n\| \, \|u\| \leq \epsilon \, \|u\| \quad \text{for all } m, n \geq N,$$

that is,

$$\|f_m(u) - f_n(u)\| \leq \epsilon \, \|u\| \quad \text{for all } m, n \geq N. \tag{$*_1$}$$

If $u = 0$, then $f_m(0) = f_n(0) = 0$ for all m, n, so the sequence $(f_n(0))$ is a Cauchy sequence in F converging to 0. If $u \neq 0$, by replacing ϵ by $\epsilon / \|u\|$, we see that the sequence $(f_n(u))$ is a Cauchy sequence in F. Since F is complete, the sequence $(f_n(u))$ has a limit which we denote by $f(u)$. This defines our candidate limit function f by

$$f(u) = \lim_{n \mapsto \infty} f_n(u).$$

It remains to prove that

1. f is linear.
2. f is continous.
3. f is the limit of (f_n) for the operator norm.

Step 2. The function f is linear.

Recall that in a normed vector space, addition and multiplication by a fixed scalar are continuous (since $\|u + v\| \leq \|u\| + \|v\|$ and $\|\lambda u\| \leq |\lambda| \, \|u\|$). Thus by definition of f and since the f_n are linear we have

$$\begin{aligned}
f(u+v) &= \lim_{n \mapsto \infty} f_n(u+v) && \text{by definition of } f \\
&= \lim_{n \mapsto \infty} (f_n(u) + f_n(v)) && \text{by linearity of } f_n \\
&= \lim_{n \mapsto \infty} f_n(u) + \lim_{n \mapsto \infty} f_n(v) && \text{since } + \text{ is continuous} \\
&= f(u) + f(v) && \text{by definition of } f.
\end{aligned}$$

Similarly,

$$\begin{aligned}
f(\lambda u) &= \lim_{n \mapsto \infty} f_n(\lambda u) && \text{by definition of } f \\
&= \lim_{n \mapsto \infty} \lambda f_n(u) && \text{by linearity of } f_n \\
&= \lambda \lim_{n \mapsto \infty} f_n(u) && \text{by continuity of scalar multiplication} \\
&= \lambda f(u) && \text{by definition of } f.
\end{aligned}$$

Therefore, f is linear.

Step 3. The function f is continuous.

Since $(f_n)_{n \geq 1}$ is a Cauchy sequence, for every $\epsilon > 0$, there is some $N > 0$ such that $\|f_m - f_n\| < \epsilon$ for all $m, n \geq N$. Since $f_m = f_n + f_m - f_n$, we get $\|f_m\| \leq \|f_n\| + \|f_m - f_n\|$, which implies that

$$\|f_m\| \leq \|f_n\| + \epsilon \quad \text{for all } m, n \geq N. \tag{$*_2$}$$

Using $(*_2)$, we also have

$$\|f_m(u)\| \leq \|f_m\| \, \|u\| \leq (\|f_n\| + \epsilon) \, \|u\| \quad \text{for all } m, n \geq N,$$

that is,

$$\|f_m(u)\| \leq (\|f_n\| + \epsilon) \, \|u\| \quad \text{for all } m, n \geq N. \tag{$*_3$}$$

Hold $n \geq N$ fixed and let m tend to $+\infty$ in $(*_3)$. Since the norm is continuous, we get

$$\|f(u)\| \leq (\|f_n\| + \epsilon) \, \|u\| \, ,$$

which shows that f is continuous.

Step 4. The function f is the limit of (f_n) for the operator norm.

Recall $(*_1)$:

$$\|f_m(u) - f_n(u)\| \leq \epsilon \, \|u\| \quad \text{for all } m, n \geq N. \tag{$*_1$}$$

Hold $n \geq N$ fixed but this time let m tend to $+\infty$ in $(*_1)$. By continuity of the norm we get

$$\|f(u) - f_n(u)\| = \|(f - f_n)(u)\| \leq \epsilon \, \|u\| \, .$$

By definition of the operator norm,

$$\|f - f_n\| = \sup\{\|(f - f_n)(u)\| \mid \|u\| = 1\} \leq \epsilon \quad \text{for all } n \geq N,$$

which proves that f_n converges to f for the operator norm. $\square$

As a special case of Theorem 37.62, if we let $F = \mathbb{R}$ (or $F = \mathbb{C}$ in the case of complex vector spaces) we see that $E' = \mathcal{L}(E; \mathbb{R})$ (or $E' = \mathcal{L}(E; \mathbb{C})$) is complete (since $\mathbb{R}$ and $\mathbb{C}$ are complete). The space E' of continuous linear forms on E is called the *dual* of E. It is a subspace of the *algebraic dual* E^* of E which consists of *all* linear forms on E, not necessarily continuous.

It can also be shown that if E, F and G are normed vector spaces, and if G is a Banach space, then $\mathcal{L}_2(E, F; G)$ is a Banach space. The proof is essentially identical.

37.12 Completion of a Normed Vector Space

An easy corollary of Theorem 37.53 and Theorem 37.52 is that every normed vector space can be embedded in a complete normed vector space, that is, a Banach space.

Theorem 37.63. *If $(E, \| \,\|)$ is a normed vector space, then its completion $(\widehat{E}, \widehat{d})$ as a metric space (where E is given the metric $d(x, y) = \|x - y\|$) can be given a unique vector space structure extending the vector space structure on E, and a norm $\| \,\|_{\widehat{E}}$, so that $(\widehat{E}, \| \,\|_{\widehat{E}})$ is a Banach space, and the metric $\widehat{d}$ is associated with the norm $\| \,\|_{\widehat{E}}$. Furthermore, the isometry $\varphi \colon E \to \widehat{E}$ is a linear isometry.*

Proof. The addition operation $+ \colon E \times E \to E$ is uniformly continuous because

$$\|(u' + v') - (u'' + v'')\| \leq \|u' - u''\| + \|v' - v''\|.$$

It is not hard to show that $\widehat{E} \times \widehat{E}$ is a complete metric space and that $E \times E$ is dense in $\widehat{E} \times \widehat{E}$. Then, by Theorem 37.52, the uniformly continuous function $+$ has a unique continuous extension $+ \colon \widehat{E} \times \widehat{E} \to \widehat{E}$.

The map $\cdot \colon \mathbb{R} \times E \to E$ is not uniformly continuous, but for any fixed $\lambda \in \mathbb{R}$, the map $L_\lambda \colon E \to E$ given by $L_\lambda(u) = \lambda \cdot u$ is uniformly continuous, so by Theorem 37.52 the function L_λ has a unique continuous extension $L_\lambda \colon \widehat{E} \to \widehat{E}$, which we use to define the scalar multiplication $\cdot \colon \mathbb{R} \times \widehat{E} \to \widehat{E}$. It is easily checked that with the above addition and scalar multiplication, $\widehat{E}$ is a vector space.

Since the norm $\| \;\|$ on E is uniformly continuous, it has a unique continuous extension $\| \;\|_{\widehat{E}} \colon \widehat{E} \to \mathbb{R}_+$. The identities $\|u + v\| \leq \|u\| + \|v\|$ and $\|\lambda u\| \leq |\lambda| \, \|u\|$ extend to $\widehat{E}$ by continuity. The equation

$$d(u, v) = \|u - v\|$$

also extends to $\widehat{E}$ by continuity and yields

$$\widehat{d}(\alpha, \beta) = \|\alpha - \beta\|_{\widehat{E}},$$

which shows that $\| \;\|_{\widehat{E}}$ is indeed a norm, and that the metric $\widehat{d}$ is associated to it. Finally, it is easy to verify that the map φ is linear. The uniqueness of the structure of normed vector space follows from the uniqueness of continuous extensions in Theorem 37.52. ∎

Theorem 37.63 and Theorem 37.52 will be used to show that every Hermitian space can be embedded in a Hilbert space.

The following version of Theorem 37.52 for normed vector spaces is needed in the theory of integration.

Theorem 37.64. *Let E and F be two normed vector spaces, let E_0 be a dense subspace of E, and let $f_0 \colon E_0 \to F$ be a continuous function. If f_0 is uniformly continuous and if F is complete, then there is a unique uniformly continuous function $f \colon E \to F$ extending f_0. Furthermore, if f_0 is a continuous linear map, then f is also a linear continuous map, and $\|f\| = \|f_0\|$.*

Proof. We only need to prove the second statement. Given any two vectors $x, y \in E$, since E_0 is dense on E we can pick sequences (x_n) and (y_n) of vectors $x_n, y_n \in E_0$ such that $x = \lim_{n \mapsto \infty} x_n$ and $y = \lim_{n \mapsto \infty} y_n$. Since addition and scalar multiplication are continuous, we get

$$x + y = \lim_{n \mapsto \infty} (x_n + y_n)$$
$$\lambda x = \lim_{n \mapsto \infty} (\lambda x_n)$$

for any $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$). Since $f(x)$ is defined by

$$f(x) = \lim_{n \mapsto \infty} f_0(x_n)$$

independently of the sequence (x_n) converging to x, and similarly for $f(y)$ and $f(x+y)$, since f_0 is linear, we have

$$\begin{aligned} f(x+y) &= \lim_{n \mapsto \infty} f_0(x_n + y_n) \\ &= \lim_{n \mapsto \infty} (f_0(x_n) + f_0(y_n)) \\ &= \lim_{n \mapsto \infty} f_0(x_n) + \lim_{n \mapsto \infty} f_0(y_n) \\ &= f(x) + f(y). \end{aligned}$$

Similarly,

$$\begin{aligned} f(\lambda x) &= \lim_{n \mapsto \infty} f_0(\lambda x_n) \\ &= \lim_{n \mapsto \infty} \lambda f_0(x_n) \\ &= \lambda \lim_{n \mapsto \infty} f_0(x_n) \\ &= \lambda f(x). \end{aligned}$$

Therefore, f is linear. Since the norm is continuous, we have

$$\|f(x)\| = \left\| \lim_{n \mapsto \infty} f_0(x_n) \right\| = \lim_{n \mapsto \infty} \|f_0(x_n)\|,$$

and since f_0 is continuous

$$\|f_0(x_n)\| \leq \|f_0\| \|x_n\| \quad \text{for all } n \geq 1,$$

so we get

$$\lim_{n \mapsto \infty} \|f_0(x_n)\| \leq \lim_{n \mapsto \infty} \|f_0\| \|x_n\| \quad \text{for all } n \geq 1,$$

that is,

$$\|f(x)\| \leq \|f_0\| \|x\|.$$

Since

$$\|f\| = \sup_{\|x\|=1,\, x \in E} \|f(x)\|,$$

we deduce that $\|f\| \leq \|f_0\|$. But since $E_0 \subseteq E$ and f agrees with f_0 on E_0, we also have

$$\|f_0\| = \sup_{\|x\|=1,\, x \in E_0} \|f_0(x)\| = \sup_{\|x\|=1,\, x \in E_0} \|f(x)\| \leq \sup_{\|x\|=1,\, x \in E} \|f(x)\| = \|f\|,$$

and thus $\|f\| = \|f_0\|$. $\square$

Finally, we consider normed affine spaces.

37.13 Normed Affine Spaces

For geometric applications, we will need to consider affine spaces $(E, \overrightarrow{E})$ where the associated space of translations $\overrightarrow{E}$ is a vector space equipped with a norm.

Definition 37.44. Given an affine space $(E, \overrightarrow{E})$, where the space of translations $\overrightarrow{E}$ is a vector space over $\mathbb{R}$ or $\mathbb{C}$, we say that $(E, \overrightarrow{E})$ is a *normed affine space* if $\overrightarrow{E}$ is a normed vector space with norm $\| \; \|$.

Given a normed affine space, there is a natural metric on E itself, defined such that

$$d(a, b) = \|\overrightarrow{ab}\|.$$

Observe that this metric is invariant under translation, that is,

$$d(a + u, b + u) = d(a, b).$$

Also, for every fixed $a \in E$ and $\lambda > 0$, if we consider the map $h\colon E \to E$, defined such that,

$$h(x) = a + \lambda\overrightarrow{ax},$$

then $d(h(x), h(y)) = \lambda d(x, y)$.

Note that the map $(a, b) \mapsto \overrightarrow{ab}$ from $E \times E$ to $\overrightarrow{E}$ is continuous, and similarly for the map $a \mapsto a + u$ from $E \times \overrightarrow{E}$ to E. In fact, the map $u \mapsto a + u$ is a homeomorphism from $\overrightarrow{E}$ to E_a.

Of course, $\mathbb{R}^n$ is a normed affine space under the Euclidean metric, and it is also complete.

If an affine space E is a finite direct sum $(E_1, a_1) \oplus \cdots \oplus (E_m, a_m)$, and each E_i is also a normed affine space with norm $\| \; \|_i$, we make $(E_1, a_1) \oplus \cdots \oplus (E_m, a_m)$ into a normed affine space, by giving it the norm

$$\|(x_1, \ldots, x_n)\| = \max(\|x_1\|_1, \ldots, \|x_n\|_n).$$

Similarly, the finite product $E_1 \times \cdots \times E_m$ is made into a normed affine space, under the same norm.

We are now ready to define the derivative (or differential) of a map between two normed affine spaces. This will lead to tangent spaces to curves and surfaces (in normed affine spaces).

37.14 Futher Readings

A thorough treatment of general topology can be found in Munkres [131, 130], Dixmier [51], Lang [111, 112], Schwartz [150, 149], Bredon [30], and the classic, Seifert and Threlfall [155].

Chapter 38

A Detour On Fractals

38.1 Iterated Function Systems and Fractals

A pleasant application of the Hausdorff distance and of the fixed point theorem for contracting mappings is a method for defining a class of "self-similar" fractals. For this, we can use iterated function systems.

Definition 38.1. Given a metric space, (X, d), an *iterated function system*, for short, an *ifs*, is a finite sequence of functions, $(f_1, \ldots, f_n)$, where each $f_i \colon X \to X$ is a contracting mapping. A nonempty compact subset, K, of X is an *invariant set (or attractor)* for the ifs, $(f_1, \ldots, f_n)$, if

$$K = f_1(K) \cup \cdots \cup f_n(K).$$

The major result about ifs's is the following:

Theorem 38.1. *If (X, d) is a nonempty complete metric space, then every iterated function system, $(f_1, \ldots, f_n)$, has a unique invariant set, A, which is a nonempty compact subset of X. Furthermore, for every nonempty compact subset, A_0, of X, this invariant set, A, if the limit of the sequence, (A_m), where $A_{m+1} = f_1(A_m) \cup \cdots \cup f_n(A_m)$.*

Proof. Since X is complete, by Theorem 37.55, the space $(\mathcal{K}(X), D)$ is a complete metric space. The theorem will follow from Theorem 37.54 if we can show that the map, $F \colon \mathcal{K}(X) \to \mathcal{K}(X)$, defined such that

$$F(K) = f_1(K) \cup \cdots \cup f_n(K),$$

for every nonempty compact set, K, is a contracting mapping. Let A, B be any two nonempty compact subsets of X and consider any $\eta \geq D(A, B)$. Since each $f_i \colon X \to X$ is a contracting mapping, there is some λ_i, with $0 \leq \lambda_i < 1$, such that

$$d(f_i(a), f_i(b)) \leq \lambda_i d(a, b),$$

for all $a, b \in X$. Let $\lambda = \max\{\lambda_1, \ldots, \lambda_n\}$. We claim that

$$D(F(A), F(B)) \leq \lambda D(A, B).$$

For any $x \in F(A) = f_1(A) \cup \cdots \cup f_n(A)$, there is some $a_i \in A_i$ such that $x = f_i(a_i)$ and since $\eta \geq D(A, B)$, there is some $b_i \in B$ such that

$$d(a_i, b_i) \leq \eta,$$

and thus,

$$d(x, f_i(b_i)) = d(f_i(a_i), f_i(b_i)) \leq \lambda_i d(a_i, b_i) \leq \lambda\eta.$$

This show that

$$F(A) \subseteq V_{\lambda\eta}(F(B)).$$

Similarly, we can prove that

$$F(B) \subseteq V_{\lambda\eta}(F(A)),$$

and since this holds for all $\eta \geq D(A, B)$, we proved that

$$D(F(A), F(B)) \leq \lambda D(A, B)$$

where $\lambda = \max\{\lambda_1, \ldots, \lambda_n\}$. Since $0 \leq \lambda_i < 1$, we have $0 \leq \lambda < 1$ and F is indeed a contracting mapping. $\square$

Theorem 38.1 justifies the existence of many familiar "self-similar" fractals. One of the best known fractals is the *Sierpinski gasket*.

Example 38.1. Consider an equilateral triangle with vertices a, b, c, and let f_1, f_2, f_3 be the dilatations of centers a, b, c and ratio $1/2$. The Sierpinski gasket is the invariant set of the ifs (f_1, f_2, f_3). The dilations f_1, f_2, f_3 can be defined explicitly as follows, assuming that $a = (-1/2, 0)$, $b = (1/2, 0)$, and $c = (0, \sqrt{3}/2)$. The contractions f_1, f_2, f_3 are specified by

$$\begin{aligned} x' &= \frac{1}{2}x - \frac{1}{4}, \\ y' &= \frac{1}{2}y, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{2}x + \frac{1}{4}, \\ y' &= \frac{1}{2}y, \end{aligned}$$

and

$$\begin{aligned} x' &= \frac{1}{2}x, \\ y' &= \frac{1}{2}y + \frac{\sqrt{3}}{4}. \end{aligned}$$

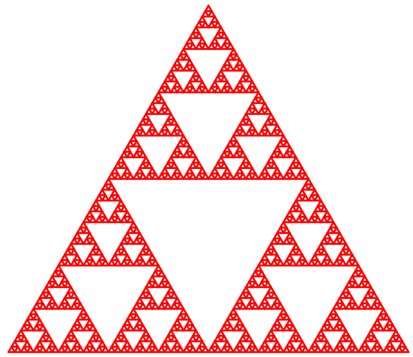


Figure 38.1: The Sierpinski gasket

We wrote a *Mathematica* program that iterates any finite number of affine maps on any input figure consisting of combinations of points, line segments, and polygons (with their interior points). Starting with the edges of the triangle a, b, c, after 6 iterations, we get the picture shown in Figure 38.1.

It is amusing that the same fractal is obtained no matter what the initial nonempty compact figure is. It is interesting to see what happens if we start with a solid triangle (with its interior points). The result after 6 iterations is shown in Figure 38.2. The convergence towards the Sierpinski gasket is very fast. Incidently, there are many other ways of defining the Sierpinski gasket.

A nice variation on the theme of the Sierpinski gasket is the *Sierpinski dragon*.

Example 38.2. The Sierpinski dragon is specified by the following three contractions:

$$\begin{aligned} x' &= -\frac{1}{4}x - \frac{\sqrt{3}}{4}y + \frac{3}{4}, \\ y' &= \frac{\sqrt{3}}{4}x - \frac{1}{4}y + \frac{\sqrt{3}}{4}, \end{aligned}$$

$$\begin{aligned} x' &= -\frac{1}{4}x + \frac{\sqrt{3}}{4}y - \frac{3}{4}, \\ y' &= -\frac{\sqrt{3}}{4}x - \frac{1}{4}y + \frac{\sqrt{3}}{4}, \end{aligned}$$

$$\begin{aligned} x' &= \frac{1}{2}x, \\ y' &= \frac{1}{2}y + \frac{\sqrt{3}}{2}. \end{aligned}$$

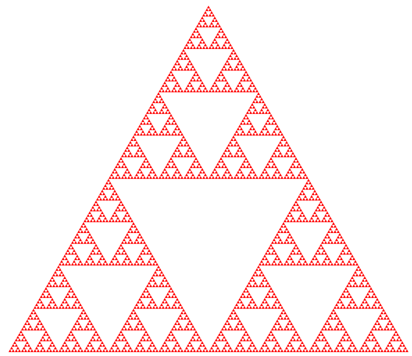


Figure 38.2: The Sierpinski gasket, version 2

The result of 7 iterations starting from the line segment $(-1, 0), (1, 0))$, is shown in Figure 38.3. This curve converges to the boundary of the Sierpinski gasket.

A different kind of fractal is the *Heighway dragon*.

Example 38.3. The Heighway dragon is specified by the following two contractions:

$$\begin{aligned}
x' &= \frac{1}{2}x - \frac{1}{2}y, \\
y' &= \frac{1}{2}x + \frac{1}{2}y,
\end{aligned}$$

$$\begin{aligned}
x' &= -\frac{1}{2}x - \frac{1}{2}y, \\
y' &= \frac{1}{2}x - \frac{1}{2}y + 1.
\end{aligned}$$

It can be shown that for any number of iterations, the polygon does not cross itself. This means that no edge is traversed twice and that if a point is traversed twice, then this point is the endpoint of some edge. The result of 13 iterations, starting with the line segment $((0, 0), (0, 1))$, is shown in Figure 38.4.

The Heighway dragon turns out to fill a closed and bounded set. It can also be shown that the plane can be tiled with copies of the Heighway dragon.

Another well known example is the *Koch curve*.

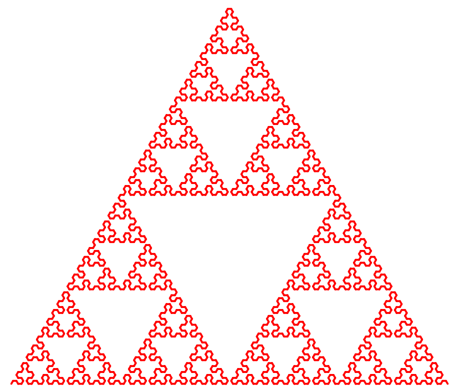


Figure 38.3: The Sierpinski dragon

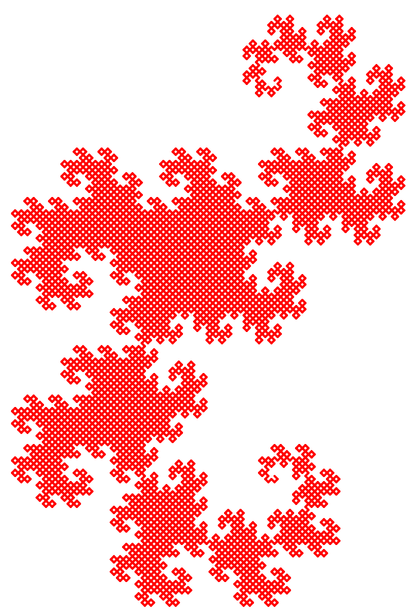


Figure 38.4: The Heighway dragon

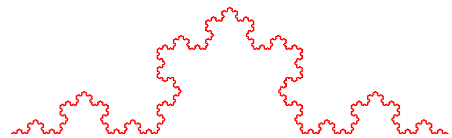


Figure 38.5: The Koch curve

Example 38.4. The Koch curve is specified by the following four contractions:

$$\begin{aligned}
x' &= \frac{1}{3}x - \frac{2}{3},\\
y' &= \frac{1}{3}y,\\[2mm]
x' &= \frac{1}{6}x - \frac{\sqrt{3}}{6}y - \frac{1}{6},\\
y' &= \frac{\sqrt{3}}{6}x + \frac{1}{6}y + \frac{\sqrt{3}}{6},\\[2mm]
x' &= \frac{1}{6}x + \frac{\sqrt{3}}{6}y + \frac{1}{6},\\
y' &= -\frac{\sqrt{3}}{6}x + \frac{1}{6}y + \frac{\sqrt{3}}{6},\\[2mm]
x' &= \frac{1}{3}x + \frac{2}{3},\\
y' &= \frac{1}{3}y.
\end{aligned}$$

The Koch curve is an example of a continuous curve which is nowhere differentiable (because it "wiggles" too much). It is a curve of infinite length. The result of 6 iterations, starting with the line segment $((-1,0),(1,0))$, is shown in Figure 38.5.

The curve obtained by putting three Kock curves together on the sides of an equilateral triangle is known as the *snowflake curve* (for obvious reasons, see below!).

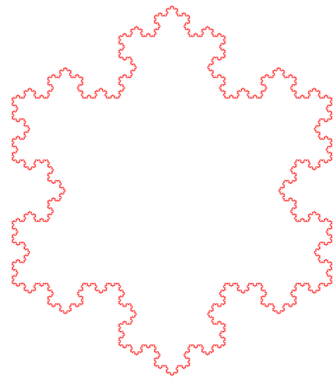


Figure 38.6: The snowflake curve

Example 38.5. The snowflake curve obtained after 5 iterations is shown in Figure 38.6.

The snowflake curve is an example of a closed curve of infinite length bounding a finite area.

We conclude with another famous example, a variant of the *Hilbert curve*.

Example 38.6. This version of the Hilbert curve is defined by the following four contractions:

$$
\begin{aligned}
x' &= \frac{1}{2}x - \frac{1}{2},\\
y' &= \frac{1}{2}y + 1,
\end{aligned}
$$

$$
\begin{aligned}
x' &= \frac{1}{2}x + \frac{1}{2},\\
y' &= \frac{1}{2}y + 1,
\end{aligned}
$$

$$
\begin{aligned}
x' &= -\frac{1}{2}y + 1,\\
y' &= \frac{1}{2}x + \frac{1}{2},
\end{aligned}
$$

$$
\begin{aligned}
x' &= \frac{1}{2}y - 1,\\
y' &= -\frac{1}{2}x + \frac{1}{2}.
\end{aligned}
$$

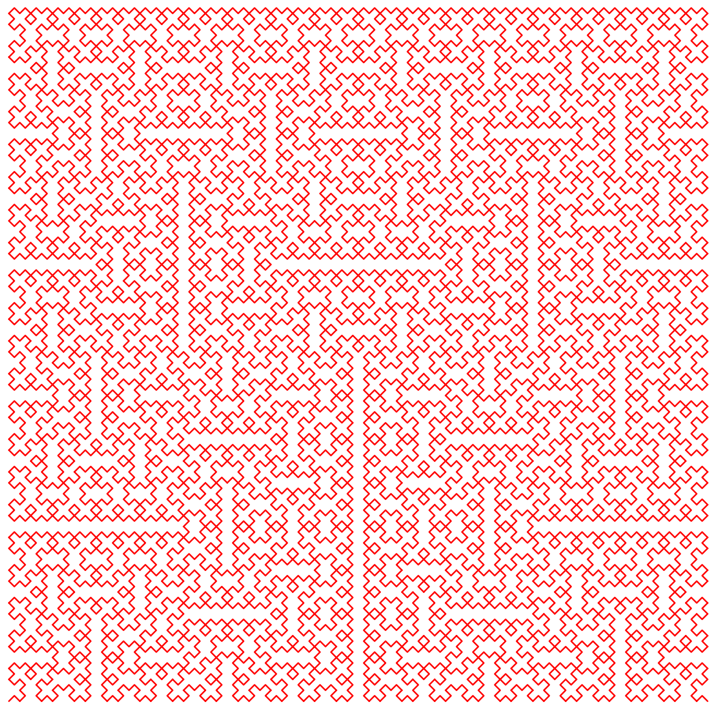


Figure 38.7: A Hilbert curve

This continuous curve is a space-filling curve, in the sense that its image is the entire unit square. The result of 6 iterations, starting with the two lines segments $((-1,0),(0,1))$ and $((0,1),(1,0))$, is shown in Figure 38.7.

For more on iterated function systems and fractals, we recommend Edgar [55].

Chapter 39

Differential Calculus

39.1 Directional Derivatives, Total Derivatives

This chapter contains a review of basic notions of differential calculus. First, we review the definition of the derivative of a function $f\colon \mathbb{R} \to \mathbb{R}$. Next, we define directional derivatives and the total derivative of a function $f\colon E \to F$ between normed affine spaces. Basic properties of derivatives are shown, including the chain rule. We show how derivatives are represented by Jacobian matrices. The mean value theorem is stated, as well as the implicit function theorem and the inverse function theorem. Diffeomorphisms and local diffeomorphisms are defined. Tangent spaces are defined. Higher-order derivatives are defined, as well as the Hessian. Schwarz's lemma (about the commutativity of partials) is stated. Several versions of Taylor's formula are stated, and a famous formula due to Faà di Bruno's is given.

We first review the notion of the derivative of a real-valued function whose domain is an open subset of $\mathbb{R}$.

Let $f\colon A \to \mathbb{R}$, where A is a nonempty open subset of $\mathbb{R}$, and consider any $a \in A$. The main idea behind the concept of the derivative of f at a, denoted by $f'(a)$, is that locally around a (that is, in some small open set $U \subseteq A$ containing a), the function f is approximated linearly[1] by the map

$$x \mapsto f(a) + f'(a)(x - a).$$

As pointed out by Dieudonné in the early 1960s, it is an "unfortunate accident" that if V is vector space of dimension one, then there is a bijection between the space V^* of linear forms defined on V and the field of scalars. As a consequence, the derivative of a real-valued function f defined on an open subset A of the reals can be defined as the scalar $f'(a)$ (for any $a \in A$). But as soon as f is a function of several arguments, the scalar interpretation of the derivative breaks down.

[1] Actually, the approximation is affine, but everybody commits this abuse of language.

Part of the difficulty in extending the idea of derivative to more complex spaces is to give an adequate notion of linear approximation. The key idea is to use linear maps. This could be carried out in terms of matrices but it turns out that this neither shortens nor simplifies proofs. In fact, this is often the opposite.

We admit that the more intrinsic definition of the notion of derivative f'_a at a point a of a function $f\colon E \to F$ between two normed (affine) spaces E and F as a linear map requires a greater effort to be grasped, but we feel that the advantages of this definition outweight its degree of abstraction. In particular, it yields a clear notion of the derivative of a function $f\colon \mathrm{M}_m(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ defined from $m \times m$ matrices to $n \times n$ matrices (many definitions make use of partial derivatives with respect to matrices that do make any sense). But more importantly, the definition of the derivative as a linear map makes it clear that whether the space E or the space F is infinite dimensional does not matter. This is important in optimization theory where the natural space of solutions of the problem is often an infinite dimensional function space. Of course, to carry out computations one need to pick finite bases and to use Jacobian matrices, but this is a different matter.

Let us now review the formal definition of the derivative of a real-valued function.

Definition 39.1. Let A be any nonempty open subset of $\mathbb{R}$, and let $a \in A$. For any function $f\colon A \to \mathbb{R}$, the *derivative of f at $a \in A$* is the limit (if it exists)

$$\lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\, h \neq 0\}$. This limit is denoted by $f'(a)$, or $\mathrm{D}f(a)$, or $\frac{df}{dx}(a)$. If $f'(a)$ exists for every $a \in A$, we say that f *is differentiable on A*. In this case, the map $a \mapsto f'(a)$ is denoted by f', or $\mathrm{D}f$, or $\frac{df}{dx}$.

Note that since A is assumed to be open, $A - \{a\}$ is also open, and since the function $h \mapsto a + h$ is continuous and U is the inverse image of $A - \{a\}$ under this function, U is indeed open and the definition makes sense.

We can also define $f'(a)$ as follows: there is some function ϵ, such that,

$$f(a+h) = f(a) + f'(a) \cdot h + \epsilon(h)h,$$

whenever $a + h \in A$, where $\epsilon(h)$ is defined for all h such that $a + h \in A$, and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Remark: We can also define the notion of *derivative of f at a on the left*, and *derivative of f at a on the right*. For example, we say that the *derivative of f at a on the left* is the limit $f'(a_-)$ (if it exists)

$$\lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\ h < 0\}$.

If a function f as in Definition 39.1 has a derivative $f'(a)$ at a, then it is continuous at a. If f is differentiable on A, then f is continuous on A. The composition of differentiable functions is differentiable.

Remark: A function f has a derivative $f'(a)$ at a iff the derivative of f on the left at a and the derivative of f on the right at a exist, and if they are equal. Also, if the derivative of f on the left at a exists, then f is continuous on the left at a (and similarly on the right).

We would like to extend the notion of derivative to functions $f \colon A \to F$, where E and F are normed affine spaces, and A is some nonempty open subset of E. The first difficulty is to make sense of the quotient

$$\frac{f(a+h) - f(a)}{h}.$$

If E and F are normed affine spaces, it will be notationally convenient to assume that the vector space associated with E is denoted by $\overrightarrow{E}$, and that the vector space associated with F is denoted as $\overrightarrow{F}$.

Since F is a normed affine space, making sense of $f(a+h)-f(a)$ is easy: we can define this as $\overrightarrow{f(a)f(a+h)}$, the unique vector translating $f(a)$ to $f(a+h)$. We should note however, that this quantity is a vector and not a point. Nevertheless, in defining derivatives, it is notationally more pleasant to denote $\overrightarrow{f(a)f(a+h)}$ by $f(a+h) - f(a)$. Thus, in the rest of this chapter, the vector $\overrightarrow{ab}$ will be denoted by $b - a$. But now, how do we define the quotient by a vector? Well, we don't!

A first possibility is to consider the *directional derivative* with respect to a vector $u \neq 0$ in $\overrightarrow{E}$. We can consider the vector $f(a + tu) - f(a)$, where $t \in \mathbb{R}$ (or $t \in \mathbb{C}$). Now,

$$\frac{f(a+tu) - f(a)}{t}$$

makes sense. The idea is that in E, the points of the form $a + tu$ for t in some small interval $[-\epsilon, +\epsilon]$ in $\mathbb{R}$ (or $\mathbb{C}$) form a line segment $[r, s]$ in A containing a, and that the image of this line segment defines a small curve segment on $f(A)$. This curve segment is defined by the map $t \mapsto f(a + tu)$, from $[r, s]$ to F, and the directional derivative $\mathrm{D}_u f(a)$ defines the direction of the tangent line at a to this curve; see Figure 39.1. This leads us to the following definition.

Definition 39.2. Let E and F be two normed affine spaces, let A be a nonempty open subset of E, and let $f \colon A \to F$ be any function. For any $a \in A$, for any $u \neq 0$ in $\overrightarrow{E}$, the *directional derivative of f at a w.r.t. the vector u*, denoted by $\mathrm{D}_u f(a)$, is the limit (if it exists)

$$\lim_{t \to 0,\, t \in U} \frac{f(a+tu) - f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tu \in A,\ t \neq 0\}$ (or $U = \{t \in \mathbb{C} \mid a + tu \in A,\ t \neq 0\}$).

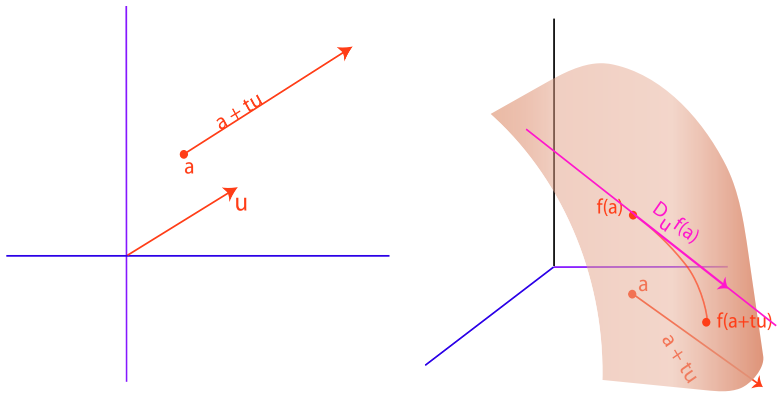


Figure 39.1: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the peach surface in $\mathbb{R}^3$, and $t \mapsto f(a+tu)$ is the embedded orange curve connecting $f(a)$ to $f(a+tu)$. Then $\mathrm{D}_u f(a)$ is the slope of the pink tangent line in the direction of u.

Since the map $t \mapsto a + tu$ is continuous, and since $A - \{a\}$ is open, the inverse image U of $A - \{a\}$ under the above map is open, and the definition of the limit in Definition 39.2 makes sense.

Remark: Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

The directional derivative is sometimes called the *Gâteaux derivative*.

In the special case where $E = \mathbb{R}$ and $F = \mathbb{R}$, and we let $u = 1$ (i.e., the real number 1, viewed as a vector), it is immediately verified that $\mathrm{D}_1 f(a) = f'(a)$, in the sense of Definition 39.1. When $E = \mathbb{R}$ (or $E = \mathbb{C}$) and F is any normed vector space, the derivative $\mathrm{D}_1 f(a)$, also denoted by $f'(a)$, provides a suitable generalization of the notion of derivative.

However, when E has dimension ≥ 2, directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonnull vectors u share something in common. As a consequence, a function can have all directional derivatives at a, and yet not be continuous at a. Two functions may have all directional derivatives in some open sets, and yet their composition may not.

Example 39.1. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be the function given by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^4 + y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

The graph of $f(x,y)$ is illustrated in Figure 39.2.

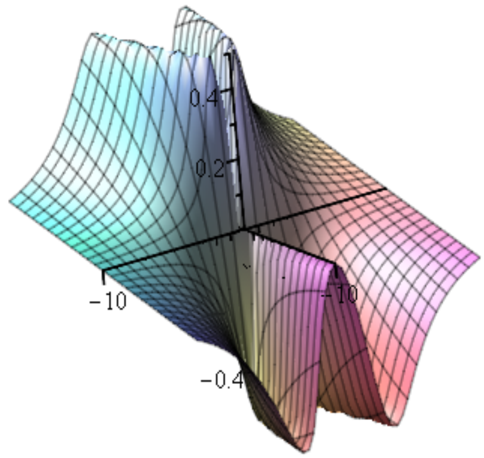


Figure 39.2: The graph of the function from Example 39.1. Note that f is not continuous at $(0,0)$, despite the existence of $\mathrm{D}_u f(0,0)$ for all $u \neq 0$.

For any $u \neq 0$, letting $u = \begin{pmatrix} h \\ k \end{pmatrix}$, we have

$$\frac{f(0+tu) - f(0)}{t} = \frac{h^2 k}{t^2 h^4 + k^2},$$

so that

$$\mathrm{D}_u f(0,0) = \begin{cases} \frac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0. \end{cases}$$

Thus, $\mathrm{D}_u f(0,0)$ exists for all $u \neq 0$.

On the other hand, if $\mathrm{D}f(0,0)$ existed, it would be a linear map $\mathrm{D}f(0,0)\colon \mathbb{R}^2 \to \mathbb{R}$ represented by a row matrix $(\alpha\ \beta)$, and we would have $\mathrm{D}_u f(0,0) = \mathrm{D}f(0,0)(u) = \alpha h + \beta k$, but the explicit formula for $\mathrm{D}_u f(0,0)$ is not linear. As a matter of fact, the function f is not continuous at $(0,0)$. For example, on the parabola $y = x^2$, $f(x,y) = \frac{1}{2}$, and when we approach the origin on this parabola, the limit is $\frac{1}{2}$, but $f(0,0) = 0$.

To avoid the problems arising with directional derivatives we introduce a more uniform notion.

Given two normed spaces E and F, recall that a linear map $f\colon E \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u)\| \leq C\,\|u\| \quad \text{for all } u \in E.$$

Definition 39.3. Let E and F be two normed affine spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, we say that f is *differentiable at* $a \in A$ if there is a linear continuous map $L\colon \overrightarrow{E} \to \overrightarrow{F}$ and a function ϵ, such that

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for every $a + h \in A$, where $\epsilon(h)$ is defined for every h such that $a + h \in A$ and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0,$$

where $U = \{h \in \overrightarrow{E} \mid a + h \in A,\ h \neq 0\}$. The linear map L is denoted by $\mathrm{D}f(a)$, or $\mathrm{D}f_a$, or $df(a)$, or df_a, or $f'(a)$, and it is called the *Fréchet derivative*, or *derivative*, or *total derivative*, or *total differential*, or *differential*, of f at a; see Figure 39.3.

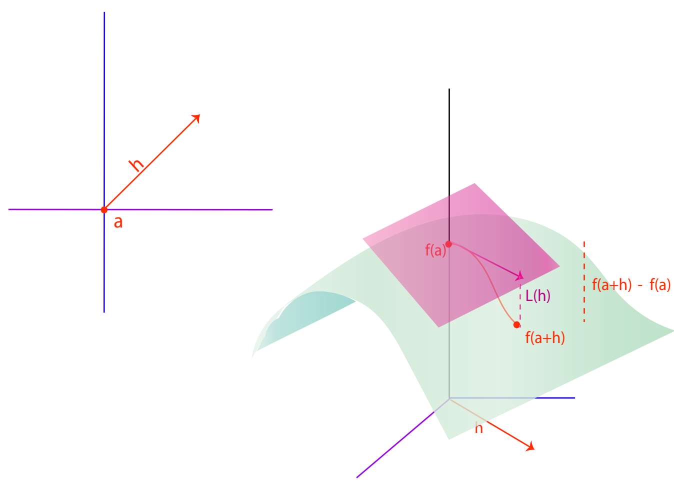


Figure 39.3: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the green surface in $\mathbb{R}^3$. The linear map $L = \mathrm{D}f(a)$ is the pink tangent plane. For any vector $h \in \mathbb{R}^2$, $L(h)$ is approximately equal to $f(a+h) - f(a)$. Note that $L(h)$ is also the direction tangent to the curve $t \mapsto f(a+tu)$.

Since the map $h \mapsto a + h$ from $\overrightarrow{E}$ to E is continuous, and since A is open in E, the inverse image U of $A - \{a\}$ under the above map is open in $\overrightarrow{E}$, and it makes sense to say that

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Note that for every $h \in U$, since $h \neq 0$, $\epsilon(h)$ is uniquely determined since

$$\epsilon(h) = \frac{f(a+h) - f(a) - L(h)}{\|h\|},$$

and that the value $\epsilon(0)$ plays absolutely no role in this definition. The condition for f to be differentiable at a amounts to the fact that

$$\lim_{h\mapsto 0} \frac{\|f(a+h) - f(a) - L(h)\|}{\|h\|} = 0$$

as $h \neq 0$ approaches 0, when $a + h \in A$. However, it does no harm to assume that $\epsilon(0) = 0$, and we will assume this from now on.

Again, we note that the derivative $\mathrm{D}f(a)$ of f at a provides an affine approximation of f, locally around a.

Remarks:

(1) Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

(2) If $h\colon (-a, a) \to \mathbb{R}$ is a real-valued function defined on some open interval containing 0, we say that h *is* $o(t)$ *for* $t \to 0$, and we write $h(t) = o(t)$, if

$$\lim_{t\mapsto 0,\, t\neq 0} \frac{h(t)}{t} = 0.$$

With this notation (the *little o notation*), the function f is differentiable at a iff

$$f(a+h) - f(a) - L(h) = o(\|h\|),$$

which is also written as

$$f(a+h) = f(a) + L(h) + o(\|h\|).$$

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that *the continuous linear map L is unique*, if it exists.

Proposition 39.1. *Let E and F be two normed affine spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f(a)$ is defined, then f is continuous at a and f has a directional derivative $\mathrm{D}_u f(a)$ for every $u \neq 0$ in $\overrightarrow{E}$, and furthermore,*

$$\mathrm{D}_u f(a) = \mathrm{D}f(a)(u).$$

Proof. If $L = \mathrm{D}f(a)$ exists, then for any nonzero vector $u \in \overrightarrow{E}$, because A is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$\begin{aligned} f(a+tu) &= f(a) + L(tu) + \epsilon(tu)\|tu\| \\ &= f(a) + tL(u) + |t|\epsilon(tu)\|u\| \end{aligned}$$

which implies that
$$L(u) = \frac{f(a+tu) - f(a)}{t} - \frac{|t|}{t}\epsilon(tu)\|u\|,$$
and since $\lim_{t\mapsto 0}\epsilon(tu) = 0$, we deduce that
$$L(u) = \mathrm{D}f(a)(u) = \mathrm{D}_u f(a).$$
Because
$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$
for all h such that $\|h\|$ is small enough, L is continuous, and $\lim_{h\mapsto 0}\epsilon(h)\|h\| = 0$, we have $\lim_{h\mapsto 0} f(a+h) = f(a)$, that is, f is continuous at a. $\square$

When E is of finite dimension, every linear map is continuous (see Proposition 9.8 or Theorem 37.58), and this assumption is then redundant.

It is important to note that the derivative $\mathrm{D}f(a)$ of f at a is a continuous linear map from the *vector space* $\overrightarrow{E}$ to the *vector space* $\overrightarrow{F}$, and not a function from the affine space E to the affine space F.

Although this may not be immediately obvious, the reason for requiring the linear map $\mathrm{D}f_a$ to be continuous is to ensure that if a function f is differentiable at a, then it is continuous at a. This is certainly a desirable property of a differentiable function. In finite dimension this holds, but in infinite dimension this is not the case. The following proposition shows that if $\mathrm{D}f_a$ exists at a and if f is continuous at a, then $\mathrm{D}f_a$ must be a continuous map. So if a function is differentiable at a, then it is continuous iff the linear map $\mathrm{D}f_a$ is continuous. We chose to include the second condition rather that the first in the definition of a differentiable function.

Proposition 39.2. *Let E and F be two normed affine spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f_a$ is defined, then f is continuous at a iff $\mathrm{D}f_a$ is a continuous linear map.*

Proof. Proposition 39.1 shows that if $\mathrm{D}f_a$ is defined and continuous then f is continuous at a. Conversely, assume that $\mathrm{D}f_a$ exists and that f is continuous at a. Since f is continuous at a and since $\mathrm{D}f_a$ exists, for any $\eta > 0$ there is some ρ with $0 < \rho < 1$ such that if $\|h\| \leq \rho$ then
$$\|f(a+h) - f(a)\| \leq \frac{\eta}{2},$$
and
$$\|f(a+h) - f(a) - \mathrm{D}_a(h)\| \leq \frac{\eta}{2}\,\|h\| \leq \frac{\eta}{2},$$
so we have
$$\begin{aligned}\|\mathrm{D}_a(h)\| &= \|\mathrm{D}_a(h) - (f(a+h) - f(a)) + f(a+h) - f(a)\| \\ &\leq \|f(a+h) - f(a) - \mathrm{D}_a(h)\| + \|f(a+h) - f(a)\| \\ &\leq \frac{\eta}{2} + \frac{\eta}{2} = \eta,\end{aligned}$$

which proves that $\mathrm{D}f_a$ is continuous at 0. By Proposition 37.56, $\mathrm{D}f_a$ is a continuous linear map. $\square$

Example 39.2. Consider the map, $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$, given by

$$f(A) = A^\top A - I,$$

where $\mathrm{M}_n(\mathbb{R})$ is equipped with any matrix norm, since they are all equivalent; for example, pick the Frobenius norm, $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$. We claim that

$$Df(A)(H) = A^\top H + H^\top A, \quad \text{for all } A \text{ and } H \text{ in } \mathrm{M}_n(\mathbb{R}).$$

We have

$$\begin{aligned} f(A+H) - f(A) - (A^\top H + H^\top A) &= (A+H)^\top (A+H) - I - (A^\top A - I) - A^\top H - H^\top A \\ &= A^\top A + A^\top H + H^\top A + H^\top H - A^\top A - A^\top H - H^\top A \\ &= H^\top H. \end{aligned}$$

It follows that

$$\epsilon(H) = \frac{f(A+H) - f(A) - (A^\top H + H^\top A)}{\|H\|} = \frac{H^\top H}{\|H\|},$$

and since our norm is the Frobenius norm,

$$\|\epsilon(H)\| = \left\| \frac{H^\top H}{\|H\|} \right\| \leq \frac{\left\|H^\top\right\| \|H\|}{\|H\|} = \left\|H^\top\right\| = \|H\|,$$

so

$$\lim_{H \mapsto 0} \epsilon(H) = 0,$$

and we conclude that

$$Df(A)(H) = A^\top H + H^\top A.$$

If $\mathrm{D}f(a)$ exists for every $a \in A$, we get a map

$$\mathrm{D}f\colon A \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{F}),$$

called the *derivative of f on A*, and also denoted by df. Recall that $\mathcal{L}(\overrightarrow{E}; \overrightarrow{F})$ denotes the vector space of all continuous maps from $\overrightarrow{E}$ to $\overrightarrow{F}$.

We now consider a number of standard results about derivatives.

39.2 Properties of Derivatives

Proposition 39.3. *Given two normed affine spaces E and F, if $f\colon E \to F$ is a constant function, then $\mathrm{D}f(a) = 0$, for every $a \in E$. If $f\colon E \to F$ is a continuous affine map, then $\mathrm{D}f(a) = \overrightarrow{f}$, for every $a \in E$, the linear map associated with f.*

Proof. Straightforward. ☐

Proposition 39.4. *Given a normed affine space E and a normed vector space F, for any two functions $f, g\colon E \to F$, for every $a \in E$, if $\mathrm{D}f(a)$ and $\mathrm{D}g(a)$ exist, then $\mathrm{D}(f+g)(a)$ and $\mathrm{D}(\lambda f)(a)$ exist, and*

$$\begin{aligned} \mathrm{D}(f+g)(a) &= \mathrm{D}f(a) + \mathrm{D}g(a), \\ \mathrm{D}(\lambda f)(a) &= \lambda \mathrm{D}f(a). \end{aligned}$$

Proof. Straightforward. ☐

Given two normed vector spaces $(E_1, \|\ \|_1)$ and $(E_2, \|\ \|_2)$, there are three natural and equivalent norms that can be used to make $E_1 \times E_2$ into a normed vector space:

1. $\|(u_1, u_2)\|_1 = \|u_1\|_1 + \|u_2\|_2$.
2. $\|(u_1, u_2)\|_2 = (\|u_1\|_1^2 + \|u_2\|_2^2)^{1/2}$.
3. $\|(u_1, u_2)\|_\infty = \max(\|u_1\|_1, \|u_2\|_2)$.

We usually pick the first norm. If E_1, E_2, and F are three normed vector spaces, recall that a bilinear map $f\colon E_1 \times E_2 \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u_1, u_2)\| \leq C \|u_1\|_1 \|u_2\|_2 \quad \text{for all } u_1 \in E_1 \text{ and all } u_2 \in E_2.$$

Proposition 39.5. *Given three normed vector spaces E_1, E_2, and F, for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $\mathrm{D}f(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,*

$$\mathrm{D}f(a, b)(u, v) = f(u, b) + f(a, v).$$

Proof. Since f is bilinear, a simple computation implies that

$$\begin{aligned} f((a, b) + (u, v)) - f(a, b) - (f(u, b) + f(a, v)) &= f(a + u, b + v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(a + u, b) + f(a + u, v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(a, b) + f(u, b) + f(a, v) + f(u, v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(u, v). \end{aligned}$$

We define
$$\epsilon(u,v) = \frac{f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))}{\|(u,v)\|_1},$$
and observe that the continuity of f implies
$$\begin{aligned}\|f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))\| &= \|f(u,v)\| \\ &\leq C\,\|u\|_1\,\|v\|_2 \leq C\left(\|u\|_1 + \|v\|_2\right)^2.\end{aligned}$$
Hence
$$\|\epsilon(u,v)\| = \left\|\frac{f(u,v)}{\|(u,v)\|_1}\right\| = \frac{\|f(u,v)\|}{\|(u,v)\|_1} \leq \frac{C\left(\|u\|_1+\|v\|_2\right)^2}{\|u\|_1+\|v\|_2} = C\left(\|u\|_1+\|v\|_2\right) = C\,\|(u,v)\|_1,$$
which in turn implies
$$\lim_{(u,v)\mapsto(0,0)} \epsilon(u,v) = 0.$$
□

We now state the very useful *chain rule*.

Theorem 39.6. *Given three normed affine spaces E, F, and G, let A be an open set in E, and let B an open set in F. For any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(f(a))$ exists, then $\mathrm{D}(g\circ f)(a)$ exists, and*
$$\mathrm{D}(g\circ f)(a) = \mathrm{D}g(f(a)) \circ \mathrm{D}f(a).$$

Proof. Since f is differentiable at a and g is differentiable at $b = f(a)$ for every η such that $0 < \eta < 1$ there is some $\rho > 0$ such that for all s, t, if $\|s\| \leq \rho$ and $\|t\| \leq \rho$ then
$$\begin{aligned}f(a+s) &= f(a) + \mathrm{D}f_a(s) + \epsilon_1(s)\\ g(b+t) &= g(b) + \mathrm{D}g_b(t) + \epsilon_2(t),\end{aligned}$$
with $\|\epsilon_1(s)\| \leq \eta\,\|s\|$ and $\|\epsilon_2(t)\| \leq \eta\,\|t\|$. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, we have
$$\|\mathrm{D}f_a(s)\| \leq \|\mathrm{D}f_a\|\,\|s\| \quad \text{and} \quad \|\mathrm{D}g_b(t)\| \leq \|\mathrm{D}g_b\|\,\|t\|,$$
which, since $\|\epsilon_1(s)\| \leq \eta\,\|s\|$ and $\eta < 1$, implies that
$$\|\mathrm{D}f_a(s) + \epsilon_1(s)\| \leq \|\mathrm{D}f_a\|\,\|s\| + \|\epsilon_1(s)\| \leq \|\mathrm{D}f_a\|\,\|s\| + \eta\,\|s\| \leq (\|\mathrm{D}f_a\|+1)\,\|s\|.$$
Consequently, if $\|s\| < \rho/(\|\mathrm{D}f_a\|+1)$, we have
$$\|\epsilon_2(\mathrm{D}f_a(s)+\epsilon_1(s))\| \leq \eta(\|\mathrm{D}f_a\|+1)\,\|s\| \tag{$*_1$}$$
and
$$\|\mathrm{D}g_b(\epsilon_1(s))\| \leq \|\mathrm{D}g_b\|\,\|\epsilon_1(s)\| \leq \eta\,\|\mathrm{D}g_b\|\,\|s\|. \tag{$*_2$}$$

Then since $b = f(a)$, using the above we have

$$\begin{aligned}(g \circ f)(a+s) &= g(f(a+s)) = g(b + \mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + \mathrm{D}g_b(\mathrm{D}f_a(s) + \epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)).\end{aligned}$$

Now by $(*_1)$ and $(*_2)$ we have

$$\begin{aligned}\|\mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| &\leq \|\mathrm{D}g_b(\epsilon_1(s))\| + \|\epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| \\ &\leq \eta \, \|\mathrm{D}g_b\| \, \|s\| + \eta(\|\mathrm{D}f_a\| + 1) \, \|s\| \\ &= \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \, \|s\| \, ,\end{aligned}$$

so if we write $\epsilon_3(s) = \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))$ we proved that

$$(g \circ f)(a+s) = g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \epsilon_3(s)$$

with $\epsilon_3(s) \leq \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \, \|s\|$, which proves that $\mathrm{D}g_b \circ \mathrm{D}f_a$ is the derivative of $g \circ f$ at a. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, so is $\mathrm{D}g_b \circ \mathrm{D}f_a$, which proves our proposition. $\square$

Theorem 39.6 has many interesting consequences. We mention two corollaries.

Proposition 39.7. *Given three normed affine spaces E, F, and G, for any open subset A in E, for any $a \in A$, let $f\colon A \to F$ such that $\mathrm{D}f(a)$ exists, and let $g\colon F \to G$ be a continuous affine map. Then, $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \overrightarrow{g} \circ \mathrm{D}f(a),$$

where $\overrightarrow{g}$ is the linear map associated with the affine map g.

Proposition 39.8. *Given two normed affine spaces E and F, let A be some open subset in E, let B be some open subset in F, let $f\colon A \to B$ be a bijection from A to B, and assume that $\mathrm{D}f$ exists on A and that $\mathrm{D}f^{-1}$ exists on B. Then, for every $a \in A$,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

Proposition 39.8 has the remarkable consequence that the two vector spaces $\overrightarrow{E}$ and $\overrightarrow{F}$ have the same dimension. In other words, a local property, the existence of a bijection f between an open set A of E and an open set B of F, such that f is differentiable on A and f^{-1} is differentiable on B, implies a global property, that the two vector spaces $\overrightarrow{E}$ and $\overrightarrow{F}$ have the same dimension.

Let us mention two more rules about derivatives that are used all the time.

Let $\iota\colon \mathbf{GL}(n, \mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the function (inversion) defined on invertible $n \times n$ matrices by

$$\iota(A) = A^{-1}.$$

Observe that $\mathbf{GL}(n, \mathbb{C})$ is indeed an open subset of the normed vector space $\mathrm{M}_n(\mathbb{C})$ of complex $n \times n$ matrices, since its complement is the closed set of matrices $A \in \mathrm{M}_n(\mathbb{C})$ satisfying $\det(A) = 0$. Then we have

$$d\iota_A(H) = -A^{-1}HA^{-1},$$

for all $A \in \mathbf{GL}(n, \mathbb{C})$ and for all $H \in \mathrm{M}_n(\mathbb{C})$.

To prove the preceding line observe that for H with sufficiently small norm, we have

$$\begin{aligned}\iota(A+H) - \iota(A) + A^{-1}HA^{-1} &= (A+H)^{-1} - A^{-1} + A^{-1}HA^{-1} \\ &= (A+H)^{-1}[I - (A+H)A^{-1} + (A+H)A^{-1}HA^{-1}] \\ &= (A+H)^{-1}[I - I - HA^{-1} + HA^{-1} + HA^{-1}HA^{-1}] \\ &= (A+H)^{-1}HA^{-1}HA^{-1}.\end{aligned}$$

Consequently, we get

$$\epsilon(H) = \frac{\iota(A+H) - \iota(A) + A^{-1}HA^{-1}}{\|H\|} = \frac{(A+H)^{-1}HA^{-1}HA^{-1}}{\|H\|},$$

and since

$$\left\|(A+H)^{-1}HA^{-1}HA^{-1}\right\| \le \|H\|^2 \left\|A^{-1}\right\|^2 \left\|(A+H)^{-1}\right\|,$$

it is clear that $\lim_{H \mapsto 0} \epsilon(H) = 0$, which proves that

$$d\iota_A(H) = -A^{-1}HA^{-1}.$$

In particular, if $A = I$, then $d\iota_I(H) = -H$.

Next, if $f\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ and $g\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ are differentiable matrix functions, then

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathbf{M}_n(\mathbb{C})$. This is known as the *product rule*.

When E is of finite dimension n, for any frame $(a_0, (u_1, \ldots, u_n))$ of E, where $(u_1, \ldots, u_n)$ is a basis of $\overrightarrow{E}$, we can define the directional derivatives with respect to the vectors in the basis $(u_1, \ldots, u_n)$ (actually, we can also do it for an infinite frame). This way, we obtain the definition of partial derivatives, as follows.

Definition 39.4. For any two normed affine spaces E and F, if E is of finite dimension n, for every frame $(a_0, (u_1, \ldots, u_n))$ for E, for every $a \in E$, for every function $f\colon E \to F$, the directional derivatives $\mathrm{D}_{u_j} f(a)$ (if they exist) are called the *partial derivatives of f with respect to the frame* $(a_0, (u_1, \ldots, u_n))$. The partial derivative $\mathrm{D}_{u_j} f(a)$ is also denoted by $\partial_j f(a)$, or $\dfrac{\partial f}{\partial x_j}(a)$.

The notation $\dfrac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a "logical obscenity." Indeed, the variable x_j really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!

We now consider the situation where the normed affine space F is a finite direct sum $F = (F_1, b_1) \oplus \cdots \oplus (F_m, b_m)$.

Proposition 39.9. *Given normed affine spaces E and $F = (F_1, b_1) \oplus \cdots \oplus (F_m, b_m)$, given any open subset A of E, for any $a \in A$, for any function $f \colon A \to F$, letting $f = (f_1, \ldots, f_m)$, $\mathrm{D}f(a)$ exists iff every $\mathrm{D}f_i(a)$ exists, and*

$$\mathrm{D}f(a) = in_1 \circ \mathrm{D}f_1(a) + \cdots + in_m \circ \mathrm{D}f_m(a).$$

Proof. Observe that $f(a+h) - f(a) = (f(a+h) - b) - (f(a) - b)$, where $b = (b_1, \ldots, b_m)$, and thus, as far as dealing with derivatives, $\mathrm{D}f(a)$ is equal to $\mathrm{D}f_b(a)$, where $f_b \colon E \to \overrightarrow{F}$ is defined such that $f_b(x) = f(x) - b$, for every $x \in E$. Thus, we can work with the vector space $\overrightarrow{F}$ instead of the affine space F. The proposition is then a simple application of Theorem 39.6. ☐

In the special case where F is a normed affine space of finite dimension m, for any frame $(b_0, (v_1, \ldots, v_m))$ of F, where $(v_1, \ldots, v_m)$ is a basis of $\overrightarrow{F}$, every point $x \in F$ can be expressed uniquely as

$$x = b_0 + x_1 v_1 + \cdots + x_m v_m,$$

where $(x_1, \ldots, x_m) \in K^m$, the coordinates of x in the frame $(b_0, (v_1, \ldots, v_m))$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$). Thus, letting F_i be the standard normed affine space K with its natural structure, we note that F is isomorphic to the direct sum $F = (K, 0) \oplus \cdots \oplus (K, 0)$. Then, every function $f \colon E \to F$ is represented by m functions $(f_1, \ldots, f_m)$, where $f_i \colon E \to K$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$), and

$$f(x) = b_0 + f_1(x) v_1 + \cdots + f_m(x) v_m,$$

for every $x \in E$. The following proposition is an immediate corollary of Proposition 39.9.

Proposition 39.10. *For any two normed affine spaces E and F, if F is of finite dimension m, for any frame $(b_0, (v_1, \ldots, v_m))$ of F, where $(v_1, \ldots, v_m)$ is a basis of $\overrightarrow{F}$, for every $a \in E$, a function $f \colon E \to F$ is differentiable at a iff each f_i is differentiable at a, and*

$$\mathrm{D}f(a)(u) = \mathrm{D}f_1(a)(u) v_1 + \cdots + \mathrm{D}f_m(a)(u) v_m,$$

for every $u \in \overrightarrow{E}$.

We now consider the situation where E is a finite direct sum. Given a normed affine space $E = (E_1, a_1) \oplus \cdots \oplus (E_n, a_n)$ and a normed affine space F, given any open subset A of E, for any $c = (c_1, \ldots, c_n) \in A$, we define the continuous functions $i_j^c \colon E_j \to E$, such that

$$i_j^c(x) = (c_1, \ldots, c_{j-1}, x, c_{j+1}, \ldots, c_n).$$

For any function $f \colon A \to F$, we have functions $f \circ i_j^c \colon E_j \to F$, defined on $(i_j^c)^{-1}(A)$, which contains c_j. If $\mathrm{D}(f \circ i_j^c)(c_j)$ exists, we call it the *partial derivative of f w.r.t. its jth argument, at c.* We also denote this derivative by $\mathrm{D}_j f(c)$. Note that $\mathrm{D}_j f(c) \in \mathcal{L}(\overrightarrow{E_j}; \overrightarrow{F})$.

This notion is a generalization of the notion defined in Definition 39.4. In fact, when E is of dimension n, and a frame $(a_0, (u_1, \ldots, u_n))$ has been chosen, we can write $E = (E_1, a_1) \oplus \cdots \oplus (E_n, a_n)$, for some obvious (E_j, a_j) (as explained just after Proposition 39.9), and then

$$\mathrm{D}_j f(c)(\lambda u_j) = \lambda \partial_j f(c),$$

and the two notions are consistent.

The definition of i_j^c and of $\mathrm{D}_j f(c)$ also makes sense for a finite product $E_1 \times \cdots \times E_n$ of affine spaces E_i. We will use freely the notation $\partial_j f(c)$ instead of $\mathrm{D}_j f(c)$.

The notion $\partial_j f(c)$ introduced in Definition 39.4 is really that of the vector derivative, whereas $\mathrm{D}_j f(c)$ is the corresponding linear map. Although perhaps confusing, we identify the two notions. The following proposition holds.

Proposition 39.11. *Given a normed affine space $E = (E_1, a_1) \oplus \cdots \oplus (E_n, a_n)$, and a normed affine space F, given any open subset A of E, for any function $f \colon A \to F$, for every $c \in A$, if $\mathrm{D}f(c)$ exists, then each $\mathrm{D}_j f(c)$ exists, and*

$$\mathrm{D}f(c)(u_1, \ldots, u_n) = \mathrm{D}_1 f(c)(u_1) + \cdots + \mathrm{D}_n f(c)(u_n),$$

for every $u_i \in E_i$, $1 \leq i \leq n$. The same result holds for the finite product $E_1 \times \cdots \times E_n$.

Proof. Since every $c \in E$ can be written as $c = a + c - a$, where $a = (a_1, \ldots, a_n)$, defining $f_a \colon \overrightarrow{E} \to F$ such that, $f_a(u) = f(a + u)$, for every $u \in \overrightarrow{E}$, clearly, $\mathrm{D}f(c) = \mathrm{D}f_a(c - a)$, and thus, we can work with the function f_a whose domain is the vector space $\overrightarrow{E}$. The proposition is then a simple application of Theorem 39.6. $\square$

39.3 Jacobian Matrices

If both E and F are of finite dimension, for any frame $(a_0, (u_1, \ldots, u_n))$ of E and any frame $(b_0, (v_1, \ldots, v_m))$ of F, every function $f \colon E \to F$ is determined by m functions $f_i \colon E \to \mathbb{R}$ (or $f_i \colon E \to \mathbb{C}$), where

$$f(x) = b_0 + f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. From Proposition 39.1, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}_{u_j} f(a) = \partial_j f(a),$$

and from Proposition 39.10, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}f_1(a)(u_j)v_1 + \cdots + \mathrm{D}f_i(a)(u_j)v_i + \cdots + \mathrm{D}f_m(a)(u_j)v_m,$$

that is,

$$\mathrm{D}f(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.$$

Since the j-th column of the $m\times n$-matrix representing $\mathrm{D}f(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ is equal to the components of the vector $\mathrm{D}f(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $\mathrm{D}f(a)$ is determined by the $m \times n$-matrix $J(f)(a) = (\partial_j f_i(a))$, (or $J(f)(a) = (\frac{\partial f_i}{\partial x_j}(a))$):

$$J(f)(a) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \ldots & \partial_n f_m(a) \end{pmatrix}$$

or

$$J(f)(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \ldots & \frac{\partial f_1}{\partial x_n}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \ldots & \frac{\partial f_2}{\partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \frac{\partial f_m}{\partial x_2}(a) & \ldots & \frac{\partial f_m}{\partial x_n}(a) \end{pmatrix}$$

Definition 39.5. The matrix $J(f)(a)$ is called the *Jacobian matrix* of $\mathrm{D}f$ at a. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the *Jacobian* of $\mathrm{D}f(a)$.

From a previous result, we know that this determinant in fact only depends on $\mathrm{D}f(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f\colon \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $\frac{\partial f_i}{\partial x_j}(a)$. We simply treat the function $f_i\colon \mathbb{R}^n \to \mathbb{R}$ as a function of its j-th argument, leaving the others fixed, and compute the derivative as in Definition 39.1, that is, the usual derivative.

Example 39.3. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, defined such that

$$f(r, \theta) = (r\cos(\theta), r\sin(\theta)).$$

Then, we have

$$J(f)(r,\theta) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}$$

and the Jacobian (determinant) has value $\det(J(f)(r,\theta)) = r$.

In the case where $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), the Jacobian matrix of $\mathrm{D}f(a)$ is a column vector. In fact, this column vector is just $\mathrm{D}_1 f(a)$. Then, for every $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$),

$$\mathrm{D}f(a)(\lambda) = \lambda \mathrm{D}_1 f(a).$$

This case is sufficiently important to warrant a definition.

Definition 39.6. Given a function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), where F is a normed affine space, the vector

$$\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$$

is called the *vector derivative or velocity vector (in the real case)* at a. We usually identify $\mathrm{D}f(a)$ with its Jacobian matrix $\mathrm{D}_1 f(a)$, which is the column vector corresponding to $\mathrm{D}_1 f(a)$. By abuse of notation, we also let $\mathrm{D}f(a)$ denote the vector $\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$.

When $E = \mathbb{R}$, the physical interpretation is that f defines a (parametric) curve that is the trajectory of some particle moving in $\mathbb{R}^m$ as a function of time, and the vector $\mathrm{D}_1 f(a)$ is the *velocity* of the moving particle $f(t)$ at $t = a$.

It is often useful to consider functions $f\colon [a,b] \to F$ from a closed interval $[a,b] \subseteq \mathbb{R}$ to a normed affine space F, and its derivative $\mathrm{D}f(a)$ on $[a,b]$, even though $[a,b]$ is not open. In this case, as in the case of a real-valued function, we define the right derivative $\mathrm{D}_1 f(a_+)$ at a, and the left derivative $\mathrm{D}_1 f(b_-)$ at b, and we assume their existence.

Example 39.4.

1. When $A = (0,1)$ and $F = \mathbb{R}^3$, a function $f\colon (0,1) \to \mathbb{R}^3$ defines a (parametric) curve in $\mathbb{R}^3$. If $f = (f_1, f_2, f_3)$, its Jacobian matrix at $a \in \mathbb{R}$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial t}(a) \\ \dfrac{\partial f_2}{\partial t}(a) \\ \dfrac{\partial f_3}{\partial t}(a) \end{pmatrix}.$$

See Figure 39.4.

The velocity vectors $J(f)(a) = \begin{pmatrix} -\sin(t) \\ \cos(t) \\ 1 \end{pmatrix}$ are represented by the blue arrows.

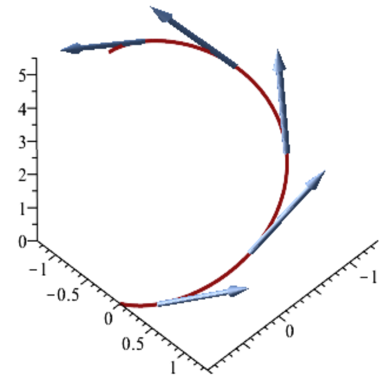


Figure 39.4: The red space curve $f(t) = (\cos(t), \sin(t), t)$.

2. When $E = \mathbb{R}^2$ and $F = \mathbb{R}^3$, a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ defines a parametric surface. Letting $\varphi = (f, g, h)$, its Jacobian matrix at $a \in \mathbb{R}^2$ is

$$J(\varphi)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial u}(a) & \dfrac{\partial f}{\partial v}(a) \\ \dfrac{\partial g}{\partial u}(a) & \dfrac{\partial g}{\partial v}(a) \\ \dfrac{\partial h}{\partial u}(a) & \dfrac{\partial h}{\partial v}(a) \end{pmatrix}.$$

See Figure 39.5. The Jacobian matrix is $J(f)(a) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2u & 2v \end{pmatrix}$. The first column is the vector tangent to the pink u-direction curve, while the second column is the vector tangent to the blue v-direction curve.

3. When $E = \mathbb{R}^3$ and $F = \mathbb{R}$, for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(a) & \dfrac{\partial f}{\partial y}(a) & \dfrac{\partial f}{\partial z}(a) \end{pmatrix}.$$

More generally, when $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) & \cdots & \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix}.$$

Its transpose is a column vector called the *gradient* of f at a, denoted by $\mathrm{grad} f(a)$ or $\nabla f(a)$. Then, given any $v \in \mathbb{R}^n$, note that

$$\mathrm{D}f(a)(v) = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n = \mathrm{grad} f(a) \cdot v,$$

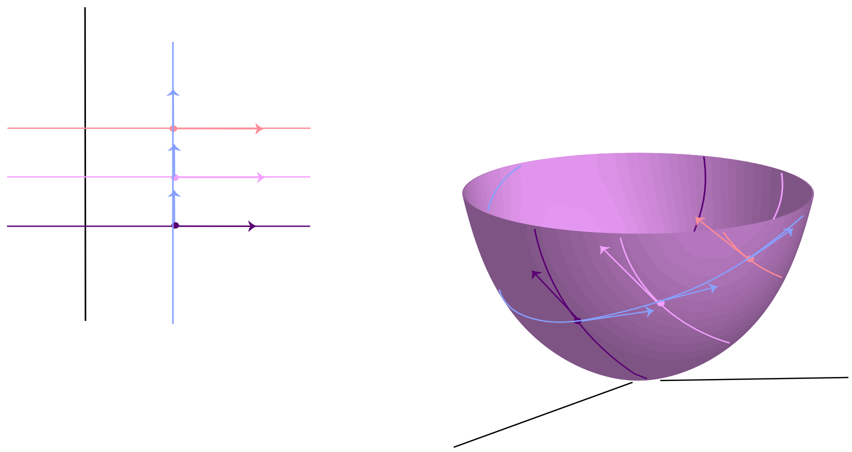


Figure 39.5: The parametric surface $x = u, y = v, z = u^2 + v^2$.

the scalar product of $\mathrm{grad} f(a)$ and v.

Example 39.5. Consider the quadratic function $f\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = x^\top A x, \quad x \in \mathbb{R}^n,$$

where A is a real $n \times n$ symmetric matrix. We claim that

$$df_u(h) = 2u^\top A h \quad \text{for all } u, h \in \mathbb{R}^n.$$

Since A is symmetric, we have

$$\begin{aligned} f(u + h) &= (u^\top + h^\top) A (u + h) \\ &= u^\top A u + u^\top A h + h^\top A u + h^\top A h \\ &= u^\top A u + 2u^\top A h + h^\top A h, \end{aligned}$$

so we have

$$f(u + h) - f(u) - 2u^\top A h = h^\top A h.$$

If we write

$$\epsilon(h) = \frac{h^\top A h}{\|h\|}$$

for $h \notin 0$ where $\| \, \|$ is the 2-norm, by Cauchy–Schwarz we have

$$|\epsilon(h)| \leq \frac{\|h\| \, \|Ah\|}{\|h\|} \leq \frac{\|h\|^2 \, \|A\|}{\|h\|} = \|h\| \, \|A\| \, ,$$

which shows that $\lim_{h \mapsto 0} \epsilon(h) = 0$. Therefore,

$$df_u(h) = 2u^\top A h \quad \text{for all } u, h \in \mathbb{R}^n,$$

as claimed. This formula shows that the gradient ∇f_u of f at u is given by

$$\nabla f_u = 2Au.$$

As a first corollary we obtain the gradient of a function of the form

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

where A is a symmetric $n \times n$ matrix and b is some vector $b \in \mathbb{R}^n$. Since the derivative of a linear function is itself, we obtain

$$df_u(h) = u^\top Ah - b^\top h,$$

and the gradient of f is given by

$$\nabla f_u = Au - b.$$

As a second corollary we obtain the gradient of the function

$$f(x) = \|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b) = (x^\top A^\top - b^\top)(Ax - b)$$

which is the function to minimize in a least squares problem, where A is an $m \times n$ matrix. We have

$$f(x) = x^\top A^\top Ax - x^\top A^\top b - b^\top Ax + b^\top b = x^\top A^\top Ax - 2b^\top Ax + b^\top b,$$

and since the derivative of a constant function is 0 and the derivative of a linear function is itself, we get

$$df_u(h) = 2u^\top A^\top Ah - 2b^\top Ah.$$

Consequently, the gradient of f is given by

$$\nabla f_u = 2A^\top Au - 2A^\top b.$$

When E, F, and G have finite dimensions, and $(a_0, (u_1, \ldots, u_p))$ is an affine frame for E, $(b_0, (v_1, \ldots, v_n))$ is an affine frame for F, and $(c_0, (w_1, \ldots, w_m))$ is an affine frame for G, if A is an open subset of E, B is an open subset of F, for any functions $f \colon A \to F$ and $g \colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(b)$ exists, by Theorem 39.6, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \partial_1 g_1(b) & \partial_2 g_1(b) & \ldots & \partial_n g_1(b) \\ \partial_1 g_2(b) & \partial_2 g_2(b) & \ldots & \partial_n g_2(b) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 g_m(b) & \partial_2 g_m(b) & \ldots & \partial_n g_m(b) \end{pmatrix} \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_p f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_p f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_n(a) & \partial_2 f_n(a) & \ldots & \partial_p f_n(a) \end{pmatrix}$$

or

$$J(h)(a) = \begin{pmatrix} \frac{\partial g_1}{\partial y_1}(b) & \frac{\partial g_1}{\partial y_2}(b) & \dots & \frac{\partial g_1}{\partial y_n}(b) \\ \frac{\partial g_2}{\partial y_1}(b) & \frac{\partial g_2}{\partial y_2}(b) & \dots & \frac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial y_1}(b) & \frac{\partial g_m}{\partial y_2}(b) & \dots & \frac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \dots & \frac{\partial f_1}{\partial x_p}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \dots & \frac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \frac{\partial f_n}{\partial x_2}(a) & \dots & \frac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$

Thus, we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^{k=n} \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed affine spaces E and F of finite dimension, given an open subset A of E, if a function $f\colon A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. There are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at a, and not even continuous at a. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$, defined such that $f(0,0) = 0$, and

$$f(x,y) = \frac{x^2 y}{x^4 + y^2} \qquad \text{if } (x,y) \neq (0,0).$$

For any $u \neq 0$, letting $u = \begin{pmatrix} h \\ k \end{pmatrix}$, we have

$$\frac{f(0 + tu) - f(0)}{t} = \frac{h^2 k}{t^2 h^4 + k^2},$$

so that

$$\mathrm{D}_u f(0,0) = \begin{cases} \frac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0. \end{cases}$$

Thus, $\mathrm{D}_u f(0,0)$ exists for all $u \neq 0$. On the other hand, if $\mathrm{D}f(0,0)$ existed, it would be a linear map $\mathrm{D}f(0,0)\colon \mathbb{R}^2 \to \mathbb{R}$ represented by a row matrix $(\alpha\ \ \beta)$, and we would have $\mathrm{D}_u f(0,0) = \mathrm{D}f(0,0)(u) = \alpha h + \beta k$, but the explicit formula for $\mathrm{D}_u f(0,0)$ is not linear. As a matter of fact, the function f is not continuous at $(0,0)$. For example, on the parabola $y = x^2$, $f(x,y) = \frac{1}{2}$, and when we approach the origin on this parabola, the limit is $\frac{1}{2}$, when in fact, $f(0,0) = 0$.

However, there are sufficient conditions on the partial derivatives for $\mathrm{D}f(a)$ to exist, namely, continuity of the partial derivatives.

If f is differentiable on A, then f defines a function $\mathrm{D}f\colon A \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{F})$. It turns out that the continuity of the partial derivatives on A is a necessary and sufficient condition for $\mathrm{D}f$ to exist and to be continuous on A.

If $f\colon [a,b] \to \mathbb{R}$ is a function which is continuous on $[a,b]$ and differentiable on (a,b), then there is some c with $a < c < b$ such that

$$f(b) - f(a) = (b-a)f'(c).$$

This result is known as the *mean value theorem* and is a generalization of *Rolle's theorem*, which corresponds to the case where $f(a) = f(b)$.

Unfortunately, the mean value theorem fails for vector-valued functions. For example, the function $f\colon [0, 2\pi] \to \mathbb{R}^2$ given by

$$f(t) = (\cos t, \sin t)$$

is such that $f(2\pi) - f(0) = (0,0)$, yet its derivative $f'(t) = (-\sin t, \cos t)$ does not vanish in $(0, 2\pi)$.

A suitable generalization of the mean value theorem to vector-valued functions is possible if we consider an inequality (an upper bound) instead of an equality. This generalized version of the mean value theorem plays an important role in the proof of several major results of differential calculus.

If E is an affine space (over $\mathbb{R}$ or $\mathbb{C}$), given any two points $a, b \in E$, the *closed segment* $[a,b]$ is the set of all points $a + \lambda(b-a)$, where $0 \leq \lambda \leq 1$, $\lambda \in \mathbb{R}$, and the *open segment* (a,b) is the set of all points $a + \lambda(b-a)$, where $0 < \lambda < 1$, $\lambda \in \mathbb{R}$.

Proposition 39.12. *Let E and F be two normed affine spaces, let A be an open subset of E, and let $f\colon A \to F$ be a continuous function on A. Given any $a \in A$ and any $h \neq 0$ in $\overrightarrow{E}$, if the closed segment $[a, a+h]$ is contained in A, if $f\colon A \to F$ is differentiable at every point of the open segment $(a, a+h)$, and*

$$\sup_{x \in (a, a+h)} \|\mathrm{D}f(x)\| \leq M,$$

for some $M \geq 0$, then

$$\|f(a+h) - f(a)\| \leq M\|h\|.$$

As a corollary, if $L\colon \overrightarrow{E} \to \overrightarrow{F}$ is a continuous linear map, then

$$\|f(a+h) - f(a) - L(h)\| \leq M\|h\|,$$

where $M = \sup_{x \in (a, a+h)} \|\mathrm{D}f(x) - L\|$.

The above proposition is sometimes called the "mean value theorem." Proposition 39.12 can be used to show the following important result.

Theorem 39.13. *Given two normed affine spaces E and F, where E is of finite dimension n, and where $(a_0, (u_1, \ldots, u_n))$ is a frame of E, given any open subset A of E, given any function $f\colon A \to F$, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{F})$ is defined and continuous on A iff every partial derivative $\partial_j f$ (or $\frac{\partial f}{\partial x_j}$) is defined and continuous on A, for all j, $1 \leq j \leq n$. As a corollary, if F is of finite dimension m, and $(b_0, (v_1, \ldots, v_m))$ is a frame of F, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{F})$ is defined and continuous on A iff every partial derivative $\partial_j f_i$ (or $\frac{\partial f_i}{\partial x_j}$) is defined and continuous on A, for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$.*

Theorem 39.13 gives a necessary and sufficient condition for the existence and continuity of the derivative of a function on an open set. It should be noted that a more general version of Theorem 39.13 holds, assuming that $E = (E_1, a_1) \oplus \cdots \oplus (E_n, a_n)$, or $E = E_1 \times \cdots \times E_n$, and using the more general partial derivatives $\mathrm{D}_j f$ introduced before Proposition 39.11.

Definition 39.7. Given two normed affine spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is *of class C^0 on A or a C^0-function on A* if f is continuous on A. We say that $f\colon A \to F$ is *of class C^1 on A or a C^1-function on A* if $\mathrm{D}f$ exists and is continuous on A.

Since the existence of the derivative on an open set implies continuity, a C^1-function is of course a C^0-function. Theorem 39.13 gives a necessary and sufficient condition for a function f to be a C^1-function (when E is of finite dimension). It is easy to show that the composition of C^1-functions (on appropriate open sets) is a C^1-function.

39.4 The Implicit and The Inverse Function Theorems

Given three normed affine spaces E, F, and G, given a function $f\colon E \times F \to G$, given any $c \in G$, it may happen that the equation

$$f(x, y) = c$$

has the property that, for some open sets $A \subseteq E$, and $B \subseteq F$, there is a function $g\colon A \to B$, such that

$$f(x, g(x)) = c,$$

for all $x \in A$. Such a situation is usually very rare, but if some solution $(a, b) \in E \times F$ such that $f(a, b) = c$ is known, under certain conditions, for some small open sets $A \subseteq E$ containing a and $B \subseteq F$ containing b, the existence of a unique $g\colon A \to B$, such that

$$f(x, g(x)) = c,$$

for all $x \in A$, can be shown. Under certain conditions, it can also be shown that g is continuous, and differentiable. Such a theorem, known as the *implicit function theorem*, can be proven.

Example 39.6. Let $E = \mathbb{R}^2$, $F = G = \mathbb{R}$, $\Omega = \mathbb{R}^2 \times \mathbb{R} \cong \mathbb{R}^3$, $f \colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ given by

$$f((x_1, x_2), x_3) = x_1^2 + x_2^2 + x_3^2 - 1,$$

$a = \big(\sqrt{3}/(2\sqrt{2}), \sqrt{3}/(2\sqrt{2})\big)$, $b = 1/2$, and $c = 0$. The set of vectors $(x_1, x_2, x_3) \in \mathbb{R}^2$ such that

$$f((x_1, x_2), x_3) = x_1^2 + x_2^2 + x_3^2 - 1 = 0$$

is the unit sphere in $\mathbb{R}^3$. The vector (a, b) belongs to the unit sphere since $\|a\|_2^2 + b^2 - 1 = 0$. The function $g \colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$g(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2}$$

satisfies the equation

$$f(x_1, x_2, g(x_1, x_2)) = 0$$

all for (x_1, x_2) in the open disk $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 < 1\}$, and $g(a) = b$. Observe that if we had picked $b = -1/2$, then we would need to consider the function

$$g(x_1, x_2) = -\sqrt{1 - x_1^2 - x_2^2}.$$

We now state a very general version of the implicit function theorem. The proof of this theorem is fairly involved and uses a fixed-point theorem for contracting mappings in complete metric spaces; it is given in Schwartz [151]. Other proofs can be found in Lang [111] and Cartan [34].

Theorem 39.14. *Let E, F, and G, be normed affine spaces, let Ω be an open subset of $E \times F$, let $f \colon \Omega \to G$ be a function defined on Ω, let $(a, b) \in \Omega$, let $c \in G$, and assume that $f(a, b) = c$. If the following assumptions hold*

(1) The function $f \colon \Omega \to G$ is continuous on Ω;

(2) F is a complete normed affine space (and so is G);

(3) $\dfrac{\partial f}{\partial y}(x, y)$ exists for every $(x, y) \in \Omega$, and $\dfrac{\partial f}{\partial y} \colon \Omega \to \mathcal{L}(\overrightarrow{F}; \overrightarrow{G})$ is continuous;

(4) $\dfrac{\partial f}{\partial y}(a, b)$ is a bijection of $\mathcal{L}(\overrightarrow{F}; \overrightarrow{G})$, and $\left(\dfrac{\partial f}{\partial y}(a, b)\right)^{-1} \in \mathcal{L}(\overrightarrow{G}; \overrightarrow{F})$;

then the following properties hold:

(a) There exist some open subset $A \subseteq E$ containing a and some open subset $B \subseteq F$ containing b, such that $A \times B \subseteq \Omega$, and for every $x \in A$, the equation $f(x, y) = c$ has a single solution $y = g(x)$, and thus, there is a unique function $g \colon A \to B$ such that $f(x, g(x)) = c$, for all $x \in A$;

(b) *The function* $g\colon A \to B$ *is continuous.*

If we also assume that

(5) *The derivative* $\mathrm{D}f(a,b)$ *exists;*

then

(c) *The derivative* $\mathrm{D}g(a)$ *exists, and*

$$\mathrm{D}g(a) = -\Big(\frac{\partial f}{\partial y}(a,b)\Big)^{-1} \circ \frac{\partial f}{\partial x}(a,b);$$

and if in addition

(6) $\dfrac{\partial f}{\partial x}\colon \Omega \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{G})$ *is also continuous (and thus, in view of (3),* f *is* C^1 *on* Ω*);*

then

(d) *The derivative* $\mathrm{D}g\colon A \to \mathcal{L}(\overrightarrow{E}; \overrightarrow{F})$ *is continuous, and*

$$\mathrm{D}g(x) = -\Big(\frac{\partial f}{\partial y}(x,g(x))\Big)^{-1} \circ \frac{\partial f}{\partial x}(x,g(x)),$$

for all $x \in A$.

Example 39.7. Going back to Example 39.6, write $x = (x_1, x_2)$ and $y = x_3$, so that the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ are given in terms of their Jacobian matrices by

$$\begin{aligned}\frac{\partial f}{\partial x}(x,y) &= \begin{pmatrix} 2x_1 & 2x_2 \end{pmatrix}\\ \frac{\partial f}{\partial y}(x,y) &= 2x_3.\end{aligned}$$

If $0 < |b| \leq 1$ and $\|a\|_2^2 + b^2 - 1 = 0$, then Conditions (3) and (4) are satisfied. Conditions (1) and (2) obviously hold. Since $df_{(a,b)}$ is given by its Jacobian matrix as

$$df_{(a,b)} = \begin{pmatrix} 2a_1 & 2a_2 & 2b \end{pmatrix},$$

Condition (5) holds, and clearly, Condition (6) also holds.

Theorem 39.14 implies that there is some open subset A of $\mathbb{R}^2$ containing a, some open subset B of $\mathbb{R}$ containing b, and a unique function $g\colon A \to B$ such that

$$f(x, g(x)) = 0$$

for all $x \in A$. In fact, we can pick A to be the open unit disk in $\mathbb{R}$, $B = (0,2)$, and if $0 < b \leq 1$, then

$$g(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2},$$

else if $-1 \leq b < 0$, then

$$g(x_1, x_2) = -\sqrt{1 - x_1^2 - x_2^2}.$$

Assuming $0 < b \leq 1$, We have

$$\frac{\partial f}{\partial x}(x, g(x)) = (2x_1 \ \ 2x_2),$$

and

$$\left(\frac{\partial f}{\partial y}(x, g(x))\right)^{-1} = \frac{1}{2\sqrt{1 - x_1^2 - x_2^2}},$$

so according to the theorem,

$$dg_x = -\frac{1}{\sqrt{1 - x_1^2 - x_2^2}}(x_1 \ \ x_2),$$

which matches the derivative of g computed directly.

Observe that the functions $(x_1, x_2) \mapsto \sqrt{1 - x_1^2 - x_2^2}$ and $(x_1, x_2) \mapsto -\sqrt{1 - x_1^2 - x_2^2}$ are two differentiable parametrizations of the sphere, but the union of their ranges does not cover the entire sphere. Since $b \neq 0$, none of the points on the unit circle in the (x_1, x_2)-plane are covered. Our function f views b as lying on the x_3-axis. In order to cover the entire sphere using this method, we need four more maps, which correspond to b lying on the x_1-axis or on the x_2 axis. Then we get the additional (implicit) maps $(x_2, x_3) \mapsto \pm\sqrt{1 - x_2^2 - x_3^2}$ and $(x_1, x_3) \mapsto \pm\sqrt{1 - x_1^2 - x_3^2}$.

The implicit function theorem plays an important role in the calculus of variations.

We now consider another very important notion, that of a (local) diffeomorphism.

Definition 39.8. Given two topological spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is a *local homeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a homeomorphism from U to $V = f(U)$. If B is an open subset of F, we say that $f\colon A \to F$ is a *(global) homeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$. If E and F are normed affine spaces, we say that $f\colon A \to F$ is a *local diffeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a bijection from U to V, f is a C^1-function on U, and f^{-1} is a C^1-function on $V = f(U)$. We say that $f\colon A \to F$ is a *(global) diffeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$, f is a C^1-function on A, and f^{-1} is a C^1-function on B.

Note that a local diffeomorphism is a local homeomorphism. Also, as a consequence of Proposition 39.8, if f is a diffeomorphism on A, then $\mathrm{D}f(a)$ is a linear isomorphism for every $a \in A$. The following theorem can be shown. In fact, there is a fairly simple proof using Theorem 39.14.

Theorem 39.15. *(Inverse Function Theorem) Let E and F be complete normed affine spaces, let A be an open subset of E, and let $f\colon A \to F$ be a C^1-function on A. The following properties hold:*

(1) For every $a \in A$, if $\mathrm{D}f(a)$ is a linear isomorphism (which means that both $\mathrm{D}f(a)$ and $(\mathrm{D}f(a))^{-1}$ are linear and continuous),[2] *then there exist some open subset $U \subseteq A$ containing a, and some open subset V of F containing $f(a)$, such that f is a diffeomorphism from U to $V = f(U)$. Furthermore,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

For every neighborhood N of a, its image $f(N)$ is a neighborhood of $f(a)$, and for every open ball $U \subseteq A$ of center a, its image $f(U)$ contains some open ball of center $f(a)$.

(2) If $\mathrm{D}f(a)$ is invertible for every $a \in A$, then $B = f(A)$ is an open subset of F, and f is a local diffeomorphism from A to B. Furthermore, if f is injective, then f is a diffeomorphism from A to B.

Proofs of the inverse function theorem can be found in Schwartz [151], Lang [111], Abraham and Marsden [1], and Cartan [34].

The idea of Schwartz's proof is that if we define the function $f_1\colon F \times \Omega \to F$ by

$$f_1(y, z) = f(z) - y,$$

then an inverse $g = f^{-1}$ of f is an implicit solution of the equation $f_1(y, z) = 0$, since $f_1(y, g(y)) = f(g(y)) - y = 0$. Observe that the roles of E and F are switched, but this is not a problem since F is complete. The proof consists in checking that the conditions of Theorem 39.14 apply.

Part (1) of Theorem 39.15 is often referred to as the "(local) inverse function theorem." It plays an important role in the study of manifolds and (ordinary) differential equations.

If E and F are both of finite dimension, and some frames have been chosen, the invertibility of $\mathrm{D}f(a)$ is equivalent to the fact that the Jacobian determinant $\det(J(f)(a))$ is nonnull. The case where $\mathrm{D}f(a)$ is just injective or just surjective is also important for defining manifolds, using implicit definitions.

Definition 39.9. Let E and F be normed affine spaces, where E and F are of finite dimension (or both E and F are complete), and let A be an open subset of E. For any $a \in A$, a C^1-function $f\colon A \to F$ is an *immersion at a* if $\mathrm{D}f(a)$ is injective. A C^1-function $f\colon A \to F$ is a *submersion at a* if $\mathrm{D}f(a)$ is surjective. A C^1-function $f\colon A \to F$ is an *immersion on A* (resp. *a submersion on A*) if $\mathrm{D}f(a)$ is injective (resp. surjective) for every $a \in A$.

[2]Actually, since E and F are Banach spaces, by the Open Mapping Theorem, it is sufficient to assume that $\mathrm{D}f(a)$ is continuous and bijective; see Lang [111].

When E and F are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then f is an immersion iff the Jacobian matrix, $J(f)(a)$, has full rank n for all $a \in E$ and if $n \geq m$, then f is a submersion iff the Jacobian matrix, $J(f)(a)$, has full rank m for all $a \in E$.

Example 39.8. For example, $f \colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (\cos(t), \sin(t))$ is an immersion since $J(f)(t) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}$ has rank 1 for all t. On the other hand, $f \colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (t^2, t^2)$ is not an immersion since $J(f)(t) = \begin{pmatrix} 2t \\ 2t \end{pmatrix}$ vanishes at $t = 0$. See Figure 39.6. An example of a submersion is given by the projection map $f \colon \mathbb{R}^2 \to \mathbb{R}$, where $f(x, y) = x$, since $J(f)(x, y) = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

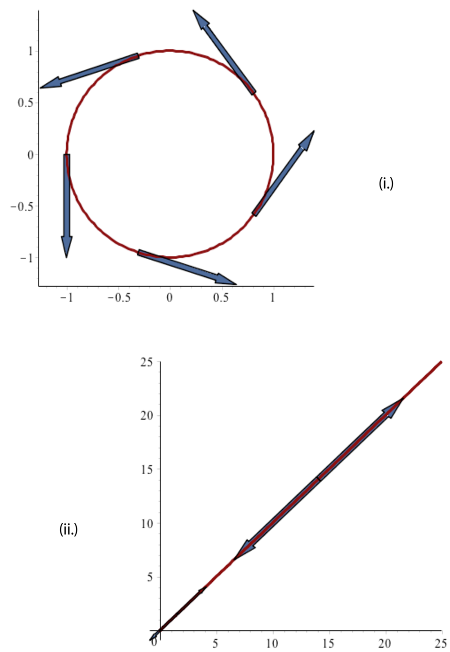


Figure 39.6: Figure $(i.)$ is the immersion of $\mathbb{R}$ into $\mathbb{R}^2$ given by $f(t) = (\cos(t), \sin(t))$. Figure $(ii.)$, the parametric curve $f(t) = (t^2, t^2)$, is not an immersion since the tangent vanishes at the origin.

The following results can be shown.

Proposition 39.16. *Let A be an open subset of $\mathbb{R}^n$, and let $f \colon A \to \mathbb{R}^m$ be a function. For every $a \in A$, $f \colon A \to \mathbb{R}^m$ is a submersion at a iff there exists an open subset U of A containing a, an open subset $W \subseteq \mathbb{R}^{n-m}$, and a diffeomorphism $\varphi \colon U \to f(U) \times W$, such that,*

$$f = \pi_1 \circ \varphi,$$

where $\pi_1 \colon f(U) \times W \to f(U)$ is the first projection. Equivalently,

$$(f \circ \varphi^{-1})(y_1, \ldots, y_m, \ldots, y_n) = (y_1, \ldots, y_m).$$

$$\begin{tikzcd} U \subseteq A \arrow[r, "\varphi"] \arrow[dr, "f"'] & f(U) \times W \arrow[d, "\pi_1"] \\ & f(U) \subseteq \mathbb{R}^m \end{tikzcd}$$

Futhermore, the image of every open subset of A under f is an open subset of F. (The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$). See Figure 39.7.

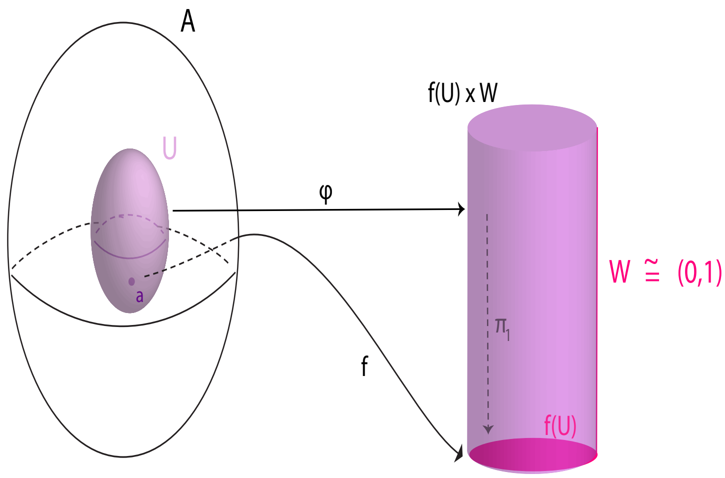


Figure 39.7: Let $n = 3$ and $m = 2$. The submersion maps the solid lavender egg in $\mathbb{R}^3$ onto the bottom pink circular face of the solid cylinder $f(U) \times W$.

Proposition 39.17. *Let A be an open subset of $\mathbb{R}^n$, and let $f \colon A \to \mathbb{R}^m$ be a function. For every $a \in A$, $f \colon A \to \mathbb{R}^m$ is an immersion at a iff there exists an open subset U of A containing a, an open subset V containing $f(a)$ such that $f(U) \subseteq V$, an open subset W containing 0 such that $W \subseteq \mathbb{R}^{m-n}$, and a diffeomorphism $\varphi \colon V \to U \times W$, such that,*

$$\varphi \circ f = in_1,$$

where $in_1 \colon U \to U \times W$ is the injection map such that $in_1(u) = (u, 0)$, or equivalently,

$$(\varphi \circ f)(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$

$$\begin{tikzcd} U \subseteq A \arrow[r, "f"] \arrow[dr, "in_1"'] & f(U) \subseteq V \arrow[d, "\varphi"] \\ & U \times W \end{tikzcd}$$

(The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$). See Figure 39.8.

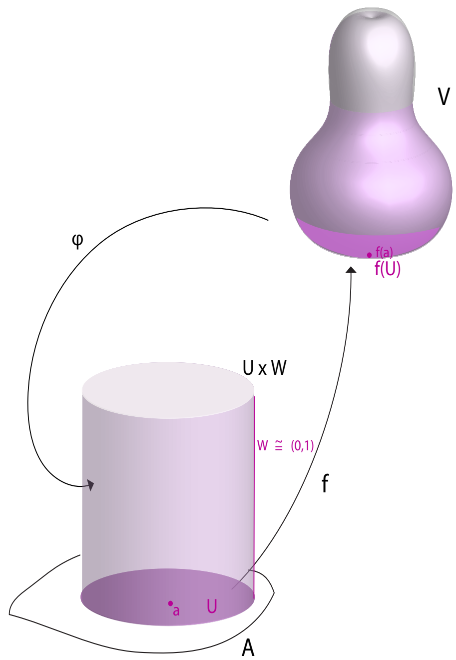


Figure 39.8: Let $n = 2$ and $m = 3$. The immersion maps the purple circular base of the cylinder $U \times W$ to circular cup on the surface of the solid purple gourd.

39.5 Tangent Spaces and Differentials

In this section, we discuss briefly a geometric interpretation of the notion of derivative. We consider sets of points defined by a differentiable function. This is a special case of the notion of a (differential) manifold.

Given two normed affine spaces E and F, let A be an open subset of E, and let $f\colon A \to F$ be a function.

Definition 39.10. Given $f\colon A \to F$ as above, its *graph* $\Gamma(f)$ is the set of all points

$$\Gamma(f) = \{(x, y) \in E \times F \mid x \in A,\, y = f(x)\}.$$

If $\mathrm{D}f$ is defined on A, we say that $\Gamma(f)$ is a *differential submanifold of* $E \times F$ *of equation* $y = f(x)$.

It should be noted that this is a very particular kind of differential manifold.

Example 39.9. If $E = \mathbb{R}$ and $F = \mathbb{R}^2$, letting $f = (g, h)$, where $g\colon \mathbb{R} \to \mathbb{R}$ and $h\colon \mathbb{R} \to \mathbb{R}$, $\Gamma(f)$ is a curve in $\mathbb{R}^3$, of equations $y = g(x)$, $z = h(x)$. When $E = \mathbb{R}^2$ and $F = \mathbb{R}$, $\Gamma(f)$ is a surface in $\mathbb{R}^3$, of equation $z = f(x, y)$.

We now define the notion of affine tangent space in a very general way. Next, we will see what it means for manifolds $\Gamma(f)$, as in Definition 39.10.

Definition 39.11. Given a normed affine space E, given any nonempty subset M of E, given any point $a \in M$, we say that a vector $u \in \overrightarrow{E}$ is *tangent at a to M* if there exist a sequence $(a_n)_{n\in\mathbb{N}}$ of points in M converging to a, and a sequence $(\lambda_n)_{n\in\mathbb{N}}$, with $\lambda_i \in \mathbb{R}$ and $\lambda_n \geq 0$, such that the sequence $(\lambda_n(a_n - a))_{n\in\mathbb{N}}$ converges to u.

The set of all vectors tangent at a to M is called the *family of tangent vectors at a to M* and the set of all points of E of the form $a + u$ where u belongs to the family of tangent vectors at a to M is called the *affine tangent family at a to M*.

Clearly, 0 is always tangent, and if u is tangent, then so is every λu, for $\lambda \in \mathbb{R}$, $\lambda \geq 0$. If $u \neq 0$, then the sequence $(\lambda_n)_{n\in\mathbb{N}}$ must tend towards $+\infty$. We have the following proposition.

Proposition 39.18. *Let E and F be two normed affine spaces, let A be an open subset of E, let $a \in A$, and let $f\colon A \to F$ be a function. If $\mathrm{D}f(a)$ exists, then the family of tangent vectors at $(a, f(a))$ to Γ is a subspace $T_a(\Gamma)$ of $\overrightarrow{E} \times \overrightarrow{F}$, defined by the condition (equation)*

$$(u, v) \in T_a(\Gamma) \quad \textit{iff} \quad v = \mathrm{D}f(a)(u),$$

and the affine tangent family at $(a, f(a))$ to Γ is an affine variety $T_a(\Gamma)$ of $E \times F$, defined by the condition (equation)

$$(x, y) \in T_a(\Gamma) \quad \textit{iff} \quad y = f(a) + \mathrm{D}f(a)(x - a),$$

where Γ is the graph of f.

The proof is actually rather simple. We have $T_a(\Gamma) = a + T_a(\Gamma)$, and since $T_a(\Gamma)$ is a subspace of $\overrightarrow{E} \times \overrightarrow{F}$, the set $T_a(\Gamma)$ is an affine variety. Thus, the affine tangent space at a point $(a, f(a))$ is a familar object, a line, a plane, etc.

As an illustration, when $E = \mathbb{R}^2$ and $F = \mathbb{R}$, the affine tangent plane at the point (a, b, c) to the surface of equation $z = f(x, y)$, is defined by the equation

$$z = c + \frac{\partial f}{\partial x}(a, b)(x - a) + \frac{\partial f}{\partial y}(a, b)(y - b).$$

If $E = \mathbb{R}$ and $F = \mathbb{R}^2$, the tangent line at (a, b, c), to the curve of equations $y = g(x)$, $z = h(x)$, is defined by the equations

$$\begin{aligned} y &= b + \mathrm{D}g(a)(x - a), \\ z &= c + \mathrm{D}h(a)(x - a). \end{aligned}$$

Thus, derivatives and partial derivatives have the desired intended geometric interpretation as tangent spaces. Of course, in order to deal with this topic properly, we really would have to go deeper into the study of (differential) manifolds.

We now briefly consider second-order and higher-order derivatives.

39.6 Second-Order and Higher-Order Derivatives

Given two normed affine spaces E and F, and some open subset A of E, if $\mathrm{D}f(a)$ is defined for every $a \in A$, then we have a mapping $\mathrm{D}f\colon A \to \mathcal{L}(\overrightarrow{E};\overrightarrow{F})$. Since $\mathcal{L}(\overrightarrow{E};\overrightarrow{F})$ is a normed vector space, if $\mathrm{D}f$ exists on an open subset U of A containing a, we can consider taking the derivative of $\mathrm{D}f$ at some $a \in A$.

Definition 39.12. Given a function $f\colon A \to F$ defined on some open subset A of E such that $\mathrm{D}f(a)$ is defined for every $a \in A$, If $\mathrm{D}(\mathrm{D}f)(a)$ exists for every $a \in A$, we get a mapping $\mathrm{D}^2 f\colon A \to \mathcal{L}(\overrightarrow{E};\mathcal{L}(\overrightarrow{E};\overrightarrow{F}))$ called the *second derivative of f on A*, where $\mathrm{D}^2 f(a) = \mathrm{D}(\mathrm{D}f)(a)$, for every $a \in A$.

As in the case of the first derivative $\mathrm{D}f_a$ where $\mathrm{D}f_a(u) = \mathrm{D}_u f(a)$, where $\mathrm{D}_u f(a)$ is the directional derivative of f at a in the direction u, it would be useful to express $\mathrm{D}^2 f(a)(u)(v)$ in terms of two directional derivatives. This can indeed be done. If $\mathrm{D}^2 f(a)$ exists, then for every $u \in \overrightarrow{E}$,

$$\mathrm{D}^2 f(a)(u) = \mathrm{D}(\mathrm{D}f)(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a) \in \mathcal{L}(\overrightarrow{E};\overrightarrow{F}).$$

We have the following result.

Proposition 39.19. *If $\mathrm{D}^2 f(a)$ exists, then $\mathrm{D}_u(\mathrm{D}_v f)(a)$ exists and*

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a), \quad \textit{for all } u, v \in \overrightarrow{E}.$$

Proof. Recall from Proposition 37.61, that the map app from $\mathcal{L}(\overrightarrow{E};\overrightarrow{F}) \times \overrightarrow{E}$ to $\overrightarrow{F}$, defined such that for every $L \in \mathcal{L}(\overrightarrow{E};\overrightarrow{F})$, for every $v \in \overrightarrow{E}$,

$$\mathrm{app}(L, v) = L(v),$$

is a continuous bilinear map. Thus, in particular, given a fixed $v \in \overrightarrow{E}$, the linear map $\mathrm{app}_v\colon \mathcal{L}(\overrightarrow{E};\overrightarrow{F}) \to \overrightarrow{F}$, defined such that $\mathrm{app}_v(L) = L(v)$, is a continuous map.

Also recall from Proposition 39.7, that if $h\colon A \to G$ is a function such that $\mathrm{D}h(a)$ exits, and $k\colon G \to H$ is a continuous linear map, then, $\mathrm{D}(k \circ h)(a)$ exists, and

$$k(\mathrm{D}h(a)(u)) = \mathrm{D}(k \circ h)(a)(u),$$

that is,

$$k(\mathrm{D}_u h(a)) = \mathrm{D}_u(k \circ h)(a),$$

Applying these two facts to $h = \mathrm{D}f$, and to $k = \mathrm{app}_v$, we have

$$\mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{app}_v \circ \mathrm{D}f)(a).$$

But $(\mathrm{app}_v \circ \mathrm{D}f)(x) = \mathrm{D}f(x)(v) = \mathrm{D}_v f(x)$, for every $x \in A$, that is, $\mathrm{app}_v \circ \mathrm{D}f = \mathrm{D}_v f$ on A. So, we have

$$\mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a),$$

and since $\mathrm{D}^2 f(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a)$, we get

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a).$$

Thus, when $\mathrm{D}^2 f(a)$ exists, $\mathrm{D}_u(\mathrm{D}_v f)(a)$ exists, and

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a),$$

for all $u, v \in \overrightarrow{E}$. ☐

Definition 39.13. We denote $\mathrm{D}_u(\mathrm{D}_v f)(a)$ by $\mathrm{D}^2_{u,v} f(a)$ (or $\mathrm{D}_u \mathrm{D}_v f(a)$).

Recall from Proposition 37.60, that the map from $\mathcal{L}_2(\overrightarrow{E}, \overrightarrow{E}; \overrightarrow{F})$ to $\mathcal{L}(\overrightarrow{E}; \mathcal{L}(\overrightarrow{E}; \overrightarrow{F}))$ defined such that $g \mapsto \varphi$ iff for every $g \in \mathcal{L}_2(\overrightarrow{E}, \overrightarrow{E}; \overrightarrow{F})$,

$$\varphi(u)(v) = g(u, v),$$

is an isomorphism of vector spaces. Thus, we will consider $\mathrm{D}^2 f(a) \in \mathcal{L}(\overrightarrow{E}; \mathcal{L}(\overrightarrow{E}; \overrightarrow{F}))$ as a continuous bilinear map in $\mathcal{L}_2(\overrightarrow{E}, \overrightarrow{E}; \overrightarrow{F})$, and we will write $\mathrm{D}^2 f(a)(u, v)$, instead of $\mathrm{D}^2 f(a)(u)(v)$.

Then, the above discussion can be summarized by saying that when $\mathrm{D}^2 f(a)$ is defined, we have

$$\mathrm{D}^2 f(a)(u, v) = \mathrm{D}_u \mathrm{D}_v f(a).$$

Definition 39.14. When E has finite dimension and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, we denote $\mathrm{D}_{e_j} \mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(a)$, when $i \neq j$, and we denote $\mathrm{D}_{e_i} \mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i^2}(a)$.

The following important lemma attributed to Schwarz can be shown, using Proposition 39.12. Given a bilinear map $f \colon \overrightarrow{E} \times \overrightarrow{E} \to \overrightarrow{F}$, recall that f is *symmetric*, if

$$f(u, v) = f(v, u),$$

for all $u, v \in \overrightarrow{E}$.

Proposition 39.20. *(Schwarz's lemma) Given two normed affine spaces E and F, given any open subset A of E, given any $f \colon A \to F$, for every $a \in A$, if $\mathrm{D}^2 f(a)$ exists, then $\mathrm{D}^2 f(a) \in \mathcal{L}_2(\overrightarrow{E}, \overrightarrow{E}; \overrightarrow{F})$ is a continuous symmetric bilinear map. As a corollary, if E is of finite dimension n, and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, we have*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a).$$

Remark: There is a variation of the above result which does not assume the existence of $\mathrm{D}^2 f(a)$, but instead assumes that $\mathrm{D}_u \mathrm{D}_v f$ and $\mathrm{D}_v \mathrm{D}_u f$ exist on an open subset containing a and are continuous at a, and concludes that $\mathrm{D}_u \mathrm{D}_v f(a) = \mathrm{D}_v \mathrm{D}_u f(a)$. This is just a different result which does not imply Proposition 39.20, and is not a consequence of Proposition 39.20.

When $E = \mathbb{R}^2$, the only existence of $\dfrac{\partial^2 f}{\partial x \partial y}(a)$ and $\dfrac{\partial^2 f}{\partial y \partial x}(a)$ is not sufficient to insure the existence of $\mathrm{D}^2 f(a)$.

When E if of finite dimension n and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, if $\mathrm{D}^2 f(a)$ exists, for every $u = u_1 e_1 + \cdots + u_n e_n$ and $v = v_1 e_1 + \cdots + v_n e_n$ in $\overrightarrow{E}$, since $\mathrm{D}^2 f(a)$ is a symmetric bilinear form, we have

$$\mathrm{D}^2 f(a)(u, v) = \sum_{i=1, j=1}^{n} u_i v_j \frac{\partial^2 f}{\partial x_i \partial x_j}(a),$$

which can be written in matrix form as:

$$\mathrm{D}^2 f(a)(u, v) = U^\top \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \ldots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \ldots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \ldots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix} V$$

where U is the column matrix representing u, and V is the column matrix representing v, over the frame $(a_0, (e_1, \ldots, e_n))$.

Definition 39.15. The above symmetric matrix is called the *Hessian of f at a*.

Example 39.10. Consider the function f defined on real invertible 2×2 matrices such that $ad - bc > 0$ given by

$$f(a, b, c, d) = \log(ad - bc).$$

We immediately verify that the Jacobian matrix of f is given by

$$df_{a,b,c,d} = \frac{1}{ad - bc} \begin{pmatrix} d & -c & -b & a \end{pmatrix}.$$

It is easily checked that if

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix},$$

then

$$df_A(X) = \operatorname{tr}(A^{-1} X) = \frac{1}{ad - bc} \operatorname{tr}\left(\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix} \right).$$

Computing second-order derivatives, we find that the Hessian matrix of f is given by

$$Hf(A) = \frac{1}{(ad - bc)^2} \begin{pmatrix} -d^2 & cd & bd & -bc \\ cd & -c^2 & -ad & ac \\ bd & -ad & -b^2 & ab \\ -bc & ac & ab & -a^2 \end{pmatrix}.$$

Using the formula for the derivative of the inversion map and the chain rule we can show that

$$D^2 f(A)(X_1, X_2) = -\mathrm{tr}(A^{-1}X_1A^{-1}X_2),$$

and so

$$Hf(A)(X_1, X_2) = -\mathrm{tr}(A^{-1}X_1A^{-1}X_2),$$

a formula which is far from obvious.

The function f can be generalized to matrices $A \in \mathbf{GL}^+(n, \mathbb{R})$, that is, $n\times n$ real invertible matrices of positive determinants, as

$$f(A) = \log\det(A).$$

It can be shown that the formulae

$$\begin{aligned} df_A(X) &= \mathrm{tr}(A^{-1}X) \\ D^2 f(A)(X_1, X_2) &= -\mathrm{tr}(A^{-1}X_1A^{-1}X_2) \end{aligned}$$

also hold.

Example 39.11. If we restrict the function of Example 39.10 to symmetric positive definite matrices we obtain the function g defined by

$$g(a, b, c) = \log(ac - b^2).$$

We immediately verify that the Jacobian matrix of g is given by

$$dg_{a,b,c} = \frac{1}{ac - b^2}\begin{pmatrix} c & -2b & a \end{pmatrix}.$$

Computing second-order derivatives, we find that the Hessian matrix of g is given by

$$Hg(a, b, c) = \frac{1}{(ac - b^2)^2}\begin{pmatrix} -c^2 & 2bc & -b^2 \\ 2bc & -2(b^2 + ac) & 2ab \\ -b^2 & 2ab & -a^2 \end{pmatrix}.$$

Although this is not obvious, it can be shown that if $ac - b^2 > 0$ and $a, c > 0$, then the matrix $-Hg(a, b, c)$ is symmetric positive definite.

If F itself is of finite dimension, and $(b_0, (v_1, \ldots, v_m))$ is a frame for F, then $f = (f_1, \ldots, f_m)$, and each component $D^2 f(a)_i(u, v)$ of $D^2 f(a)(u, v)$ $(1 \leq i \leq m)$, can be written as

$$D^2 f(a)_i(u, v) = U^\top \begin{pmatrix} \dfrac{\partial^2 f_i}{\partial x_1^2}(a) & \dfrac{\partial^2 f_i}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f_i}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f_i}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f_i}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f_i}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f_i}{\partial x_n^2}(a) \end{pmatrix} V$$

Thus, we could describe the vector $\mathrm{D}^2 f(a)(u, v)$ in terms of an $mn \times mn$-matrix consisting of m diagonal blocks, which are the above Hessians, and the row matrix $(U^\top, \ldots, U^\top)$ (m times) and the column matrix consisting of m copies of V.

We now indicate briefly how higher-order derivatives are defined. Let $m \geq 2$. Given a function $f \colon A \to F$ as before, for any $a \in A$, if the derivatives $\mathrm{D}^i f$ exist on A for all i, $1 \leq i \leq m - 1$, by induction, $\mathrm{D}^{m-1} f$ can be considered to be a continuous function $\mathrm{D}^{m-1} f \colon A \to \mathcal{L}_{m-1}(\overrightarrow{E^{m-1}}; \overrightarrow{F})$.

Definition 39.16. Define $\mathrm{D}^m f(a)$, the *m-th derivative of f at a*, as

$$\mathrm{D}^m f(a) = \mathrm{D}(\mathrm{D}^{m-1} f)(a).$$

Then $\mathrm{D}^m f(a)$ *can be identified with a continuous m-multilinear map in* $\mathcal{L}_m(\overrightarrow{E^m}; \overrightarrow{F})$. We can then show (as we did before), that if $\mathrm{D}^m f(a)$ is defined, then

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \mathrm{D}_{u_1} \ldots \mathrm{D}_{u_m} f(a).$$

Definition 39.17. When E if of finite dimension n and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, if $\mathrm{D}^m f(a)$ exists, for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, we denote $\mathrm{D}_{e_{j_m}} \ldots \mathrm{D}_{e_{j_1}} f(a)$ by

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a).$$

Example 39.12. Going back to the function f of Example 39.10 given by $f(A) = \log\det(A)$, using the formula for the derivative of the inversion map, the chain rule and the product rule, we can show that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma \in \mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1} X_1 A^{-1} X_{\sigma(1)+1} A^{-1} X_{\sigma(2)+1} \cdots A^{-1} X_{\sigma(m-1)+1})$$

for any $m \geq 1$, where $A \in \mathbf{GL}^+(n, \mathbb{R})$ and $X_1, \ldots X_m$ are any $n \times n$ real matrices.

Given a m-multilinear map $f \in \mathcal{L}_m(\overrightarrow{E^m}; \overrightarrow{F})$, recall that f is *symmetric* if

$$f(u_{\pi(1)}, \ldots, u_{\pi(m)}) = f(u_1, \ldots, u_m),$$

for all $u_1, \ldots, u_m \in \overrightarrow{E}$, and all permutations π on $\{1, \ldots, m\}$. Then, the following generalization of Schwarz's lemma holds.

Proposition 39.21. *Given two normed affine spaces E and F, given any open subset A of E, given any $f \colon A \to F$, for every $a \in A$, for every $m \geq 1$, if $\mathrm{D}^m f(a)$ exists, then $\mathrm{D}^m f(a) \in \mathcal{L}_m(\overrightarrow{E^m}; \overrightarrow{F})$ is a continuous symmetric m-multilinear map. As a corollary, if E is of finite dimension n, and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, we have*

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a) = \frac{\partial^m f}{\partial x_{\pi(j_1)} \ldots \partial x_{\pi(j_m)}}(a),$$

for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, and for every permutation π on $\{1, \ldots, m\}$.

Because the trace function is invariant under permutation of its arguments ($\mathrm{tr}(XY) = \mathrm{tr}(YX)$), we see that the m-th derivatives in Example 39.12 are indeed symmetric multilinear maps.

If E is of finite dimension n, and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1} e_1 + \cdots + u_{j,n} e_n.$$

The concept of C^1-function is generalized to the concept of C^m-function, and Theorem 39.13 can also be generalized.

Definition 39.18. Given two normed affine spaces E and F, and an open subset A of E, for any $m \geq 1$, we say that a function $f\colon A \to F$ is *of class C^m on A or a C^m-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every k, $1 \leq k \leq m$. We say that $f\colon A \to F$ is *of class C^∞ on A or a C^∞-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every $k \geq 1$. A C^∞-function (on A) is also called a *smooth function* (on A). A *C^m-diffeomorphism $f\colon A \to B$ between A and B* (where A is an open subset of E and B is an open subset of B) is a bijection between A and $B = f(A)$, such that both $f\colon A \to B$ and its inverse $f^{-1}\colon B \to A$ are C^m-functions.

Equivalently, f is a C^m-function on A if f is a C^1-function on A and $\mathrm{D}f$ is a C^{m-1}-function on A.

We have the following theorem giving a necessary and sufficient condition for f to a C^m-function on A. A generalization to the case where $E = (E_1, a_1) \oplus \cdots \oplus (E_n, a_n)$ also holds.

Theorem 39.22. *Given two normed affine spaces E and F, where E is of finite dimension n, and where $(a_0, (u_1, \ldots, u_n))$ is a frame of E, given any open subset A of E, given any function $f\colon A \to F$, for any $m \geq 1$, the derivative $\mathrm{D}^m f$ is a C^m-function on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f$ (or $\dfrac{\partial^k f}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$. As a corollary, if F is of finite dimension p, and $(b_0, (v_1, \ldots, v_p))$ is a frame of F, the derivative $\mathrm{D}^m f$ is defined and continuous on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f_i$ (or $\dfrac{\partial^k f_i}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, for all i, $1 \leq i \leq p$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$.*

Definition 39.19. When $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any $a \in E$, $\mathrm{D}^m f(a)(1, \ldots, 1)$ is a vector in $\overrightarrow{F}$, called the *mth-order vector derivative*. As in the case $m = 1$, we will usually identify the multilinear map $\mathrm{D}^m f(a)$ with the vector $\mathrm{D}^m f(a)(1, \ldots, 1)$.

Some notational conventions can also be introduced to simplify the notation of higher-order derivatives, and we discuss such conventions very briefly.

Recall that when E is of finite dimension n, and $(a_0, (e_1, \ldots, e_n))$ is a frame for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1}e_1 + \cdots + u_{j,n}e_n.$$

We can then group the various occurrences of ∂x_{j_k} corresponding to the same variable x_{j_k}, and this leads to the notation

$$\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} f(a),$$

where $\alpha_1 + \alpha_2 + \cdots + \alpha_n = m$.

If we denote $(\alpha_1, \ldots, \alpha_n)$ simply by α, then we denote

$$\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} f$$

by

$$\partial^\alpha f, \quad \text{or} \quad \left(\frac{\partial}{\partial x}\right)^\alpha f.$$

If $\alpha = (\alpha_1, \ldots, \alpha_n)$, we let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$, and if $h = (h_1, \ldots, h_n)$, we denote $h_1^{\alpha_1} \cdots h_n^{\alpha_n}$ by h^α.

In the next section, we survey various versions of Taylor's formula.

39.7 Taylor's formula, Faà di Bruno's formula

We discuss, without proofs, several versions of Taylor's formula. The hypotheses required in each version become increasingly stronger. The first version can be viewed as a generalization of the notion of derivative. Given an m-linear map $f\colon \overrightarrow{E^m} \to \overrightarrow{F}$, for any vector $h \in \overrightarrow{E}$, we abbreviate

$$f(\underbrace{h, \ldots, h}_{m})$$

by $f(h^m)$. The version of Taylor's formula given next is sometimes referred to as the *formula of Taylor–Young*.

Theorem 39.23. *(Taylor–Young) Given two normed affine spaces E and F, for any open subset $A \subseteq E$, for any function $f \colon A \to F$, for any $a \in A$, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m-1$, and if $\mathrm{D}^m f(a)$ exists, then we have:*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \|h\|^m \epsilon(h),$$

for any h such that $a + h \in A$, and where $\lim_{h\to 0,\, h\neq 0} \epsilon(h) = 0$.

The above version of Taylor's formula has applications to the study of relative maxima (or minima) of real-valued functions. It is also used to study the local properties of curves and surfaces.

The next version of Taylor's formula can be viewed as a generalization of Proposition 39.12. It is sometimes called the *Taylor formula with Lagrange remainder* or *generalized mean value theorem*.

Theorem 39.24. *(Generalized mean value theorem) Let E and F be two normed affine spaces, let A be an open subset of E, and let $f \colon A \to F$ be a function on A. Given any $a \in A$ and any $h \neq 0$ in $\overrightarrow{E}$, if the closed segment $[a, a+h]$ is contained in A, $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, $\mathrm{D}^{m+1} f(x)$ exists at every point x of the open segment $]a, a+h[$, and*

$$\max_{x \in (a, a+h)} \left\| \mathrm{D}^{m+1} f(x) \right\| \leq M,$$

for some $M \geq 0$, then

$$\left\| f(a+h) - f(a) - \Big(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m)\Big) \right\| \leq M \frac{\|h\|^{m+1}}{(m+1)!}.$$

As a corollary, if $L \colon \overrightarrow{E^{m+1}} \to \overrightarrow{F}$ is a continuous $(m+1)$-linear map, then

$$\left\| f(a+h) - f(a) - \Big(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{L(h^{m+1})}{(m+1)!}\Big) \right\| \leq M \frac{\|h\|^{m+1}}{(m+1)!},$$

where $M = \max_{x \in (a, a+h)} \|\mathrm{D}^{m+1} f(x) - L\|$.

The above theorem is sometimes stated under the slightly stronger assumption that f is a C^m-function on A. If $f \colon A \to \mathbb{R}$ is a real-valued function, Theorem 39.24 can be refined a little bit. This version is often called the *formula of Taylor–MacLaurin*.

Theorem 39.25. *(Taylor–MacLaurin) Let E be a normed affine space, let A be an open subset of E, and let $f \colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in $\overrightarrow{E}$, if the closed segment $[a, a+h]$ is contained in A, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, and $\mathrm{D}^{m+1} f(x)$ exists at every point x of the open segment $]a, a+h[$, then there is some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, such that*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{1}{(m+1)!}\mathrm{D}^{m+1} f(a + \theta h)(h^{m+1}).$$

Example 39.13. Going back to the function f of Example 39.10 given by $f(A) = \log\det(A)$, we know from Example 39.12 that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma\in\mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1}X_1A^{-1}X_{\sigma(1)+1}\cdots A^{-1}X_{\sigma(m-1)+1}) \qquad (*)$$

for all $m \geq 1$, with $A \in \mathbf{GL}^+(n, \mathbb{R})$. If we make the stronger assumption that A is symmetric positive definite, then for any other symmetric positive definite matrix B, since the symmetric positive definite matrices form a convex set, the matrices $A + \theta(B - A) = (1-\theta)A + \theta B$ are also symmetric positive definite for $\theta \in [0, 1]$. Theorem 39.25 applies with $H = B - A$ (a symmetric matrix), and using $(*)$, we obtain

$$\begin{aligned}\log\det(A + H) = \log\det(A) + \mathrm{tr}\bigg(&A^{-1}H - \frac{1}{2}(A^{-1}H)^2 + \cdots + \frac{(-1)^{m-1}}{m}(A^{-1}H)^m \\ &+ \frac{(-1)^m}{m+1}((A+\theta H)^{-1}H)^{m+1}\bigg),\end{aligned}$$

for some θ such that $0 < \theta < 1$. In particular, if $A = I$, for any symmetric matrix H such that $I + H$ is symmetric positive definite, we obtain

$$\begin{aligned}\log\det(I + H) = \mathrm{tr}\bigg(&H - \frac{1}{2}H^2 + \cdots + \frac{(-1)^{m-1}}{m}H^m \\ &+ \frac{(-1)^m}{m+1}((I+\theta H)^{-1}H)^{m+1}\bigg),\end{aligned}$$

for some θ such that $0 < \theta < 1$. In the special case when $n = 1$, we have $I = 1$, H is a real such that $1 + H > 0$ and the trace function is the identity, so we recognize the partial sum of the series for $x \mapsto \log(1 + x)$,

$$\begin{aligned}\log(1 + H) = H &- \frac{1}{2}H^2 + \cdots + \frac{(-1)^{m-1}}{m}H^m \\ &+ \frac{(-1)^m}{m+1}(1+\theta H)^{-(m+1)}H^{m+1}.\end{aligned}$$

We also mention for "mathematical culture," a version with integral remainder, in the case of a real-valued function. This is usually called *Taylor's formula with integral remainder*.

Theorem 39.26. *(Taylor's formula with integral remainder) Let E be a normed affine space, let A be an open subset of E, and let $f\colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in $\overrightarrow{E}$, if the closed segment $[a, a + h]$ is contained in A, and if f is a C^{m+1}-function on A, then we have*

$$\begin{aligned}f(a + h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m)& \\ + \int_0^1 \frac{(1-t)^m}{m!}\Big[\mathrm{D}^{m+1} f(a + th)(h^{m+1})\Big]dt.&\end{aligned}$$

The advantage of the above formula is that it gives an explicit remainder. We now examine briefly the situation where E is of finite dimension n, and $(a_0, (e_1, \ldots, e_n))$ is a frame for E. In this case, we get a more explicit expression for the expression

$$\sum_{i=0}^{k=m} \frac{1}{k!} \mathrm{D}^k f(a)(h^k)$$

involved in all versions of Taylor's formula, where by convention, $\mathrm{D}^0 f(a)(h^0) = f(a)$. If $h = h_1 e_1 + \cdots + h_n e_n$, then we have

$$\sum_{k=0}^{k=m} \frac{1}{k!} \mathrm{D}^k f(a)(h^k) = \sum_{k_1 + \cdots + k_n \leq m} \frac{h_1^{k_1} \cdots h_n^{k_n}}{k_1! \cdots k_n!} \left(\frac{\partial}{\partial x_1}\right)^{k_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{k_n} f(a),$$

which, using the abbreviated notation introduced at the end of Section 39.6, can also be written as

$$\sum_{k=0}^{k=m} \frac{1}{k!} \mathrm{D}^k f(a)(h^k) = \sum_{|\alpha| \leq m} \frac{h^\alpha}{\alpha!} \partial^\alpha f(a).$$

The advantange of the above notation is that it is the same as the notation used when $n = 1$, i.e., when $E = \mathbb{R}$ (or $E = \mathbb{C}$). Indeed, in this case, the Taylor–MacLaurin formula reads as:

$$f(a + h) = f(a) + \frac{h}{1!} \mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!} \mathrm{D}^m f(a) + \frac{h^{m+1}}{(m+1)!} \mathrm{D}^{m+1} f(a + \theta h),$$

for some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, where $\mathrm{D}^k f(a)$ is the value of the k-th derivative of f at a (and thus, as we have already said several times, this is the kth-order vector derivative, which is just a scalar, since $F = \mathbb{R}$).

In the above formula, the assumptions are that $f \colon [a, a + h] \to \mathbb{R}$ is a C^m-function on $[a, a + h]$, and that $\mathrm{D}^{m+1} f(x)$ exists for every $x \in (a, a + h)$.

Taylor's formula is useful to study the local properties of curves and surfaces. In the case of a curve, we consider a function $f \colon [r, s] \to F$ from a closed interval $[r, s]$ of $\mathbb{R}$ to some affine space F, the derivatives $\mathrm{D}^k f(a)(h^k)$ correspond to vectors $h^k \mathrm{D}^k f(a)$, where $\mathrm{D}^k f(a)$ is the kth vector derivative of f at a (which is really $\mathrm{D}^k f(a)(1, \ldots, 1)$), and for any $a \in (r, s)$, Theorem 39.23 yields the following formula:

$$f(a + h) = f(a) + \frac{h}{1!} \mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!} \mathrm{D}^m f(a) + h^m \epsilon(h),$$

for any h such that $a + h \in (r, s)$, and where $\lim_{h \to 0,\, h \neq 0} \epsilon(h) = 0$.

In the case of functions $f \colon \mathbb{R}^n \to \mathbb{R}$, it is convenient to have formulae for the Taylor–Young formula and the Taylor–MacLaurin formula in terms of the gradient and the Hessian.

Recall that the *gradient* $\nabla f(a)$ of f at $a \in \mathbb{R}^n$ is the column vector

$$\nabla f(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) \\ \dfrac{\partial f}{\partial x_2}(a) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix},$$

and that

$$f'(a)(u) = \mathrm{D}f(a)(u) = \nabla f(a) \cdot u,$$

for any $u \in \mathbb{R}^n$ (where $\cdot$ means inner product). The above equation shows that *the direction of the gradient* $\nabla f(a)$ *is the direction of maximal increase of the function* f *at* a and that $\|\nabla f(a)\|$ *is the rate of change of* f *in its direction of maximal increase*. This is the reason why methods of "gradient descent" pick the direction *opposite* to the gradient (we are trying to minimize f).

The *Hessian matrix* $\nabla^2 f(a)$ of f at $a \in \mathbb{R}^n$ is the $n \times n$ symmetric matrix

$$\nabla^2 f(a) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix},$$

and we have

$$\mathrm{D}^2 f(a)(u, v) = u^\top \nabla^2 f(a)\, v = u \cdot \nabla^2 f(a) v = \nabla^2 f(a) u \cdot v,$$

for all $u, v \in \mathbb{R}^n$. Then, we have the following three formulations of the formula of Taylor–Young of order 2:

$$\begin{aligned} f(a + h) &= f(a) + \mathrm{D}f(a)(h) + \frac{1}{2}\mathrm{D}^2 f(a)(h, h) + \|h\|^2\, \epsilon(h) \\ f(a + h) &= f(a) + \nabla f(a) \cdot h + \frac{1}{2}(h \cdot \nabla^2 f(a) h) + (h \cdot h)\epsilon(h) \\ f(a + h) &= f(a) + (\nabla f(a))^\top h + \frac{1}{2}(h^\top \nabla^2 f(a)\, h) + (h^\top h)\epsilon(h). \end{aligned}$$

with $\lim_{h \mapsto 0} \epsilon(h) = 0$.

One should keep in mind that only the first formula is intrinsic (i.e., does not depend on the choice of a basis), whereas the other two depend on the basis and the inner product chosen on $\mathbb{R}^n$. As an exercise, the reader should write similar formulae for the Taylor–MacLaurin formula of order 2.

Another application of Taylor's formula is the derivation of a formula which gives the m-th derivative of the composition of two functions, usually known as "Faà di Bruno's formula." This formula is useful when dealing with geometric continuity of splines curves and surfaces.

Proposition 39.27. *Given any normed affine space E, for any function $f\colon \mathbb{R} \to \mathbb{R}$ and any function $g\colon \mathbb{R} \to E$, for any $a \in \mathbb{R}$, letting $b = f(a)$, $f^{(i)}(a) = \mathrm{D}^i f(a)$, and $g^{(i)}(b) = \mathrm{D}^i g(b)$, for any $m \geq 1$, if $f^{(i)}(a)$ and $g^{(i)}(b)$ exist for all i, $1 \leq i \leq m$, then $(g \circ f)^{(m)}(a) = \mathrm{D}^m(g \circ f)(a)$ exists and is given by the following formula:*

$$(g \circ f)^{(m)}(a) = \sum_{0 \leq j \leq m} \sum_{\substack{i_1+i_2+\cdots+i_m=j \\ i_1+2i_2+\cdots+mi_m=m \\ i_1,i_2,\cdots,i_m \geq 0}} \frac{m!}{i_1! \cdots i_m!}\, g^{(j)}(b) \left(\frac{f^{(1)}(a)}{1!}\right)^{i_1} \cdots \left(\frac{f^{(m)}(a)}{m!}\right)^{i_m}.$$

When $m = 1$, the above simplifies to the familiar formula

$$(g \circ f)'(a) = g'(b)f'(a),$$

and for $m = 2$, we have

$$(g \circ f)^{(2)}(a) = g^{(2)}(b)(f^{(1)}(a))^2 + g^{(1)}(b)f^{(2)}(a).$$

39.8 Vector Fields, Covariant Derivatives, Lie Brackets

In this section, we briefly consider vector fields and covariant derivatives of vector fields. Such derivatives play an important role in continuous mechanics. Given a normed affine space $(E, \overrightarrow{E})$, a *vector field over* $(E, \overrightarrow{E})$ is a function $X\colon E \to \overrightarrow{E}$. Intuitively, a vector field assigns a vector to every point in E. Such vectors could be forces, velocities, accelerations, etc.

Given two vector fields X, Y defined on some open subset Ω of E, for every point $a \in \Omega$, we would like to define the derivative of X with respect to Y at a. This is a type of directional derivative that gives the variation of X as we move along Y, and we denote it by $\mathrm{D}_Y X(a)$. The derivative $\mathrm{D}_Y X(a)$ is defined as follows.

Definition 39.20. Let $(E, \overrightarrow{E})$ be a normed affine space. Given any open subset Ω of E, given any two vector fields X and Y defined over Ω, for any $a \in \Omega$, the *covariant derivative (or Lie derivative) of X w.r.t. the vector field Y at a*, denoted by $\mathrm{D}_Y X(a)$, is the limit (if it exists)

$$\lim_{t \to 0,\, t \in U} \frac{X(a + tY(a)) - X(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tY(a) \in \Omega,\ t \neq 0\}$.

If Y is a constant vector field, it is immediately verified that the map

$$X \mapsto \mathrm{D}_Y X(a)$$

is a linear map called the *derivative* of the vector field X, and denoted by $\mathrm{D}X(a)$. If $f\colon E \to \mathbb{R}$ is a function, we define $\mathrm{D}_Y f(a)$ as the limit (if it exists)

$$\lim_{t \to 0,\, t \in U} \frac{f(a + tY(a)) - f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tY(a) \in \Omega,\, t \neq 0\}$. It is the *directional derivative of f w.r.t. the vector field Y at a*, and it is also often denoted by $Y(f)(a)$, or $Y(f)_a$.

From now on, we assume that all the vector fields and all the functions under consideration are smooth (C^∞). The set $C^\infty(\Omega)$ of smooth C^∞-functions $f\colon \Omega \to \mathbb{R}$ is a ring. Given a smooth vector field X and a smooth function f (both over Ω), the vector field fX is defined such that $(fX)(a) = f(a)X(a)$, and it is immediately verified that it is smooth. Thus, the set $\mathcal{X}(\Omega)$ of smooth vector fields over Ω is a $C^\infty(\Omega)$-module.

The following proposition is left as an exercise. It shows that $\mathrm{D}_Y X(a)$ is a $\mathbb{R}$-bilinear map on $\mathcal{X}(\Omega)$, is $C^\infty(\Omega)$-linear in Y, and satisfies the Leibniz derivation rules with respect to X.

Proposition 39.28. *The covariant derivative $\mathrm{D}_Y X(a)$ satisfies the following properties:*

$$\begin{aligned}
\mathrm{D}_{(Y_1+Y_2)} X(a) &= \mathrm{D}_{Y_1} X(a) + \mathrm{D}_{Y_2} X(a),\\
\mathrm{D}_{fY} X(a) &= f(a)\mathrm{D}_Y X(a),\\
\mathrm{D}_Y (X_1 + X_2)(a) &= \mathrm{D}_Y X_1(a) + \mathrm{D}_Y X_2(a),\\
\mathrm{D}_Y fX(a) &= \mathrm{D}_Y f(a) X(a) + f(a)\mathrm{D}_Y X(a),
\end{aligned}$$

where X, Y, X_1, X_2, Y_1, Y_2 are smooth vector fields over Ω, and $f\colon E \to \mathbb{R}$ is a smooth function.

In differential geometry, the above properties are taken as the axioms of *affine connections*, in order to define covariant derivatives of vector fields over manifolds. In many cases, the vector field Y is the tangent field of some smooth curve $\gamma\colon]-\eta, \eta[\to E$. If so, the following proposition holds.

Proposition 39.29. *Given a smooth curve $\gamma\colon]-\eta, \eta[\to E$, letting Y be the vector field defined on $\gamma(]-\eta, \eta[)$ such that*

$$Y(\gamma(u)) = \frac{d\gamma}{dt}(u),$$

for any vector field X defined on $\gamma(]-\eta, \eta[)$, we have

$$\mathrm{D}_Y X(a) = \frac{d}{dt}\Big[X(\gamma(t))\Big](0),$$

where $a = \gamma(0)$.

The derivative $\mathrm{D}_Y X(a)$ is thus the derivative of the vector field X along the curve γ, and it is called the *covariant derivative of X along γ*.

Given an affine frame $(O, (u_1, \ldots, u_n))$ for $(E, \overrightarrow{E})$, it is easily seen that the covariant derivative $\mathrm{D}_Y X(a)$ is expressed as follows:

$$\mathrm{D}_Y X(a) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(Y_j \frac{\partial X_i}{\partial x_j} \right)(a) e_i.$$

Generally, $\mathrm{D}_Y X(a) \neq \mathrm{D}_X Y(a)$. The quantity

$$[X, Y] = \mathrm{D}_X Y - \mathrm{D}_Y X$$

is called the *Lie bracket* of the vector fields X and Y. The Lie bracket plays an important role in differential geometry. In terms of coordinates,

$$[X, Y] = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(X_j \frac{\partial Y_i}{\partial x_j} - Y_j \frac{\partial X_i}{\partial x_j} \right) e_i.$$

39.9 Futher Readings

A thorough treatment of differential calculus can be found in Munkres [130], Lang [112], Schwartz [151], Cartan [34], and Avez [9]. The techniques of differential calculus have many applications, especially to the geometry of curves and surfaces and to differential geometry in general. For this, we recommend do Carmo [52, 53] (two beautiful classics on the subject), Kreyszig [106], Stoker [166], Gray [81], Berger and Gostiaux [13], Milnor [126], Lang [110], Warner [186] and Choquet-Bruhat [38].

39.10 Summary

The main concepts and results of this chapter are listed below:

- *Directional derivative* ($\mathrm{D}_u f(a)$).
- *Total derivative*, *Fréchet derivative*, *derivative*, *total differential*, *differential* ($df(a), df_a$).
- *Partial derivatives*.
- *Affine* functions.
- The *chain rule*.
- *Jacobian matrices* ($J(f)(a)$), *Jacobians*.

- *Gradient* of a function ($\operatorname{grad} f(a)$, $\nabla f(a)$).
- *Mean value theorem.*
- *C^0-functions, C^1-functions.*
- The *implicit function theorem.*
- *Local homeomorphisms, local diffeomorphisms, diffeomorphisms.*
- The *inverse function theorem.*
- *Immersions, submersions.*
- Second-order and higher-order derivatives.
- *Schwarz's lemma.*
- *Hessian matrix.*
- *C^∞-functions, smooth functions.*
- *Taylor–Young's formula.*
- Generalized mean value theorem.
- *Taylor–MacLaurin's formula.*
- *Taylor's formula with integral remainder.*
- *Faà di Bruno's formula.*

39.11 Problems

Problem 39.1. Let $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ be the function defined on $n \times n$ matrices by

$$f(A) = A^2.$$

Prove that

$$\mathrm{D}f_A(H) = AH + HA,$$

for all $A, H \in \mathrm{M}_n(\mathbb{R})$.

Problem 39.2. Let $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ be the function defined on $n \times n$ matrices by

$$f(A) = A^3.$$

Prove that

$$\mathrm{D}f_A(H) = A^2H + AHA + HA^2,$$

for all $A, H \in \mathrm{M}_n(\mathbb{R})$.

Problem 39.3. If $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ and $g\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ are differentiable matrix functions, prove that

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathrm{M}_n(\mathbb{R})$.

Problem 39.4. Recall that $\mathfrak{so}(3)$ denotes the vector space of real skew-symmetric $n \times n$ matrices ($B^\top = -B$). Let $C\colon \mathfrak{so}(n) \to \mathrm{M}_n(\mathbb{R})$ be the function given by

$$C(B) = (I - B)(I + B)^{-1}.$$

(1) Prove that if B is skew-symmetric, then $I - B$ and $I + B$ are invertible, and so C is well-defined. Prove that

(2) Prove that

$$dC(B)(A) = -[I + (I - B)(I + B)^{-1}]A(I + B)^{-1} = -2(I + B)^{-1}A(I + B)^{-1}.$$

(3) Prove that $dC(B)$ is injective for every skew-symmetric matrix B.

Problem 39.5. Prove that

$$\begin{aligned} &d^m C_B(H_1, \ldots, H_m) \\ &\quad = 2(-1)^m \sum_{\pi \in \mathfrak{S}_m} (I + B)^{-1} H_{\pi(1)} (I + B)^{-1} H_{\pi(2)} (I + B)^{-1} \cdots (I + B)^{-1} H_{\pi(m)} (I + B)^{-1}. \end{aligned}$$

Problem 39.6. Consider the function g defined for all $A \in \mathbf{GL}(n, \mathbb{R})$, that is, all $n \times n$ real invertible matrices, given by

$$g(A) = \det(A).$$

(1) Prove that

$$dg_A(X) = \det(A)\mathrm{tr}(A^{-1}X),$$

for all $n \times n$ real matrices X.

(2) Consider the function f defined for all $A \in \mathbf{GL}^+(n, \mathbb{R})$, that is, $n \times n$ real invertible matrices of positive determinants, given by

$$f(A) = \log g(A) = \log\det(A).$$

Prove that

$$\begin{aligned} df_A(X) &= \mathrm{tr}(A^{-1}X) \\ \mathrm{D}^2 f(A)(X_1, X_2) &= -\mathrm{tr}(A^{-1}X_1A^{-1}X_2), \end{aligned}$$

for all $n \times n$ real matrices X, X_1, X_2.

(3) Prove that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma \in \mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1}X_1A^{-1}X_{\sigma(1)+1}A^{-1}X_{\sigma(2)+1} \cdots A^{-1}X_{\sigma(m-1)+1})$$

for any $m \geq 1$, where $X_1, \ldots X_m$ are any $n \times n$ real matrices.

Part VI

Preliminaries for Optimization Theory

Chapter 40

Extrema of Real-Valued Functions

This chapter deals with extrema of real-valued functions. In most optimization problems, we need to find necessary conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open). This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*. The case of equality constraints is much easier to deal with and is treated in this chapter.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 50.

40.1 Local Extrema, Constrained Local Extrema, and Lagrange Multipliers

Let $J\colon E \to \mathbb{R}$ be a real-valued function defined on a normed vector space E (or more generally, any topological space). Ideally we would like to find where the function J reaches a minimum or a maximum value, at least locally. In this chapter we will usually use the notations $dJ(u)$ or $J'(u)$ (or dJ_u or J'_u) for the derivative of J at u, instead of $\mathrm{D}J(u)$. Our presentation follows very closely that of Ciarlet [41] (Chapter 7), which we find to be one of the clearest.

Definition 40.1. If $J\colon E \to \mathbb{R}$ is a real-valued function defined on a normed vector space E, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in W.$$

In either case, we say that J has a *local extremum* (or *relative extremum*) at u. We say that J has a *strict local minimum* (resp. *strict local maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) < J(w) \quad \text{for all } w \in W - \{u\}$$

(resp.

$$J(u) > J(w) \quad \text{for all } w \in W - \{u\}).$$

By abuse of language, we often say that the point u itself "is a local minimum" or a "local maximum," even though, strictly speaking, this does not make sense.

We begin with a well-known necessary condition for a local extremum.

Proposition 40.1. *Let E be a normed vector space and let $J\colon \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J has a local extremum at some point $u \in \Omega$ and if J is differentiable at u, then*

$$dJ_u = J'(u) = 0.$$

Proof. Pick any $v \in E$. Since Ω is open, for t small enough we have $u + tv \in \Omega$, so there is an open interval $I \subseteq \mathbb{R}$ such that the function φ given by

$$\varphi(t) = J(u + tv)$$

for all $t \in I$ is well-defined. By applying the chain rule, we see that φ is differentiable at $t = 0$, and we get

$$\varphi'(0) = dJ_u(v).$$

Without loss of generality, assume that u is a local minimum. Then we have

$$\varphi'(0) = \lim_{t \mapsto 0_-} \frac{\varphi(t) - \varphi(0)}{t} \leq 0$$

and

$$\varphi'(0) = \lim_{t \mapsto 0_+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0,$$

which shows that $\varphi'(0) = dJ_u(v) = 0$. As $v \in E$ is arbitrary, we conclude that $dJ_u = 0$. $\square$

Definition 40.2. A point $u \in \Omega$ such that $J'(u) = 0$ is called a *critical point* of J.

If $E = \mathbb{R}^n$, then the condition $dJ_u = 0$ is equivalent to the system

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u_1, \ldots, u_n) &= 0 \\
&\vdots \\
\frac{\partial J}{\partial x_n}(u_1, \ldots, u_n) &= 0.
\end{aligned}$$

The condition of Proposition 40.1 is only a *necessary* condition for the existence of an extremum, but not a sufficient condition.

Here are some counter-examples. If $f : \mathbb{R} \to \mathbb{R}$ is the function given by $f(x) = x^3$, since $f'(x) = 3x^2$, we have $f'(0) = 0$, but 0 is neither a minimum nor a maximum of f as evidenced by the graph shown in Figure 40.1.

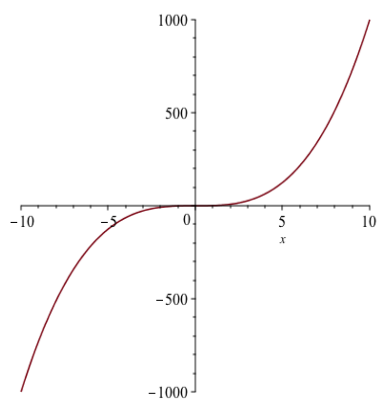


Figure 40.1: The graph of $f(x) = x^3$. Note that $x = 0$ is a saddle point and not a local extremum.

If $g\colon \mathbb{R}^2 \to \mathbb{R}$ is the function given by $g(x, y) = x^2 - y^2$, then $g'_{(x,y)} = (2x \quad -2y)$, so $g'_{(0,0)} = (0 \;\; 0)$, yet near $(0, 0)$ the function g takes negative and positive values. See Figure 40.2.

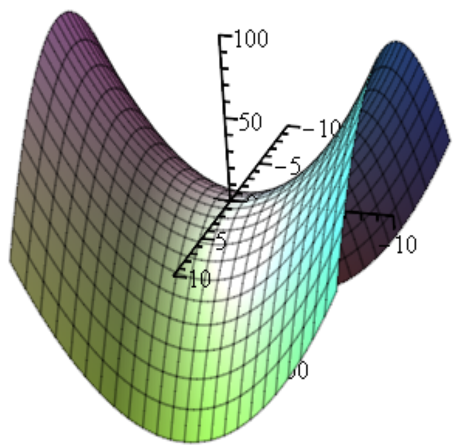


Figure 40.2: The graph of $g(x, y) = x^2 - y^2$. Note that $(0, 0)$ is a saddle point and not a local extremum.

It is very important to note that the hypothesis that Ω *is open* is crucial for the validity of Proposition 40.1.

For example, if J is the identity function on $\mathbb{R}$ and $U = [0, 1]$, a closed subset, then $J'(x) = 1$ for all $x \in [0, 1]$, even though J has a minimum at $x = 0$ and a maximum at $x = 1$.

In many practical situations, we need to look for local extrema of a function J *under additional constraints*. This situation can be formalized conveniently as follows. We have a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space, but we also have some subset U of Ω, and we are looking for the local extrema of J *with respect to the set U*.

The elements $u \in U$ are often called *feasible solutions* of the optimization problem consisting in finding the local extrema of some objective function J with respect to some subset U of Ω defined by a set of constraints. Note that in most cases, U is *not* open. In fact, U is usually closed.

Definition 40.3. If $J\colon \Omega \to \mathbb{R}$ is a real-valued function defined on some open subset Ω of a normed vector space E and if U is some subset of Ω, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in U$ *with respect to U* if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \le J(w) \quad \text{for all } w \in U \cap W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in U \cap W.$$

In either case, we say that J has a *local extremum* at u *with respect to* U.

In order to find necessary conditions for a function $J \colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open), we need to somehow incorporate the definition of U into these conditions. This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*.

An inequality constraint of the form $\varphi_i(x) \geq 0$ is equivalent to the inequality constraint $-\varphi_x(x) \leq 0$. An equality constraint $\varphi_i(x) = 0$ is equivalent to the conjunction of the two inequality constraints $\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$, so the case of inequality constraints subsumes the case of equality constraints. However, the case of equality constraints is easier to deal with, and in this chapter we will restrict our attention to this case.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we will discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 50.

We begin by considering the case where $\Omega \subseteq E_1 \times E_2$ is an open subset of a product of normed vector spaces and where U is the zero locus of some continuous function $\varphi \colon \Omega \to E_2$, which means that

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

For the sake of brevity, we say that J has a *constrained local extremum* at u instead of saying that J has a *local extremum* at the point $u \in U$ *with respect to* U.

In most applications, we have $E_1 = \mathbb{R}^{n-m}$ and $E_2 = \mathbb{R}^m$ for some integers m, n such that $1 \leq m < n$, Ω is an open subset of $\mathbb{R}^n$, $J \colon \Omega \to \mathbb{R}$, and we have m functions $\varphi_i \colon \Omega \to \mathbb{R}$ defining the subset

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\}.$$

Fortunately, there is a necessary condition for constrained local extrema in terms of *Lagrange multipliers*.

Theorem 40.2. *(Necessary condition for a constrained extremum in terms of Lagrange multipliers) Let Ω be an open subset of $\mathbb{R}^n$, consider m C^1-functions $\varphi_i \colon \Omega \to \mathbb{R}$ (with $1 \leq m < n$), let*

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\},$$

and let $u \in U$ be a point such that the derivatives $d\varphi_i(u) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R})$ are linearly independent; equivalently, assume that the $m \times n$ matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. If $J \colon \Omega \to \mathbb{R}$ is a function which is differentiable at $u \in U$ and if J has a local constrained extremum at u, then there exist m numbers $\lambda_i(u) \in \mathbb{R}$, uniquely defined, such that

$$dJ(u) + \lambda_1(u) d\varphi_1(u) + \cdots + \lambda_m(u) d\varphi_m(u) = 0;$$

equivalently,

$$\nabla J(u) + \lambda_1(u) \nabla\varphi_1(u) + \cdots + \lambda_m(u) \nabla\varphi_m(u) = 0.$$

Theorem 40.2 will be proven as a corollary of Theorem 40.4, which gives a more general formulation that applies to the situation where E_1 is an infinite-dimensional Banach space. To simplify the exposition we postpone a discussion of this theorem until we have presented several examples illustrating the method of Lagrange multipliers.

Definition 40.4. The numbers $\lambda_i(u)$ involved in Theorem 40.2 are called the *Lagrange multipliers* associated with the constrained extremum u (again, with some minor abuse of language).

The linear independence of the linear forms $d\varphi_i(u)$ is equivalent to the fact that the Jacobian matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ of $\varphi = (\varphi_1, \ldots, \varphi_m)$ at u has rank m. If $m = 1$, the linear independence of the $d\varphi_i(u)$ reduces to the condition $\nabla\varphi_1(u) \neq 0$.

A fruitful way to reformulate the use of Lagrange multipliers is to introduce the notion of the Lagrangian associated with our constrained extremum problem.

Definition 40.5. The *Lagrangian* associated with our constrained extremum problem is the function $L \colon \Omega \times \mathbb{R}^m \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \lambda_1 \varphi_1(v) + \cdots + \lambda_m \varphi_m(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$.

We have the following simple but important proposition.

Proposition 40.3. *There exists some $\mu = (\mu_1, \ldots, \mu_m)$ and some $u \in U$ such that*

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

if and only if

$$dL(u, \mu) = 0,$$

or equivalently

$$\nabla L(u, \mu) = 0;$$

that is, iff (u, μ) is a critical point of the Lagrangian L.

Proof. Indeed $dL(u, \mu) = 0$ is equivalent to

$$\begin{aligned}
\frac{\partial L}{\partial v}(u, \mu) &= 0 \\
\frac{\partial L}{\partial \lambda_1}(u, \mu) &= 0 \\
&\vdots \\
\frac{\partial L}{\partial \lambda_m}(u, \mu) &= 0,
\end{aligned}$$

and since

$$\frac{\partial L}{\partial v}(u, \mu) = dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u)$$

and

$$\frac{\partial L}{\partial \lambda_i}(u, \mu) = \varphi_i(u),$$

we get

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

and

$$\varphi_1(u) = \cdots = \varphi_m(u) = 0,$$

that is, $u \in U$. The converse is proven in a similar fashion (essentially by reversing the argument). $\square$

If we write out explicitly the condition

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0,$$

we get the $n \times m$ system

$$\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) = 0$$
$$\vdots$$
$$\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_n}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) = 0,$$

and it is important to note that the matrix of this system is the *transpose* of the Jacobian matrix of φ at u. If we write $\mathrm{Jac}(\varphi)(u) = \big((\partial \varphi_i/\partial x_j)(u)\big)$ for the Jacobian matrix of φ (at u), then the above system is written in matrix form as

$$\nabla J(u) + (\mathrm{Jac}(\varphi)(u))^{\top} \lambda = 0,$$

where λ is viewed as a column vector, and the Lagrangian is equal to

$$L(u, \lambda) = J(u) + (\varphi_1(u), \ldots, \varphi_m(u))\lambda.$$

The beauty of the Lagrangian is that the constraints $\{\varphi_i(v) = 0\}$ have been incorporated into the function $L(v, \lambda)$, and that the necessary condition for the existence of a constrained local extremum of J is reduced to the necessary condition for the existence of a local extremum of the *unconstrained* L.

However, one should be careful to check that the assumptions of Theorem 40.2 are satisfied (in particular, the linear independence of the linear forms $d\varphi_i$).

Example 40.1. For example, let $J\colon \mathbb{R}^3 \to \mathbb{R}$ be given by

$$J(x, y, z) = x + y + z^2$$

and $g\colon \mathbb{R}^3 \to \mathbb{R}$ by

$$g(x, y, z) = x^2 + y^2.$$

Since $g(x, y, z) = 0$ iff $x = y = 0$, we have $U = \{(0, 0, z) \mid z \in \mathbb{R}\}$ and the restriction of J to U is given by

$$J(0, 0, z) = z^2,$$

which has a minimum for $z = 0$. However, a "blind" use of Lagrange multipliers would require that there is some λ so that

$$\frac{\partial J}{\partial x}(0,0,z) = \lambda \frac{\partial g}{\partial x}(0,0,z), \quad \frac{\partial J}{\partial y}(0,0,z) = \lambda \frac{\partial g}{\partial y}(0,0,z), \quad \frac{\partial J}{\partial z}(0,0,z) = \lambda \frac{\partial g}{\partial z}(0,0,z),$$

and since

$$\frac{\partial g}{\partial x}(x,y,z) = 2x, \quad \frac{\partial g}{\partial y}(x,y,z) = 2y, \quad \frac{\partial g}{\partial z}(0,0,z) = 0,$$

the partial derivatives above all vanish for $x = y = 0$, so at a local extremum we should also have

$$\frac{\partial J}{\partial x}(0,0,z) = 0, \quad \frac{\partial J}{\partial y}(0,0,z) = 0, \quad \frac{\partial J}{\partial z}(0,0,z) = 0,$$

but this is absurd since

$$\frac{\partial J}{\partial x}(x,y,z) = 1, \quad \frac{\partial J}{\partial y}(x,y,z) = 1, \quad \frac{\partial J}{\partial z}(x,y,z) = 2z.$$

The reader should enjoy finding the reason for the flaw in the argument.

One should also keep in mind that Theorem 40.2 gives only a necessary condition. The (u, λ) may *not* correspond to local extrema! Thus, it is always necessary to analyze the local behavior of J near a critical point u. This is generally difficult, but in the case where J is affine or quadratic and the constraints are affine or quadratic, this is possible (although not always easy).

Example 40.2. Let us apply the above method to the following example in which $E_1 = \mathbb{R}$, $E_2 = \mathbb{R}$, $\Omega = \mathbb{R}^2$, and

$$\begin{aligned} J(x_1, x_2) &= -x_2 \\ \varphi(x_1, x_2) &= x_1^2 + x_2^2 - 1. \end{aligned}$$

Observe that

$$U = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}$$

is the unit circle, and since

$$\nabla\varphi(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix},$$

it is clear that $\nabla\varphi(x_1, x_2) \neq 0$ for every point $= (x_1, x_2)$ on the unit circle. If we form the Lagrangian

$$L(x_1, x_2, \lambda) = -x_2 + \lambda(x_1^2 + x_2^2 - 1),$$

Theorem 40.2 says that a necessary condition for J to have a constrained local extremum is that $\nabla L(x_1, x_2, \lambda) = 0$, so the following equations must hold:

$$\begin{aligned} 2\lambda x_1 &= 0 \\ -1 + 2\lambda x_2 &= 0 \\ x_1^2 + x_2^2 &= 1. \end{aligned}$$

The second equation implies that $\lambda \neq 0$, and then the first yields $x_1 = 0$, so the third yields $x_2 = \pm 1$, and we get two solutions:

$$\begin{aligned} \lambda &= \frac{1}{2}, & (x_1, x_2) &= (0, 1) \\ \lambda &= -\frac{1}{2}, & (x_1', x_2') &= (0, -1). \end{aligned}$$

We can check immediately that the first solution is a minimum and the second is a maximum. The reader should look for a geometric interpretation of this problem.

Example 40.3. Let us now consider the case in which J is a quadratic function of the form

$$J(v) = \frac{1}{2} v^\top A v - v^\top b,$$

where A is an $n \times n$ symmetric matrix, $b \in \mathbb{R}^n$, and the constraints are given by a linear system of the form

$$Cv = d,$$

where C is an $m \times n$ matrix with $m < n$ and $d \in \mathbb{R}^m$. We also assume that C has rank m. In this case the function φ is given by

$$\varphi(v) = (Cv - d)^\top,$$

because we view $\varphi(v)$ as a row vector (and v as a column vector), and since

$$d\varphi(v)(w) = C^\top w,$$

the condition that the Jacobian matrix of φ at u have rank m is satisfied. The Lagrangian of this problem is

$$L(v, \lambda) = \frac{1}{2} v^\top A v - v^\top b + (Cv - d)^\top \lambda = \frac{1}{2} v^\top A v - v^\top b + \lambda^\top (Cv - d),$$

where λ is viewed as a column vector. Now because A is a symmetric matrix, it is easy to show that

$$\nabla L(v, \lambda) = \begin{pmatrix} Av - b + C^\top \lambda \\ Cv - d \end{pmatrix}.$$

Therefore, the necessary condition for constrained local extrema is

$$\begin{aligned} Av + C^\top \lambda &= b \\ Cv &= d, \end{aligned}$$

which can be expressed in matrix form as

$$\begin{pmatrix} A & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix},$$

where the matrix of the system is a symmetric matrix. We should not be surprised to find the system discussed later in Chapter 42, except for some renaming of the matrices and vectors involved. As we will show in Section 42.2, the function J has a minimum iff A is positive definite, so in general, if A is only a symmetric matrix, the critical points of the Lagrangian do *not* correspond to extrema of J.

Remark: If the Jacobian matrix $\mathrm{Jac}(\varphi)(v) = \big((\partial\varphi_i/\partial x_j)(v)\big)$ has rank m for all $v \in U$ (which is equivalent to the linear independence of the linear forms $d\varphi_i(v)$), then we say that $0 \in \mathbb{R}^m$ is a *regular value* of φ. In this case, it is known that

$$U = \{v \in \Omega \mid \varphi(v) = 0\}$$

is a *smooth submanifold of dimension* $n - m$ *of* $\mathbb{R}^n$. Furthermore, the set

$$T_vU = \{w \in \mathbb{R}^n \mid d\varphi_i(v)(w) = 0,\ 1 \leq i \leq m\} = \bigcap_{i=1}^{m} \mathrm{Ker}\, d\varphi_i(v)$$

is the *tangent space* to U at v (a vector space of dimension $n - m$). Then, the condition

$$dJ(v) + \mu_1 d\varphi_1(v) + \cdots + \mu_m d\varphi_m(v) = 0$$

implies that $dJ(v)$ vanishes on the tangent space T_vU. Conversely, if $dJ(v)(w) = 0$ for all $w \in T_vU$, this means that $dJ(v)$ is orthogonal (in the sense of Definition 11.3) to T_vU. Since (by Theorem 11.4 (b)) the orthogonal of T_vU is the space of linear forms spanned by $d\varphi_1(v), \ldots, d\varphi_m(v)$, it follows that $dJ(v)$ must be a linear combination of the $d\varphi_i(v)$. Therefore, when 0 is a regular value of φ, Theorem 40.2 asserts that if $u \in U$ is a local extremum of J, then $dJ(u)$ must vanish on the tangent space T_uU. We can say even more. The subset $Z(J)$ of Ω given by

$$Z(J) = \{v \in \Omega \mid J(v) = J(u)\}$$

(the *level set of level* $J(u)$) is a hypersurface in Ω, and if $dJ(u) \neq 0$, the zero locus of $dJ(u)$ is the tangent space $T_uZ(J)$ to $Z(J)$ at u (a vector space of dimension $n - 1$), where

$$T_uZ(J) = \{w \in \mathbb{R}^n \mid dJ(u)(w) = 0\}.$$

Consequently, Theorem 40.2 asserts that

$$T_uU \subseteq T_uZ(J);$$

this is a geometric condition.

We now return to the general situation where E_1 and E_2 may be infinite-dimensional normed vector spaces (with E_1 a Banach space) and we state and prove the following general result about the method of Lagrange multipliers.

Theorem 40.4. *(Necessary condition for a constrained extremum) Let $\Omega \subseteq E_1 \times E_2$ be an open subset of a product of normed vector spaces, with E_1 a Banach space (E_1 is complete), let $\varphi\colon \Omega \to E_2$ be a C^1-function (which means that $d\varphi(\omega)$ exists and is continuous for all $\omega \in \Omega$), and let*

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

Moreover, let $u = (u_1, u_2) \in U$ *be a point such that*

$$\frac{\partial \varphi}{\partial x_2}(u_1, u_2) \in \mathcal{L}(E_2; E_2) \quad \textit{and} \quad \left(\frac{\partial \varphi}{\partial x_2}(u_1, u_2)\right)^{-1} \in \mathcal{L}(E_2; E_2),$$

and let $J \colon \Omega \to \mathbb{R}$ *be a function which is differentiable at* u*. If* J *has a constrained local extremum at* u*, then there is a continuous linear form* $\Lambda(u) \in \mathcal{L}(E_2; \mathbb{R})$ *such that*

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

Proof. The plan of attack is to use the implicit function theorem; Theorem 39.14. Observe that the assumptions of Theorem 39.14 are indeed met. Therefore, there exist some open subsets $U_1 \subseteq E_1$, $U_2 \subseteq E_2$, and a continuous function $g \colon U_1 \to U_2$ with $(u_1, u_2) \in U_1 \times U_2 \subseteq \Omega$ and such that

$$\varphi(v_1, g(v_1)) = 0$$

for all $v_1 \in U_1$. Moreover, g is differentiable at $u_1 \in U_1$ and

$$dg(u_1) = -\left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u).$$

It follows that the restriction of J to $(U_1 \times U_2) \cap U$ yields a function G of a single variable, with

$$G(v_1) = J(v_1, g(v_1))$$

for all $v_1 \in U_1$. Now the function G is differentiable at u_1 and it has a local extremum at u_1 on U_1, so Proposition 40.1 implies that

$$dG(u_1) = 0.$$

By the chain rule,

$$\begin{aligned} dG(u_1) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \circ dg(u_1) \\ &= \frac{\partial J}{\partial x_1}(u) - \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u). \end{aligned}$$

From $dG(u_1) = 0$, we deduce

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u),$$

and since we also have

$$\frac{\partial J}{\partial x_2}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_2}(u),$$

if we let

$$\Lambda(u) = -\frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1},$$

then we get

$$\begin{aligned} dJ(u) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \\ &= \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \left(\frac{\partial \varphi}{\partial x_1}(u) + \frac{\partial \varphi}{\partial x_2}(u)\right) \\ &= -\Lambda(u) \circ d\varphi(u), \end{aligned}$$

which yields $dJ(u) + \Lambda(u) \circ d\varphi(u) = 0$, as claimed. $\square$

Finally, we prove Theorem 40.2.

Proof of Theorem 40.2. The linear independence of the m linear forms $d\varphi_i(u)$ is equivalent to the fact that the $m \times n$ matrix $A = \big((\partial \varphi_i/\partial x_j)(u)\big)$ has rank m. By reordering the columns, we may assume that the first m columns are linearly independent. To conform to the set-up of Theorem 40.4 we define E_1 and E_2 as

$$E_1 = \left\{ \sum_{i=m+1}^{n} v_i e_i \mid (v_{m+1}, \ldots, v_n) \in \mathbb{R}^{n-m} \right\}, \quad E_2 = \left\{ \sum_{i=1}^{m} v_i e_i \mid (v_1, \ldots, v_m) \in \mathbb{R}^{m} \right\}.$$

If we let $\psi \colon \Omega \to \mathbb{R}^m$ be the function defined by

$$\psi(v_{m+1}, \ldots, v_n, v_1, \ldots, v_m) = (\varphi_1(v), \ldots, \varphi_m(v))$$

for all $(v_{m+1}, \ldots, v_n, v_1, \ldots, v_m) \in \Omega$, with $v = (v_1, \ldots, v_n)$, then we see that $\partial \psi/\partial x_2(u)$ is invertible and both $\partial \psi/\partial x_2(u)$ and its inverse are continuous, so that Theorem 40.4 applies, and there is some (continuous) linear form $\Lambda(u) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\psi(u_{m+1}, \ldots, u_n, u_1, \ldots, u_m) = 0,$$

namely

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

However, $\Lambda(u)$ is defined by some m-tuple $(\lambda_1(u), \ldots, \lambda_m(u)) \in \mathbb{R}^m$, and in view of the definition of φ, the above equation is equivalent to

$$dJ(u) + \lambda_1(u) d\varphi_1(u) + \cdots + \lambda_m(u) d\varphi_m(u) = 0.$$

The uniqueness of the $\lambda_i(u)$ is a consequence of the linear independence of the $d\varphi_i(u)$. $\square$

We now investigate conditions for the existence of extrema involving the second derivative of J.

40.2 Using Second Derivatives to Find Extrema

For the sake of brevity, we consider only the case of local minima; analogous results are obtained for local maxima (replace J by $-J$, since $\max_u J(u) = -\min_u -J(u)$). We begin with a necessary condition for an unconstrained local minimum.

Proposition 40.5. *Let E be a normed vector space and let $J\colon \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J is differentiable in Ω, if J has a second derivative $\mathrm{D}^2 J(u)$ at some point $u \in \Omega$, and if J has a local minimum at u, then*

$$\mathrm{D}^2 J(u)(w, w) \geq 0 \quad \textit{for all } w \in E.$$

Proof. Pick any nonzero vector $w \in E$. Since Ω is open, for t small enough, $u + tw \in \Omega$ and $J(u + tw) \geq J(u)$, so there is some open interval $I \subseteq \mathbb{R}$ such that

$$u + tw \in \Omega \quad \text{and} \quad J(u + tw) \geq J(u)$$

for all $t \in I$. Using the Taylor–Young formula and the fact that we must have $dJ(u) = 0$ since J has a local minimum at u, we get

$$0 \leq J(u + tw) - J(u) = \frac{t^2}{2}\mathrm{D}^2 J(u)(w, w) + t^2 \left\|w\right\|^2 \epsilon(tw),$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$, which implies that

$$\mathrm{D}^2 J(u)(w, w) \geq 0.$$

Since the argument holds for all $w \in E$ (trivially if $w = 0$), the proposition is proven. $\square$

One should be cautioned that there is no converse to the previous proposition. For example, the function $f\colon x \mapsto x^3$ has no local minimum at 0, yet $df(0) = 0$ and $\mathrm{D}^2 f(0)(u, v) = 0$. Similarly, the reader should check that the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 - 3y^3$$

has no local minimum at $(0, 0)$; yet $df(0, 0) = 0$ since $df(x, y) = (2x, -9y^2)$, and for $u = (u_1, u_2)$, $\mathrm{D}^2 f(0, 0)(u, u) = 2u_1^2 \geq 0$ since

$$D^2 f(x, y)(u, u) = \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & -18y \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

See Figure 40.3.

When $E = \mathbb{R}^n$, Proposition 40.5 says that a necessary condition for having a local minimum is that the Hessian $\nabla^2 J(u)$ be positive semidefinite (it is always symmetric).

We now give sufficient conditions for the existence of a local minimum.

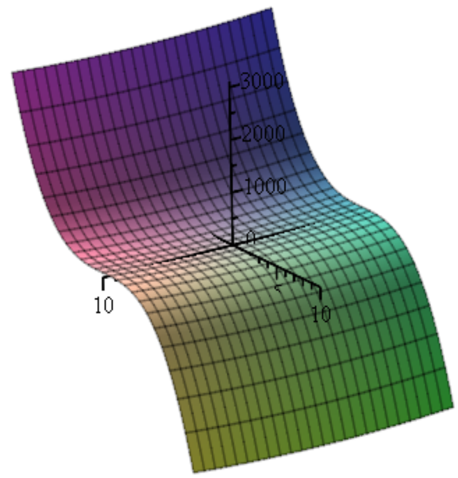


Figure 40.3: The graph of $f(x, y) = x^2 - 3y^3$. Note that $(0, 0)$ not a local extremum despite the fact that $df(0, 0) = 0$.

Theorem 40.6. *Let E be a normed vector space, let $J\colon \Omega \to \mathbb{R}$ be a function with Ω some open subset of E, and assume that J is differentiable in Ω and that $dJ(u) = 0$ at some point $u \in \Omega$. The following properties hold:*

(1) If $\mathrm{D}^2J(u)$ exists and if there is some number $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and

$$\mathrm{D}^2J(u)(w, w) \geq \alpha \left\|w\right\|^2 \quad \text{for all } w \in E,$$

then J has a strict local minimum at u.

(2) If $\mathrm{D}^2J(v)$ exists for all $v \in \Omega$ and if there is a ball $B \subseteq \Omega$ centered at u such that

$$\mathrm{D}^2J(v)(w, w) \geq 0 \quad \text{for all } v \in B \text{ and all } w \in E,$$

then J has a local minimum at u.

Proof. (1) Using the formula of Taylor–Young, for every vector w small enough, we can write

$$\begin{aligned} J(u + w) - J(u) &= \frac{1}{2}\mathrm{D}^2J(u)(w, w) + \left\|w\right\|^2 \epsilon(w) \\ &\geq \left(\frac{1}{2}\alpha + \epsilon(w)\right) \left\|w\right\|^2 \end{aligned}$$

with $\lim_{w \mapsto 0} \epsilon(w) = 0$. Consequently if we pick $r > 0$ small enough that $|\epsilon(w)| < \alpha/2$ for all w with $\|w\| < r$, then $J(u + w) > J(u)$ for all $u + w \in B$, where B is the open ball of center u and radius r. This proves that J has a local strict minimum at u.

(2) The formula of Taylor–Maclaurin shows that for all $u + w \in B$, we have

$$J(u+w) = J(u) + \frac{1}{2}\mathrm{D}^2 J(v)(w,w) \geq J(u),$$

for some $v \in (u, u+w)$ (recall that $(u, u+w) = \{(1-\lambda)(u+w) + \lambda(u+w) \mid 0 < \lambda < 1\}$). $\square$

There are no converses of the two assertions of Theorem 40.6. However, there is a condition on $\mathrm{D}^2 J(u)$ that implies the condition of Part (1). Since this condition is easier to state when $E = \mathbb{R}^n$, we begin with this case.

Recall that a $n \times n$ symmetric matrix A is *positive definite* if $x^\top A x > 0$ for all $x \in \mathbb{R}^n - \{0\}$. In particular, A must be invertible.

Proposition 40.7. *For any symmetric matrix A, if A is positive definite, then there is some $\alpha > 0$ such that*

$$x^\top A x \geq \alpha \left\|x\right\|^2 \quad \text{for all } x \in \mathbb{R}^n.$$

Proof. Pick any norm in $\mathbb{R}^n$ (recall that all norms on $\mathbb{R}^n$ are equivalent). Since the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ is compact and since the function $f(x) = x^\top A x$ is never zero on S^{n-1}, the function f has a minimum $\alpha > 0$ on S^{n-1}. Using the usual trick that $x = \|x\| (x/\|x\|)$ for every nonzero vector $x \in \mathbb{R}^n$ and the fact that the inequality of the proposition is trivial for $x = 0$, from

$$x^\top A x \geq \alpha \quad \text{for all } x \text{ with } \|x\| = 1,$$

we get

$$x^\top A x \geq \alpha \left\|x\right\|^2 \quad \text{for all } x \in \mathbb{R}^n,$$

as claimed. $\square$

We can combine Theorem 40.6 and Proposition 40.7 to obtain a useful sufficient condition for the existence of a strict local minimum. First let us introduce some terminology.

Definition 40.6. Given a function $J \colon \Omega \to \mathbb{R}$ as before, say that a point $u \in \Omega$ is a *nondegenerate critical point* if $dJ(u) = 0$ and if the Hessian matrix $\nabla^2 J(u)$ is invertible.

Proposition 40.8. *Let $J \colon \Omega \to \mathbb{R}$ be a function defined on some open subset $\Omega \subseteq \mathbb{R}^n$. If J is differentiable in Ω and if some point $u \in \Omega$ is a nondegenerate critical point such that $\nabla^2 J(u)$ is positive definite, then J has a strict local minimum at u.*

Remark: It is possible to generalize Proposition 40.8 to infinite-dimensional spaces by finding a suitable generalization of the notion of a nondegenerate critical point. Firstly, we assume that E is a Banach space (a complete normed vector space). Then we define the dual E' of E as the set of continuous linear forms on E, so that $E' = \mathcal{L}(E; \mathbb{R})$. Following Lang, we use the notation E' for the space of continuous linear forms to avoid confusion with the space $E^* = \mathrm{Hom}(E, \mathbb{R})$ of all linear maps from E to $\mathbb{R}$. A continuous bilinear map $\varphi\colon E \times E \to \mathbb{R}$ in $\mathcal{L}_2(E, E; \mathbb{R})$ yields a map Φ from E to E' given by

$$\Phi(u) = \varphi_u,$$

where $\varphi_u \in E'$ is the linear form defined by

$$\varphi_u(v) = \varphi(u, v).$$

It is easy to check that φ_u is continuous and that the map Φ is continuous. Then we say that φ is *nondegenerate* iff $\Phi\colon E \to E'$ is an isomorphism of Banach spaces, which means that Φ is invertible and that both Φ and Φ^{-1} are continuous linear maps. Given a function $J\colon \Omega \to \mathbb{R}$ differentiable on Ω as before (where Ω is an open subset of E), if $\mathrm{D}^2 J(u)$ exists for some $u \in \Omega$, we say that u is a *nondegenerate critical point* if $dJ(u) = 0$ and if $\mathrm{D}^2 J(u)$ is nondegenerate. Of course, $\mathrm{D}^2 J(u)$ is positive definite if $\mathrm{D}^2 J(u)(w, w) > 0$ for all $w \in E - \{0\}$.

Using the above definition, Proposition 40.7 can be generalized to a nondegenerate positive definite bilinear form (on a Banach space) and Theorem 40.8 can also be generalized to the situation where $J\colon \Omega \to \mathbb{R}$ is defined on an open subset of a Banach space. For details and proofs, see Cartan [34] (Part I Chapter 8) and Avez [9] (Chapter 8 and Chapter 10).

In the next section we make use of convexity; both on the domain Ω and on the function J itself.

40.3 Using Convexity to Find Extrema

We begin by reviewing the definition of a convex set and of a convex function.

Definition 40.7. Given any real vector space E, we say that a subset C of E is *convex* if either $C = \emptyset$ or if for every pair of points $u, v \in C$, the line segment connecting u and v is contained in C, i.e.,

$$(1 - \lambda)u + \lambda v \in C \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1.$$

Given any two points $u, v \in E$, the *line segment* $[u, v]$ is the set

$$[u, v] = \{(1 - \lambda)u + \lambda v \in E \mid \lambda \in \mathbb{R},\ 0 \leq \lambda \leq 1\}.$$

Clearly, a nonempty set C is convex iff $[u, v] \subseteq C$ whenever $u, v \in C$. See Figure 40.4 for an example of a convex set.

Definition 40.8. If C is a nonempty convex subset of E, a function $f\colon C \to \mathbb{R}$ is *convex* (on C) if for every pair of points $u, v \in C$,

$$f((1 - \lambda)u + \lambda v) \leq (1 - \lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1;$$

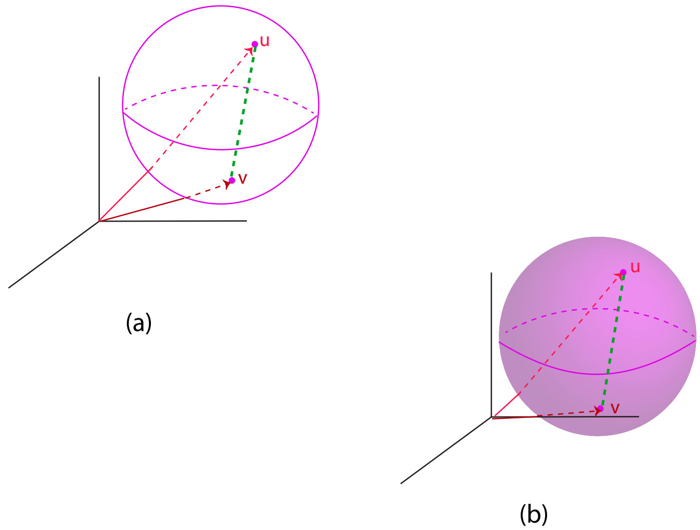


Figure 40.4: Figure (a) shows that a sphere is not convex in $\mathbb{R}^3$ since the dashed green line does not lie on its surface. Figure (b) shows that a solid ball is convex in $\mathbb{R}^3$.

the function f is *strictly convex* (on C) if for every pair of distinct points $u, v \in C$ $(u \neq v)$,

$$f((1-\lambda)u + \lambda v) < (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 < \lambda < 1;$$

see Figure 40.5. The *epigraph*[1] $\mathbf{epi}(f)$ of a function $f\colon A \to \mathbb{R}$ defined on some subset A of $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ defined as

$$\mathbf{epi}(f) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y,\ x \in A\}.$$

A function $f\colon C \to \mathbb{R}$ defined on a convex subset C is *concave* (resp. *strictly concave*) if $(-f)$ is convex (resp. strictly convex).

It is obvious that a function f is convex iff its epigraph $\mathbf{epi}(f)$ is a convex subset of $\mathbb{R}^{n+1}$.

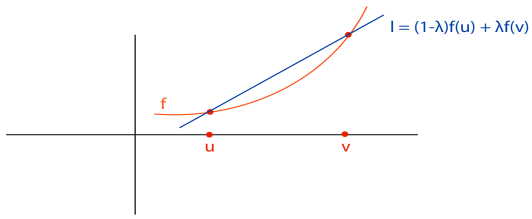
Example 40.4. Here are some common examples of convex sets.

- Subspaces $V \subseteq E$ of a vector space E are convex.
- *Affine subspaces*, that is, sets of the form $u + V$, where V is a subspace of E and $u \in E$, are convex.
- Balls (open or closed) are convex. Given any linear form $\varphi\colon E \to \mathbb{R}$, for any scalar $c \in \mathbb{R}$, the *closed half–spaces*

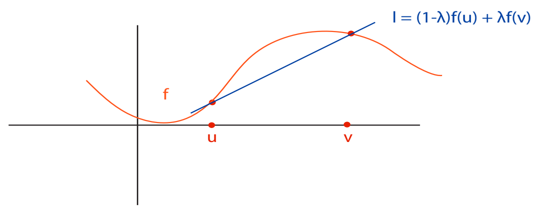
$$H^+_{\varphi,c} = \{u \in E \mid \varphi(u) \geq c\}, \qquad H^-_{\varphi,c} = \{u \in E \mid \varphi(u) \leq c\},$$

are convex.

[1] "Epi" means above.



(a)



(b)

Figure 40.5: Figures (a) and (b) are the graphs of real valued functions. Figure (a) is the graph of convex function since the blue line lies above the graph of f. Figure (b) shows the graph of a function which is not convex.

- Any intersection of half–spaces is convex.
- More generally, any intersection of convex sets is convex.

Example 40.5. Here are some common examples of convex and concave functions.

- Linear forms are convex functions (but not strictly convex).
- Any norm $\| \, \| \colon E \to \mathbb{R}_+$ is a convex function.
- The max function,
$$\max(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$$
is convex on $\mathbb{R}^n$.
- The exponential $x \mapsto e^{cx}$ is strictly convex for any $c \neq 0$ ($c \in \mathbb{R}$).
- The logarithm function is concave on $\mathbb{R}_+ - \{0\}$.
- The *log-determinant function* log det is concave on the set of symmetric positive definite matrices. This function plays an important role in convex optimization.

An excellent exposition of convexity and its applications to optimization can be found in Boyd [29].

Here is a necessary condition for a function to have a local minimum with respect to a convex subset U.

Theorem 40.9. *(Necessary condition for a local minimum on a convex subset) Let $J\colon \Omega \to \mathbb{R}$ be a function defined on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset. Given any $u \in U$, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

Proof. Let $v = u + w$ be an arbitrary point in U. Since U is convex, we have $u + tw \in U$ for all t such that $0 \leq t \leq 1$. Since $dJ(u)$ exists, we can write

$$J(u+tw) - J(u) = dJ(u)(tw) + \|tw\|\, \epsilon(tw)$$

with $\lim_{t\mapsto 0} \epsilon(tw) = 0$. However, because $0 \leq t$,

$$J(u+tw) - J(u) = t(dJ(u)(w) + \|w\|\, \epsilon(tw))$$

and since u is a local minimum with respect to U, we have $J(u+tw) - J(u) \geq 0$, so we get

$$t(dJ(u)(w) + \|w\|\, \epsilon(tw)) \geq 0.$$

The above implies that $dJ(u)(w) \geq 0$, because otherwise we could pick $t > 0$ small enough so that

$$dJ(u)(w) + \|w\|\, \epsilon(tw) < 0,$$

a contradiction. Since the argument holds for all $v = u + w \in U$, the theorem is proven. ☐

Observe that the convexity of U is a substitute for the use of Lagrange multipliers, but we now have to deal with an *inequality* instead of an equality.

In the special case where U is a subspace of E we have the following result.

Corollary 40.10. *With the same assumptions as in Theorem 40.9, if U is a subspace of E, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(w) = 0 \quad \textit{for all } w \in U.$$

Proof. In this case since $u \in U$ we have $2u \in U$, and for any $u + w \in U$, we must have $2u-(u+w) = u-w \in U$. The previous theorem implies that $dJ(u)(w) \geq 0$ and $dJ(u)(-w) \geq 0$, that is, $dJ(u)(w) \leq 0$, so $dJ(u) = 0$. Since the argument holds for $w \in U$ (because U is a subspace, if $u, w \in U$, then $u + w \in U$), we conclude that

$$dJ(u)(w) = 0 \quad \text{for all } w \in U.$$ ☐

We will now characterize convex functions when they have a first derivative or a second derivative.

Proposition 40.11. *(Convexity and first derivative) Let $f \colon \Omega \to \mathbb{R}$ be a function differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$f(v) \geq f(u) + df(u)(v-u) \quad \text{for all } u, v \in U.$$

(2) The function f is strictly convex on U iff

$$f(v) > f(u) + df(u)(v-u) \quad \text{for all } u, v \in U \text{ with } u \neq v.$$

See Figure 40.6.

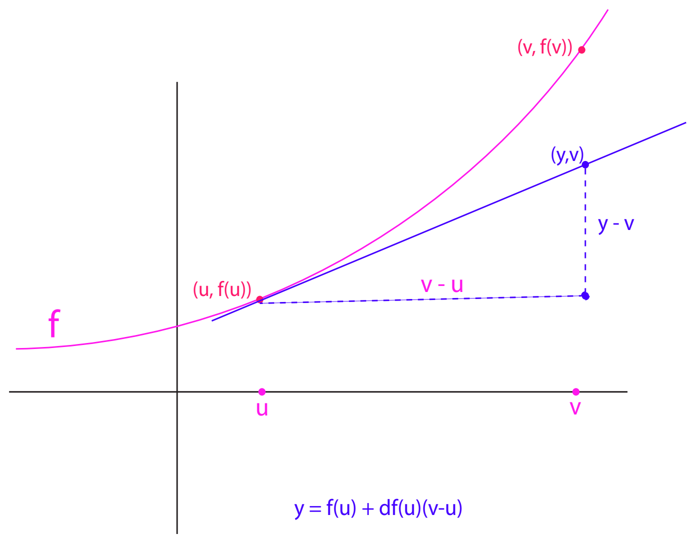


Figure 40.6: An illustration of a convex valued function f. Since f is convex it always lies above its tangent line.

Proof. Let $u, v \in U$ be any two distinct points and pick $\lambda \in \mathbb{R}$ with $0 < \lambda < 1$. If the function f is convex, then

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v),$$

which yields

$$\frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

It follows that

$$df(u)(v-u) = \lim_{\lambda \mapsto 0} \frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

If f is strictly convex, the above reasoning does not work, because a strict inequality is not necessarily preserved by "passing to the limit." We have recourse to the following trick: for any ω such that $0 < \omega < 1$, observe that

$$(1-\lambda)u + \lambda v = u + \lambda(v-u) = \frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u + \omega(v-u)).$$

If we assume that $0 < \lambda \leq \omega$, the convexity of f yields

$$f(u+\lambda(v-u)) = f\left(\left(1-\frac{\lambda}{\omega}\right)u + \frac{\lambda}{\omega}(u+\omega(v-u))\right) \leq \frac{\omega-\lambda}{\omega}f(u) + \frac{\lambda}{\omega}f(u+\omega(v-u)).$$

If we subtract $f(u)$ to both sides, we get

$$\frac{f(u+\lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega}.$$

Now since $0 < \omega < 1$ and f is strictly convex,

$$f(u+\omega(v-u)) = f((1-\omega)u + \omega v) < (1-\omega)f(u) + \omega f(v),$$

which implies that

$$\frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

and thus we get

$$\frac{f(u+\lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u).$$

If we let λ go to 0, by passing to the limit we get

$$df(u)(v-u) \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

which yields the desired strict inequality.

Let us now consider the converse of (1); that is, assume that

$$f(v) \geq f(u) + df(u)(v-u) \quad \text{for all } u, v \in U.$$

For any two distinct points $u, v \in U$ and for any λ with $0 < \lambda < 1$, we get

$$\begin{aligned} f(v) &\geq f(v+\lambda(u-v)) - \lambda df(v+\lambda(u-v))(u-v) \\ f(u) &\geq f(v+\lambda(u-v)) + (1-\lambda)df(v+\lambda(u-v))(u-v), \end{aligned}$$

and if we multiply the first inequality by $1-\lambda$ and the second inequality by λ and them add up the resulting inequalities, we get

$$(1-\lambda)f(v) + \lambda f(u) \geq f(v+\lambda(u-v)) = f((1-\lambda)v + \lambda u),$$

which proves that f is convex.

The proof of the converse of (2) is similar, except that the inequalities are replaced by strict inequalities. $\square$

We now establish a convexity criterion using the second derivative of f. This criterion is often easier to check than the previous one.

Proposition 40.12. *(Convexity and second derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function twice differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$\mathrm{D}^2 f(u)(v-u, v-u) \geq 0 \quad \text{for all } u, v \in U.$$

(2) If

$$\mathrm{D}^2 f(u)(v-u, v-u) > 0 \quad \text{for all } u, v \in U \text{ with } u \neq v,$$

then f is strictly convex.

Proof. First assume that the inequality in Condition (1) is satisfied. For any two distinct points $u, v \in U$, the formula of Taylor–Maclaurin yields

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}\mathrm{D}^2 f(w)(v-u, v-u) \\ &= \frac{\rho^2}{2}\mathrm{D}^2 f(w)(v-w, v-w), \end{aligned}$$

for some $w = (1-\lambda)u + \lambda v = u + \lambda(v-u)$ with $0 < \lambda < 1$, and with $\rho = 1/(1-\lambda) > 0$, so that $v - u = \rho(v-w)$. Since $\mathrm{D}^2 f(w)(v-w, v-w) \geq 0$ for all $u, w \in U$, we conclude by applying Proposition 40.11(1).

Similarly, if (2) holds, the above reasoning and Proposition 40.11(2) imply that f is strictly convex.

To prove the necessary condition in (1), define $g\colon \Omega \to \mathbb{R}$ by

$$g(v) = f(v) - df(u)(v),$$

where $u \in U$ is any point considered fixed. If f is convex, since

$$g(v) - g(u) = f(v) - f(u) - df(u)(v-u),$$

Proposition 40.11 implies that $f(v) - f(u) - df(u)(v-u) \geq 0$, which implies that g has a local minimum at u with respect to all $v \in U$. Therefore, we have $dg(u) = 0$. Observe that g is twice differentiable in Ω and $\mathrm{D}^2 g(u) = \mathrm{D}^2 f(u)$, so the formula of Taylor–Young yields for every $v = u + w \in U$ and all t with $0 \leq t \leq 1$,

$$\begin{aligned} 0 \leq g(u+tw) - g(u) &= \frac{t^2}{2}\mathrm{D}^2 f(u)(tw, tw) + \|tw\|^2 \epsilon(tw) \\ &= \frac{t^2}{2}(\mathrm{D}^2 f(u)(w, w) + 2\|w\|^2 \epsilon(wt)), \end{aligned}$$

with $\lim_{t \mapsto 0} \epsilon(wt) = 0$, and for t small enough, we must have $\mathrm{D}^2 f(u)(w, w) \geq 0$, as claimed. $\square$

The converse of Proposition 40.12 (2) is false as we see by considering the strictly convex function f given by $f(x) = x^4$ and its second derivative at $x = 0$.

Example 40.6. On the other hand, if f is a quadratic function of the form

$$f(u) = \frac{1}{2}u^\top Au - u^\top b$$

where A is a symmetric matrix, we know that

$$df(u)(v) = v^\top (Au - b),$$

so

$$\begin{aligned}
f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}v^\top Av - v^\top b - \frac{1}{2}u^\top Au + u^\top b - (v-u)^\top (Au - b) \\
&= \frac{1}{2}v^\top Av - \frac{1}{2}u^\top Au - (v-u)^\top Au \\
&= \frac{1}{2}v^\top Av + \frac{1}{2}u^\top Au - v^\top Au \\
&= \frac{1}{2}(v-u)^\top A(v-u).
\end{aligned}$$

Therefore, Proposition 40.11 implies that if A is positive semidefinite, then f is convex and if A is positive definite, then f is strictly convex. The converse follows by Proposition 40.12.

We conclude this section by applying our previous theorems to convex functions defined on convex subsets. In this case local minima (resp. local maxima) are global minima (resp. global maxima). The next definition is the special case of Definition 40.1 in which $W = E$ but it does not hurt to state it explicitly.

Definition 40.9. Let $f\colon E \to \mathbb{R}$ be any function defined on some normed vector space (or more generally, any set). For any $u \in E$, we say that f has a *minimum* in u (resp. *maximum* in u) if

$$f(u) \leq f(v) \;\; (\text{resp. } f(u) \geq f(v)) \quad \text{for all } v \in E.$$

We say that f has a *strict minimum* in u (resp. *strict maximum* in u) if

$$f(u) < f(v) \;\; (\text{resp. } f(u) > f(v)) \quad \text{for all } v \in E - \{u\}.$$

If $U \subseteq E$ is a subset of E and $u \in U$, we say that f has a *minimum* in u (resp. *strict minimum* in u) *with respect to U* if

$$f(u) \leq f(v) \qquad \text{for all } v \in U \quad (\text{resp. } f(u) < f(v) \qquad \text{for all } v \in U - \{u\}),$$

and similarly for a *maximum* in u (resp. *strict maximum* in u) *with respect to U* with $\leq$ changed to $\geq$ and $<$ to $>$.

Sometimes, we say *global* maximum (or minimum) to stress that a maximum (or a minimum) is not simply a local maximum (or minimum).

Theorem 40.13. *Given any normed vector space E, let U be any nonempty convex subset of E.*

(1) For any convex function $J\colon U \to \mathbb{R}$, for any $u \in U$, if J has a local minimum at u in U, then J has a (global) minimum at u in U.

(2) Any strict convex function $J\colon U \to \mathbb{R}$ has at most one minimum (in U), and if it does, then it is a strict minimum (in U).

(3) Let $J\colon \Omega \to \mathbb{R}$ be any function defined on some open subset Ω of E with $U \subseteq \Omega$ and assume that J is convex on U. For any point $u \in U$, if $dJ(u)$ exists, then J has a minimum in u with respect to U iff

$$dJ(u)(v - u) \geq 0 \quad \text{for all } v \in U.$$

(4) If the convex subset U in (3) is open, then the above condition is equivalent to

$$dJ(u) = 0.$$

Proof. (1) Let $v = u + w$ be any arbitrary point in U. Since J is convex, for all t with $0 \leq t \leq 1$, we have

$$J(u + tw) = J(u + t(v - u)) = J((1 - t)u + tv) \leq (1 - t)J(u) + tJ(v),$$

which yields

$$J(u + tw) - J(u) \leq t(J(v) - J(u)).$$

Because J has a local minimum at u, there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u) \leq t_0(J(v) - J(u)),$$

which implies that $J(v) - J(u) \geq 0$.

(2) If J is strictly convex, the above reasoning with $w \neq 0$ shows that there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u) < t_0(J(v) - J(u)),$$

which shows that u is a strict global minimum (in U), and thus that it is unique.

(3) We already know from Theorem 40.9 that the condition $dJ(u)(v-u) \geq 0$ for all $v \in U$ is necessary (even if J is not convex). Conversely, because J is convex, careful inspection of the proof of Part (1) of Proposition 40.11 shows that only the fact that $dJ(u)$ exists is needed to prove that

$$J(v) - J(u) \geq dJ(u)(v - u) \quad \text{for all } v \in U,$$

and if

$$dJ(u)(v-u) \geq 0 \quad \text{for all } v \in U,$$

then

$$J(v) - J(u) \geq 0 \quad \text{for all } v \in U,$$

as claimed.

(4) If U is open, then for every $u \in U$ we can find an open ball B centered at u of radius ϵ small enough so that $B \subseteq U$. Then for any $w \neq 0$ such that $\|w\| < \epsilon$, we have both $v = u + w \in B$ and $v' = u - w \in B$, so Condition (3) implies that

$$dJ(u)(w) \geq 0 \quad \text{and} \quad dJ(u)(-w) \geq 0,$$

which yields

$$dJ(u)(w) = 0.$$

Since the above holds for all $w \neq 0$ such such that $\|w\| < \epsilon$ and since $dJ(u)$ is linear, we leave it to the reader to fill in the details of the proof that $dJ(u) = 0$. $\square$

Example 40.7. Theorem 40.13 can be used to rederive the fact that the least squares solutions of a linear system $Ax = b$ (where A is an $m \times n$ matrix) are given by the normal equation

$$A^\top Ax = A^\top b.$$

For this, we consider the quadratic function

$$J(v) = \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2,$$

and our least squares problem is equivalent to finding the minima of J on $\mathbb{R}^n$. A computation reveals that

$$\begin{aligned}
J(v) &= \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2 \\
&= \frac{1}{2}(Av - b)^\top (Av - b) - \frac{1}{2}b^\top b \\
&= \frac{1}{2}(v^\top A^\top - b^\top)(Av - b) - \frac{1}{2}b^\top b \\
&= \frac{1}{2}v^\top A^\top Av - v^\top A^\top b,
\end{aligned}$$

and so

$$dJ(u) = A^\top Au - A^\top b.$$

Since $A^\top A$ is positive semidefinite, the function J is convex, and Theorem 40.13(4) implies that the minima of J are the solutions of the equation

$$A^\top Au - A^\top b = 0.$$

The considerations in this chapter reveal the need to find methods for finding the zeros of the derivative map

$$dJ\colon \Omega \to E',$$

where Ω is some open subset of a normed vector space E and E' is the space of all continuous linear forms on E (a subspace of E^*). Generalizations of *Newton's method* yield such methods and they are the object of the next chapter.

40.4 Summary

The main concepts and results of this chapter are listed below:

- Local minimum, local maximum, local extremum, strict local minimum, strict local maximum.
- Necessary condition for a local extremum involving the derivative; critical point.
- Local minimum with respect to a subset U, local maximum with respect to a subset U, local extremum with respect to a subset U.
- Constrained local extremum.
- Necessary condition for a constrained extremum.
- Necessary condition for a constrained extremum in terms of Lagrange multipliers.
- Lagrangian.
- Critical points of a Lagrangian.
- Necessary condition of an unconstrained local minimum involving the second-order derivative.
- Sufficient condition for a local minimum involving the second-order derivative.
- A sufficient condition involving nondegenerate critical points.
- Convex sets, convex functions, concave functions, strictly convex functions, strictly concave functions.
- Necessary condition for a local minimum on a convex set involving the derivative.
- Convexity of a function involving a condition on its first derivative.
- Convexity of a function involving a condition on its second derivative.
- Minima of convex functions on convex sets.

40.5 Problems

Problem 40.1. Find the extrema of the function $J(v_1, v_2) = v_2^2$ on the subset U given by

$$U = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1^2 + v_2^2 - 1 = 0\}.$$

Problem 40.2. Find the extrema of the function $J(v_1, v_2) = v_1 + (v_2 - 1)^2$ on the subset U given by

$$U = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1^2 = 0\}.$$

Problem 40.3. Let A be an $n \times n$ real symmetric matrix, B an $n \times n$ symmetric positive definite matrix, and let $b \in \mathbb{R}^n$.

(1) Prove that a necessary condition for the function J given by

$$J(v) = \frac{1}{2} v^\top A v - b^\top v$$

to have an extremum at $u \in U$, with U defined by

$$U = \{v \in \mathbb{R}^n \mid v^\top B v = 1\},$$

is that there is some $\lambda \in \mathbb{R}$ such that

$$Au - b = \lambda B u.$$

(2) Prove that there is a symmetric positive definite matrix S such that $B = S^2$. Prove that if $b = 0$, then λ is an eigenvalue of the symmetric matrix $S^{-1} A S^{-1}$.

(3) Prove that for all $(u, \lambda) \in U \times \mathbb{R}$, if $Au - b = \lambda B u$, then

$$J(v) - J(u) = \frac{1}{2}(v - u)^\top (A - \lambda B)(v - u)$$

for all $v \in U$. Deduce that without additional assumptions, it is not possible to conclude that u is an extremum of J on U.

Problem 40.4. Let E be a normed vector space, and let U be a subset of E such that for all $u, v \in U$, we have $(1/2)(u + v) \in U$.

(1) Prove that if U is closed, then U is convex.

Hint. Every real $\theta \in (0, 1)$ can be written as

$$\theta = \sum_{n \geq 1} \alpha_n 2^{-n},$$

with $\alpha_n \in \{0, 1\}$.

(2) Does the result in (1) hold if U is not closed?

Problem 40.5. Prove that the function f with domain $\mathrm{dom}(f) = \mathbb{R} - \{0\}$ given by $f(x) = 1/x^2$ has the property that $f''(x) > 0$ for all $x \in \mathrm{dom}(f)$, but it is not convex. Why isn't Proposition 40.12 applicable?

Problem 40.6. (1) Prove that the function $x \mapsto e^{ax}$ (on $\mathbb{R}$) is convex for any $a \in \mathbb{R}$.

(2) Prove that the function $x \mapsto x^a$ is convex on $\{x \in \mathbb{R} \mid x > 0\}$, for all $a \in \mathbb{R}$ such that $a \leq 0$ or $a \geq 1$.

Problem 40.7. (1) Prove that the function $x \mapsto |x|^p$ is convex on $\mathbb{R}$ for all $p \geq 1$.

(2) Prove that the function $x \mapsto \log x$ is concave on $\{x \in \mathbb{R} \mid x > 0\}$.

(3) Prove that the function $x \mapsto x \log x$ is convex on $\{x \in \mathbb{R} \mid x > 0\}$.

Problem 40.8. (1) Prove that the function f given by $f(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ is convex on $\mathbb{R}^n$.

(2) Prove that the function g given by $g(x_1, \ldots, x_n) = \log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbb{R}^n$.

Prove that

$$\max\{x_1, \ldots, x_n\} \leq g(x_1, \ldots, x_n) \leq \max\{x_1, \ldots, x_n\} + \log n.$$

Problem 40.9. In Problem 39.6, it was shown that

$$\begin{aligned} df_A(X) &= \mathrm{tr}(A^{-1}X) \\ \mathrm{D}^2 f(A)(X_1, X_2) &= -\mathrm{tr}(A^{-1}X_1A^{-1}X_2), \end{aligned}$$

for all $n \times n$ real matrices X, X_1, X_2, where f is the function defined on $\mathbf{GL}^+(n, \mathbb{R})$ (the $n \times n$ real invertible matrices of positive determinants), given by

$$f(A) = \log \det(A).$$

Assume that A is symmetric positive definite and that X is symmetric.

(1) Prove that the eigenvalues of $A^{-1}X$ are real (even though $A^{-1}X$ may **not** be symmetric).

Hint. Since A is symmetric positive definite, then so is A^{-1}, so we can write $A^{-1} = S^2$ for some symmetric positive definite matrix S, and then

$$A^{-1}X = S^2X = S(SXS)S^{-1}.$$

(2) Prove that the eigenvalues of $(A^{-1}X)^2$ are nonnegative. Deduce that

$$\mathrm{D}^2 f(A)(X, X) = -\mathrm{tr}((A^{-1}X)^2) < 0$$

for all nonzero symmetric matrices X and SPD matrices A. Conclude that the function $X \mapsto \log \det X$ is strictly concave on the set of symmetric positive definite matrices.

Chapter 41

Newton's Method and Its Generalizations

In Chapter 40 we investigated the problem of determining when a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum. Proposition 40.1 gives a necessary condition when J is differentiable: if J has a local extremum at $u \in \Omega$, then we must have

$$J'(u) = 0.$$

Thus we are led to the problem of finding the zeros of the derivative

$$J'\colon \Omega \to E',$$

where $E' = \mathcal{L}(E; \mathbb{R})$ is the set of linear continuous functions from E to $\mathbb{R}$; that is, the *dual* of E, as defined in the remark after Proposition 40.8.

This leads us to consider the problem in a more general form, namely, given a function $f\colon \Omega \to Y$ from an open subset Ω of a normed vector space X to a normed vector space Y, find

(i) Sufficient conditions which guarantee the *existence of a zero* of the function f; that is, an element $a \in \Omega$ such that $f(a) = 0$.

(ii) An *algorithm* for approximating such an a, that is, a sequence (x_k) of points of Ω whose limit is a.

In this chapter we discuss Newton's method and some of it generalizations to give (partial) answers to Problems (i) and (i).

41.1 Newton's Method for Real Functions of a Real Argument

When $X = Y = \mathbb{R}$, we can use *Newton's method* to find a zero of a function $f\colon \Omega \to \mathbb{R}$. We pick some initial element $x_0 \in \mathbb{R}$ "close enough" to a zero a of f, and we define the sequence

(x_k) by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

for all $k \geq 0$, provided that $f'(x_k) \neq 0$. The idea is to define x_{k+1} as the intersection of the x-axis with the tangent line to the graph of the function $x \mapsto f(x)$ at the point $(x_k, f(x_k))$. Indeed, the equation of this tangent line is

$$y - f(x_k) = f'(x_k)(x - x_k),$$

and its intersection with the x-axis is obtained for $y = 0$, which yields

$$x = x_k - \frac{f(x_k)}{f'(x_k)},$$

as claimed. See Figure 41.1.

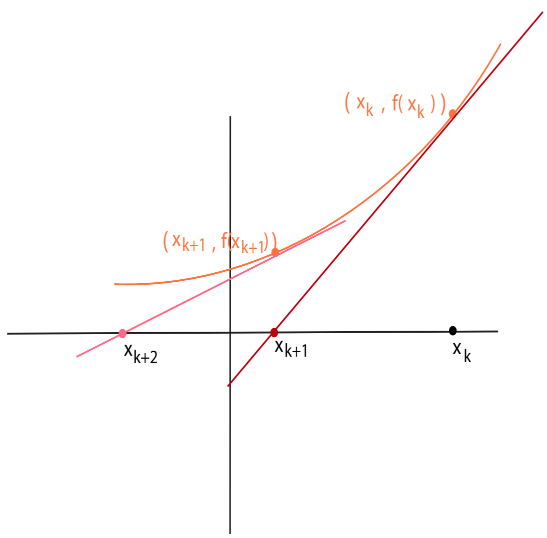


Figure 41.1: The construction of two stages in Newton's method.

Example 41.1. If $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α. It can be shown that the method converges to $\sqrt{\alpha}$ for any $x_0 > 0$; see Problem 41.1. Actually, the method also converges when $x_0 < 0$! Find out what is the limit.

The case of a real function suggests the following method for finding the zeros of a function $f\colon \Omega \to Y$, with $\Omega \subseteq X$: given a starting point $x_0 \in \Omega$, the sequence (x_k) is defined by

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)) \tag{$*$}$$

for all $k \geq 0$.

For the above to make sense, it must be ensured that

(1) All the points x_k remain within Ω.

(2) The function f is differentiable within Ω.

(3) The derivative $f'(x)$ is a bijection from X to Y for all $x \in \Omega$.

These are rather demanding conditions but there are sufficient conditions that guarantee that they are met. Another practical issue is that it may be very costly to compute $(f'(x_k))^{-1}$ at every iteration step. In the next section we investigate generalizations of Newton's method which address the issues that we just discussed.

41.2 Generalizations of Newton's Method

Suppose that $f\colon \Omega \to \mathbb{R}^n$ is given by n functions $f_i\colon \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$. In this case, finding a zero a of f is equivalent to solving the system

$$\begin{aligned} f_1(a_1 \ldots, a_n) &= 0 \\ f_2(a_1 \ldots, a_n) &= 0 \\ &\vdots \\ f_n(a_1 \ldots, a_n) &= 0. \end{aligned}$$

In the standard Newton method, the iteration step is given by $(*)$, namely

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)),$$

and if we define Δx_k as $\Delta x_k = x_{k+1} - x_k$, we see that $\Delta x_k = -(f'(x_k))^{-1}(f(x_k))$, so Δx_k is obtained by solving the equation

$$f'(x_k)\Delta x_k = -f(x_k),$$

and then we set $x_{k+1} = x_k + \Delta x_k$.

The generalization is as follows.

Variant 1. A single iteration of Newton's method consists in solving the linear system

$$(J(f)(x_k))\Delta x_k = -f(x_k),$$

and then setting

$$x_{k+1} = x_k + \Delta x_k,$$

where $J(f)(x_k) = \left(\frac{\partial f_i}{\partial x_j}(x_k)\right)$ is the Jacobian matrix of f at x_k.

In general it is very costly to compute $J(f)(x_k)$ at each iteration and then to solve the corresponding linear system. If the method converges, the consecutive vectors x_k should differ only a little, as also the corresponding matrices $J(f)(x_k)$. Thus, we are led to several variants of Newton's method.

Variant 2. This variant consists in keeping the same matrix for p consecutive steps (where p is some fixed integer ≥ 2):

$$\begin{aligned}
x_{k+1} &= x_k - (f'(x_0))^{-1}(f(x_k)), && 0 \leq k \leq p-1 \\
x_{k+1} &= x_k - (f'(x_p))^{-1}(f(x_k)), && p \leq k \leq 2p-1 \\
&\vdots \\
x_{k+1} &= x_k - (f'(x_{rp}))^{-1}(f(x_k)), && rp \leq k \leq (r+1)p-1 \\
&\vdots
\end{aligned}$$

Variant 3. Set $p = \infty$, that is, use *the same matrix $f'(x_0)$ for all iterations*, which leads to iterations of the form

$$x_{k+1} = x_k - (f'(x_0))^{-1}(f(x_k)), \quad k \geq 0,$$

Variant 4. Replace $f'(x_0)$ by a particular matrix A_0 which is easy to invert:

$$x_{k+1} = x_k - A_0^{-1} f(x_k), \quad k \geq 0.$$

In the last two cases, if possible, we use an LU factorization of $f'(x_0)$ or A_0 to speed up the method. In some cases, it may even possible to set $A_0 = I$.

The above considerations lead us to the definition of a *generalized Newton method*, as in Ciarlet [41] (Chapter 7). Recall that a linear map $f \in \mathcal{L}(E; F)$ is called an *isomorphism* iff f is continuous, bijective, and f^{-1} is also continuous.

Definition 41.1. If X and Y are two normed vector spaces and if $f\colon \Omega \to Y$ is a function from some open subset Ω of X, a *generalized Newton method* for finding zeros of f consists of

(1) A sequence of families $(A_k(x))$ of linear isomorphisms from X to Y, for all $x \in \Omega$ and all integers $k \geq 0$;

(2) Some starting point $x_0 \in \Omega$;

(3) A sequence (x_k) of points of Ω defined by

$$x_{k+1} = x_k - (A_k(x_\ell))^{-1}(f(x_k)), \quad k \geq 0, \tag{$**$}$$

where for every integer $k \geq 0$, the integer ℓ satisfies the condition

$$0 \leq \ell \leq k.$$

With $\Delta x_k = x_{k+1} - x_k$, Equation $(**)$ is equivalent to solving the equation

$$A_k(x_\ell)(\Delta x_k) = -f(x_k)$$

and setting $x_{k+1} = x_k + \Delta x_k$. The function $A_k(x)$ usually depends on f'.

Definition 41.1 gives us enough flexibility to capture all the situations that we have previously discussed:

	Function	Index
Variant 1:	$A_k(x) = f'(x)$,	$\ell = k$
Variant 2:	$A_k(x) = f'(x)$,	$\ell = \min\{rp, k\}$, if $rp \leq k \leq (r+1)p - 1, r \geq 0$
Variant 3:	$A_k(x) = f'(x)$,	$\ell = 0$
Variant 4:	$A_k(x) = A_0$,	

where A_0 is a linear isomorphism from X to Y. The first case corresponds to Newton's original method and the others to the variants that we just discussed. We could also have $A_k(x) = A_k$, a fixed linear isomorphism independent of $x \in \Omega$.

Example 41.2. Consider the matrix function f given by

$$f(X) = A - X^{-1},$$

with A and X invertible $n \times n$ matrices. If we apply Variant 1 of Newton's method starting with any $n \times n$ matrix X_0, since the derivative of the function g given by $g(X) = X^{-1}$ is $dg_X(Y) = -X^{-1}YX^{-1}$, we have

$$f'_X(Y) = X^{-1}YX^{-1},$$

so

$$(f'_X)^{-1}(Y) = XYX$$

and the Newton step is

$$X_{k+1} = X_k - (f'_{X_k})^{-1}(f(X_k)) = X_k - X_k(A - X_k^{-1})X_k,$$

which yields the sequence (X_k) with

$$X_{k+1} = X_k(2I - AX_k), \quad k \geq 0.$$

In Problem 41.5, it is shown that Newton's method converges to A^{-1} iff the spectral radius of $I - X_0A$ is strictly smaller than 1, that is, $\rho(I - X_0A) < 1$.

The following theorem inspired by the *Newton–Kantorovich theorem* gives sufficient conditions that guarantee that the sequence (x_k) constructed by a generalized Newton method converges to a zero of f close to x_0. Although quite technical, these conditions are not very surprising.

Theorem 41.1. *Let X be a Banach space, let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$, and assume that there are constants $r, M, \beta > 0$ such that if we let*

$$B = \{x \in X \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k\geq 0} \sup_{x\in B} \left\|A_k^{-1}(x)\right\|_{\mathcal{L}(Y;X)} \leq M,$$

(2) $\beta < 1$ and

$$\sup_{k\geq 0} \sup_{x,x'\in B} \|f'(x) - A_k(x')\|_{\mathcal{L}(X;Y)} \leq \frac{\beta}{M}$$

(3)

$$\|f(x_0)\| \leq \frac{r}{M}(1-\beta).$$

Then the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(f(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B and converges to a zero a of f, which is the only zero of f in B. Furthermore, the convergence is geometric, which means that

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1-\beta}\beta^k.$$

Proof. We follow Ciarlet [41] (Theorem 7.5.1, Section 7.5). The proof has three steps.

Step 1. We establish the following inequalities for all $k \geq 1$.

$$\|x_k - x_{k-1}\| \leq M\,\|f(x_{k-1})\| \tag{a}$$
$$\|x_k - x_0\| \leq r \quad (x_k \in B) \tag{b}$$
$$\|f(x_k)\| \leq \frac{\beta}{M}\,\|x_k - x_{k-1}\|\,. \tag{c}$$

We proceed by induction on k, starting with the base case $k = 1$. Since

$$x_1 = x_0 - A_0^{-1}(x_0)(f(x_0)),$$

we have $x_1 - x_0 = -A_0^{-1}(x_0)(f(x_0))$, so by (1) and (3) and since $0 < \beta < 1$, we have

$$\|x_1 - x_0\| \leq M \|f(x_0)\| \leq r(1-\beta) \leq r,$$

establishing (a) and (b) for $k = 1$. We also have $f(x_0) = -A_0(x_0)(x_1 - x_0)$, so $-f(x_0) - A_0(x_0)(x_1 - x_0) = 0$ and thus

$$f(x_1) = f(x_1) - f(x_0) - A_0(x_0)(x_1 - x_0).$$

By the mean value theorem (Proposition 39.12) applied to the function $x \mapsto f(x) - A_0(x_0)(x)$, by (2), we get

$$\|f(x_1)\| \leq \sup_{x \in B} \|f'(x) - A_0(x_0)\| \|x_1 - x_0\| \leq \frac{\beta}{M} \|x_1 - x_0\|,$$

which is (c) for $k = 1$. We now establish the induction step.

Since by definition

$$x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1})), \quad 0 \leq \ell \leq k-1,$$

by (1) and the fact that by the induction hypothesis for (b), $x_\ell \in B$, we get

$$\|x_k - x_{k-1}\| \leq M \|f(x_{k-1})\|,$$

which proves (a) for k. As a consequence, since by the induction hypothesis for (c),

$$\|f(x_{k-1})\| \leq \frac{\beta}{M} \|x_{k-1} - x_{k-2}\|,$$

we get

$$\|x_k - x_{k-1}\| \leq M \|f(x_{k-1})\| \leq \beta \|x_{k-1} - x_{k-2}\|, \tag{$*_1$}$$

and by repeating this step,

$$\|x_k - x_{k-1}\| \leq \beta^{k-1} \|x_1 - x_0\|. \tag{$*_2$}$$

Using ($*_2$) and (3), we obtain

$$\begin{aligned} \|x_k - x_0\| &\leq \sum_{j=1}^{k} \|x_j - x_{j-1}\| \leq \left(\sum_{j=1}^{k} \beta^{j-1}\right) \|x_1 - x_0\| \\ &\leq \frac{\|x_1 - x_0\|}{1-\beta} \leq \frac{M}{1-\beta} \|f(x_0)\| \leq r, \end{aligned}$$

which proves that $x_k \in B$, which is (b) for k.

Since

$$x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1}))$$

we also have $f(x_{k-1}) = -A_{k-1}(x_\ell)(x_k - x_{k-1})$, so we have

$$f(x_k) = f(x_k) - f(x_{k-1}) - A_{k-1}(x_\ell)(x_k - x_{k-1}),$$

and as in the base case, applying the mean value theorem (Proposition 39.12) to the function $x \mapsto f(x) - A_{k-1}(x_\ell)(x)$, by (2), we obtain

$$\|f(x_k)\| \leq \sup_{x \in B} \|f'(x) - A_{k-1}(x_\ell)\| \, \|x_k - x_{k-1}\| \leq \frac{\beta}{M} \|x_k - x_{k-1}\|,$$

proving (c) for k.

Step 2. Prove that f has a zero in B.

To do this we prove that (x_k) is a Cauchy sequence. This is because, using $(*_2)$, we have

$$\begin{aligned}\|x_{k+j} - x_k\| &\leq \sum_{i=0}^{j-1} \|x_{k+i+1} - x_{k+i}\| \leq \beta^k \left(\sum_{i=0}^{j-1} \beta^i\right) \|x_1 - x_0\| \\ &\leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,\end{aligned}$$

for all $k \geq 0$ and all $j \geq 0$, proving that (x_k) is a Cauchy sequence. Since B is a closed ball in a complete normed vector space, B is complete and the Cauchy sequence (x_k) converges to a limit $a \in B$. Since f is continuous on Ω (because it is differentiable), by (c) we obtain

$$\|f(a)\| = \lim_{k \mapsto \infty} \|f(x_k)\| \leq \frac{\beta}{M} \lim_{k \mapsto \infty} \|x_k - x_{k-1}\| = 0,$$

which yields $f(a) = 0$.

Since

$$\|x_{k+j} - x_k\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,$$

if we let j tend to infinity, we obtain the inequality

$$\|x_k - a\| = \|a - x_k\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,$$

which is the last statement of the theorem.

Step 3. Prove that f has a unique zero in B.

Suppose $f(a) = f(b) = 0$ with $a, b \in B$. Since $A_0^{-1}(x_0)(A_0(x_0)(b - a)) = b - a$, we have

$$b - a = -A_0^{-1}(x_0)(f(b) - f(a) - A_0(x_0)(b - a)),$$

which by (1) and (2) and the mean value theorem implies that

$$\|b - a\| \leq \left\|A_0^{-1}(x_0)\right\| \sup_{x \in B} \|f'(x) - A_0(x_0)\| \, \|b - a\| \leq \beta \|b - a\|.$$

Since $0 < \beta < 1$, the inequality $\|b - a\| \leq \beta \|b - a\|$ is only possible if $a = b$. $\square$

It should be observed that the conditions of Theorem 41.1 are typically quite stringent. It can be shown that Theorem 41.1 applies to the function f of Example 41.1 given by $f(x) = x^2 - \alpha$ with $\alpha > 0$, for any $x_0 > 0$ such that

$$\frac{6}{7}\alpha \leq x_0^2 \leq \frac{6}{5}\alpha,$$

with $\beta = 2/5$, $r = (1/6)x_0$, $M = 3/(5x_0)$. Such values of x_0 are quite close to $\sqrt{\alpha}$.

If we assume that we already know that some element $a \in \Omega$ is a zero of f, the next theorem gives sufficient conditions for a special version of a generalized Newton method to converge. For this special method the linear isomorphisms $A_k(x)$ are independent of $x \in \Omega$.

Theorem 41.2. *Let X be a Banach space and let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$. If $a \in \Omega$ is a point such that $f(a) = 0$, if $f'(a)$ is a linear isomorphism, and if there is some λ with $0 < \lambda < 1/2$ such that*

$$\sup_{k\geq 0} \|A_k - f'(a)\|_{\mathcal{L}(X;Y)} \leq \frac{\lambda}{\|(f'(a))^{-1}\|_{\mathcal{L}(Y;X)}},$$

then there is a closed ball B of center a such that for every $x_0 \in B$, the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(f(x_k)), \quad k \geq 0,$$

is entirely contained within B and converges to a, which is the only zero of f in B. Furthermore, the convergence is geometric, which means that

$$\|x_k - a\| \leq \beta^k \|x_0 - a\|,$$

for some $\beta < 1$.

A proof of Theorem 41.2 can be found in Ciarlet [41] (Section 7.5).

For the sake of completeness, we state a version of the Newton–Kantorovich theorem which corresponds to the case where $A_k(x) = f'(x)$. In this instance, a stronger result can be obtained especially regarding upper bounds, and we state a version due to Gragg and Tapia which appears in Problem 7.5-4 of Ciarlet [41].

Theorem 41.3. *(Newton–Kantorovich) Let X be a Banach space, and let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$. Assume that there exist three positive constants λ, μ, ν and a point $x_0 \in \Omega$ such that*

$$0 < \lambda\mu\nu \leq \frac{1}{2},$$

and if we let

$$\begin{aligned}\rho^- &= \frac{1-\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ \rho^+ &= \frac{1+\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ B &= \{x \in X \mid \|x - x_0\| < \rho^-\}\\ \Omega^+ &= \{x \in \Omega \mid \|x - x_0\| < \rho^+\},\end{aligned}$$

then $\overline{B} \subseteq \Omega$, $f'(x_0)$ *is an isomorphism of* $\mathcal{L}(X;Y)$, *and*

$$\begin{aligned}\left\|(f'(x_0))^{-1}\right\| &\leq \mu,\\ \left\|(f'(x_0))^{-1}f(x_0)\right\| &\leq \lambda,\\ \sup_{x,y\in\Omega^+} \|f'(x) - f'(y)\| &\leq \nu\,\|x-y\|\,.\end{aligned}$$

Then $f'(x)$ *is isomorphism of* $\mathcal{L}(X;Y)$ *for all* $x \in B$, *and the sequence defined by*

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)), \quad k \geq 0$$

is entirely contained within the ball B *and converges to a zero* a *of* f *which is the only zero of* f *in* Ω^+. *Finally, if we write* $\theta = \rho^-/\rho^+$, *then we have the following bounds:*

$$\begin{aligned}\|x_k - a\| &\leq \frac{2\sqrt{1-2\lambda\mu\nu}}{\lambda\mu\nu}\frac{\theta^{2k}}{1-\theta^{2k}}\|x_1 - x_0\| && \text{if } \lambda\mu\nu < \frac{1}{2}\\ \|x_k - a\| &\leq \frac{\|x_1 - x_0\|}{2^{k-1}} && \text{if } \lambda\mu\nu = \frac{1}{2},\end{aligned}$$

and

$$\frac{2\,\|x_{k+1} - x_k\|}{1+\sqrt{(1+4\theta^{2k}(1+\theta^{2k})^{-2})}} \leq \|x_k - a\| \leq \theta^{2k-1}\,\|x_k - x_{k-1}\|\,.$$

We can now specialize Theorems 41.1 and 41.2 to the search of zeros of the derivative $J' \colon \Omega \to E'$, of a function $J \colon \Omega \to \mathbb{R}$, with $\Omega \subseteq E$. The second derivative J'' of J is a continuous bilinear form $J'' \colon E \times E \to \mathbb{R}$, but is is convenient to view it as a linear map in $\mathcal{L}(E, E')$; the continuous linear form $J''(u)$ is given by $J''(u)(v) = J''(u,v)$. In our next theorem, which follows immediately from Theorem 41.1, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$.

Theorem 41.4. *Let* E *be a Banach space, let* $J \colon \Omega \to \mathbb{R}$ *be twice differentiable on the open subset* $\Omega \subseteq E$, *and assume that there are constants* $r, M, \beta > 0$ *such that if we let*

$$B = \{x \in E \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k\geq 0}\sup_{x\in B}\left\|A_k^{-1}(x)\right\|_{\mathcal{L}(E';E)} \leq M,$$

(2) $\beta < 1$ *and*

$$\sup_{k\geq 0}\sup_{x,x'\in B}\|J''(x) - A_k(x')\|_{\mathcal{L}(E;E')} \leq \frac{\beta}{M}$$

(3)

$$\|J'(x_0)\| \leq \frac{r}{M}(1-\beta).$$

Then the sequence (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(J'(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B *and converges to a zero* a *of* J', *which is the only zero of* J' *in* B. *Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1-\beta}\beta^k.$$

In the next theorem, which follows immediately from Theorem 41.2, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$ that are independent of $x \in \Omega$.

Theorem 41.5. *Let* E *be a Banach space and let* $J\colon \Omega \to \mathbb{R}$ *be twice differentiable on the open subset* $\Omega \subseteq E$. *If* $a \in \Omega$ *is a point such that* $J'(a) = 0$, *if* $J''(a)$ *is a linear isomorphism, and if there is some* λ *with* $0 < \lambda < 1/2$ *such that*

$$\sup_{k\geq 0}\|A_k - J''(a)\|_{\mathcal{L}(E;E')} \leq \frac{\lambda}{\|(J''(a))^{-1}\|_{\mathcal{L}(E';E)}},$$

then there is a closed ball B *of center* a *such that for every* $x_0 \in B$, *the sequence* (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(J'(x_k)), \quad k \geq 0,$$

is entirely contained within B *and converges to* a, *which is the only zero of* J' *in* B. *Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \beta^k \|x_0 - a\|,$$

for some $\beta < 1$.

When $E = \mathbb{R}^n$, the Newton method given by Theorem 41.4 yields an iteration step of the form

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)\nabla J(x_k), \quad 0 \leq \ell \leq k,$$

where $\nabla J(x_k)$ is the gradient of J at x_k (here, we identify E' with $\mathbb{R}^n$). In particular, Newton's original method picks $A_k = J''$, and the iteration step is of the form

$$x_{k+1} = x_k - (\nabla^2 J(x_k))^{-1} \nabla J(x_k), \quad k \geq 0,$$

where $\nabla^2 J(x_k)$ is the Hessian of J at x_k.

Example 41.3. Let us apply Newton's original method to the function J given by $J(x) = \frac{1}{3}x^3 - 4x$. We have $J'(x) = x^2 - 4$ and $J''(x) = 2x$, so the Newton step is given by

$$x_{k+1} = x_k - \frac{x_k^2 - 4}{2x_k} = \frac{1}{2}\left(x_k + \frac{4}{x_k}\right).$$

This is the sequence of Example 41.1 to compute the square root of 4. Starting with any $x_0 > 0$ it converges very quickly to 2.

As remarked in Ciarlet [41] (Section 7.5), generalized Newton methods have a very wide range of applicability. For example, various versions of gradient descent methods can be viewed as instances of Newton method. See Section 49.9 for an example.

Newton's method also plays an important role in convex optimization, in particular, interior-point methods. A variant of Newton's method dealing with equality constraints has been developed. We refer the reader to Boyd and Vandenberghe [29], Chapters 10 and 11, for a comprehensive exposition of these topics.

41.3 Summary

The main concepts and results of this chapter are listed below:

- Newton's method for functions $f \colon \mathbb{R} \to \mathbb{R}$.
- Generalized Newton methods.
- The *Newton-Kantorovich* theorem.

41.4 Problems

Problem 41.1. If $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α.

(1) Prove that if $x_0 > 0$, then $x_k > 0$ and

$$x_{k+1} - \sqrt{\alpha} = \frac{1}{2x_k}(x_k - \sqrt{\alpha})^2$$
$$x_{k+2} - x_{k+1} = \frac{1}{2x_{k+1}}(\alpha - x_{k+1}^2)$$

for all $k \geq 0$. Deduce that Newton's method converges to $\sqrt{\alpha}$ for any $x_0 > 0$.

(2) Prove that if $x_0 < 0$, then Newton's method converges to $-\sqrt{\alpha}$.

Problem 41.2. (1) If $\alpha > 0$ and $f(x) = x^2 - \alpha$, show that the conditions of Theorem 41.1 are satisfied by any $\beta \in (0, 1)$ and any x_0 such that

$$|x_0^2 - \alpha| \leq \frac{4\beta(1-\beta)}{(\beta+2)^2} x_0^2,$$

with

$$r = \frac{\beta}{\beta+2} x_0, \qquad M = \frac{\beta+2}{4x_0}.$$

(2) Prove that the maximum of the function defined on $[0, 1]$ by

$$\beta \mapsto \frac{4\beta(1-\beta)}{(\beta+2)^2}$$

has a maximum for $\beta = 2/5$. For this value of β, check that $r = (1/6)x_0$, $M = 3/(5x_0)$, and

$$\frac{6}{7}\alpha \leq x_0^2 \leq \frac{6}{5}\alpha.$$

Problem 41.3. Consider generalizing Problem 41.1 to the matrix function f given by $f(X) = X^2 - C$, where X and C are two real $n \times n$ matrices with C symmetric positive definite. The first step is to determine for which A does the inverse df_A^{-1} exist. Let g be the function given by $g(X) = X^2$. From Problem 39.1 we know that the derivative at A of the function g is $dg_A(X) = AX + XA$, and obviously $df_A = dg_A$. Thus we are led to figure out when the linear matrix map $X \mapsto AX + XA$ is invertible. This can be done using the Kronecker product.

Given an $m \times n$ matrix $A = (a_{ij})$ and a $p \times q$ matrix $B = (b_{ij})$, the *Kronecker product* (or *tensor product*) $A \otimes B$ of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

It can be shown (and you may use these facts without proof) that $\otimes$ is associative and that

$$\begin{aligned}(A \otimes B)(C \otimes D) &= AC \otimes BD \\ (A \otimes B)^\top &= A^\top \otimes B^\top,\end{aligned}$$

whenever AC and BD are well defined.

Given any $n \times n$ matrix X, let $\mathrm{vec}(X)$ be the vector in $\mathbb{R}^{n^2}$ obtained by concatenating the *rows* of X.

(1) Prove that $AX = Y$ iff

$$(A \otimes I_n)\mathrm{vec}(X) = \mathrm{vec}(Y)$$

and $XA = Y$ iff

$$(I_n \otimes A^\top)\mathrm{vec}(X) = \mathrm{vec}(Y).$$

Deduce that $AX + XA = Y$ iff

$$((A \otimes I_n) + (I_n \otimes A^\top))\mathrm{vec}(X) = \mathrm{vec}(Y).$$

In the case where $n = 2$ and if we write

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

check that

$$A \otimes I_2 + I_2 \otimes A^\top = \begin{pmatrix} 2a & c & b & 0 \\ b & a+d & 0 & b \\ c & 0 & a+d & c \\ 0 & c & b & 2d \end{pmatrix}.$$

The problem is determine when the matrix $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible.

Remark: The equation $AX + XA = Y$ is a special case of the equation $AX + XB = C$ (sometimes written $AX - XB = C$), called the *Sylvester equation*, where A is an $m \times m$ matrix, B is an $n \times n$ matrix, and X, C are $m \times n$ matrices; see Higham [91] (Appendix B).

(2) In the case where $n = 2$, prove that

$$\det(A \otimes I_2 + I_2 \otimes A^\top) = 4(a + d)^2(ad - bc).$$

(3) Let A and B be any two $n \times n$ complex matrices. Use Schur factorizations $A = UT_1U^*$ and $B = VT_2V^*$ (where U and V are unitary and T_1, T_2 are upper-triangular) to prove that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A and $\mu_1, \ldots, \mu_n$ are the eigenvalues of B, then the scalars $\lambda_i\mu_j$ are the eigenvalues of $A \otimes B$, for $i, j = 1, \ldots, n$.

Hint. Check that $U \otimes V$ is unitary and that $T_1 \otimes T_2$ is upper triangular.

(4) Prove that the eigenvalues of $(A \otimes I_n) + (I_n \otimes B)$ are the scalars $\lambda_i + \mu_j$, for $i, j = 1, \ldots, n$. Deduce that the eigenvalues of $(A \otimes I_n) + (I_n \otimes A^\top)$ are $\lambda_i + \lambda_j$, for $i, j = 1, \ldots, n$. Thus $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible iff $\lambda_i + \lambda_j \neq 0$, for $i, j = 1, \ldots, n$. In particular, prove that if A is symmetric positive definite, then so is $(A \otimes I_n) + (I_n \otimes A^\top)$.

Hint. Use (3).

(5) Prove that if A and B are symmetric and $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible, then the unique solution X of the equation $AX + XA = B$ is symmetric.

(6) Starting with a symmetric positive definite matrix X_0, the general step of Newton's method is

$$X_{k+1} = X_k - (f'_{X_k})^{-1}(X_k^2 - C) = X_k - (g'_{X_k})^{-1}(X_k^2 - C),$$

and since g'_{X_k} is linear, this is equivalent to

$$X_{k+1} = X_k - (g'_{X_k})^{-1}(X_k^2) + (g'_{X_k})^{-1}(C).$$

But since X_k is SPD, $(g'_{X_k})^{-1}(X_k^2)$ is the unique solution of

$$X_k Y + Y X_k = X_k^2$$

whose solution is obviously $Y = (1/2)X_k$. Therefore the Newton step is

$$X_{k+1} = X_k - (g'_{X_k})^{-1}(X_k^2) + (g'_{X_k})^{-1}(C) = X_k - \frac{1}{2}X_k + (g'_{X_k})^{-1}(C) = \frac{1}{2}X_k + (g'_{X_k})^{-1}(C),$$

so we have

$$X_{k+1} = \frac{1}{2}X_k + (g'_{X_k})^{-1}(C) = (g'_{X_k})^{-1}(X_k^2 + C).$$

Prove that if X_k and C are symmetric positive definite, then $(g'_{X_k})^{-1}(C)$ is symmetric positive definite, and if C is symmetric positive semidefinite, then $(g'_{X_k})^{-1}(C)$ is symmetric positive semidefinite.

Hint. By (5) the unique solution Z of the equation $X_k Z + Z X_k = C$ (where C is symmetric) is symmetric so it can be diagonalized as $Z = QDQ^\top$ with Q orthogonal and D a real diagonal matrix. Prove that

$$Q^\top X_k Q D + D Q^\top X_k Q = Q^\top C Q,$$

and solve the system using the diagonal elements.

Deduce that if X_k and C are SPD, then X_{k+1} is SPD.

Since $C = P\Sigma P^\top$ is SPD, it has an SPD square root (in fact unique) $C^{1/2} = P\Sigma^{1/2}P^\top$. Prove that

$$X_{k+1} - C^{1/2} = (g'_{X_k})^{-1}(X_k - C^{1/2})^2.$$

Prove that

$$\left\|(g'_{X_k})^{-1}\right\|_2 = \frac{1}{2\left\|X_k\right\|_2}.$$

Since

$$X_{k+1} - X_k = (g'_{X_k})^{-1}(C - X_k^2),$$

deduce that if $X_k \neq C^2$, then $X_k - X_{k+1}$ is SPD.

Open problem: Does Theorem 41.1 apply for some suitable r, M, β?

(7) Prove that if C and X_0 commute, provided that the equation $X_k Z + Z X_k = C$ has a unique solution for all k, then X_k and C commute for all k and Z is given by

$$Z = \frac{1}{2} X_k^{-1} C = \frac{1}{2} C X_k^{-1}.$$

Deduce that

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1} C) = \frac{1}{2}(X_k + C X_k^{-1}).$$

This is the matrix analog of the formula given in Problem 41.1(1).

Prove that if C and X_0 have positive eigenvalues and C and X_0 commute, then X_{k+1} has positive eigenvalues for all $k \geq 0$ and thus the sequence (X_k) is defined.

Hint. Because X_k and C commute, X_k^{-1} and C commute, and obviously X_k and X_k^{-1} commute. By Proposition 31.9, X_k, X_k^{-1}, and C are triangulable in a common basis, so there is some orthogonal matrix P and some upper-triangular matrices T_1, T_2 such that

$$X_k = P T_1 P^\top, \quad X_k^{-1} = P T_1^{-1} P^\top, \quad C = P T_2 P^\top.$$

It follows that

$$X_{k+1} = \frac{1}{2} P (T_1 + T_1^{-1} T_2) P^\top.$$

Also recall that the diagonal entries of an upper-triangular matrix are the eigenvalues of that matrix.

We conjecture that if C has positive eigenvalues, then the Newton sequence converges starting with any X_0 of the form $X_0 = \mu I_n$, with $\mu > 0$.

(8) Implement the above method in `Matlab` (there is a command `kron(A, B)` to form the Kronecker product of A and B). Test your program on diagonalizable and nondiagonalizable matrices, including

$$W = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix},$$

and

$$A_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 0.01 & 0 & 0 \\ -1 & -1 & 100 & 100 \\ -1 & -1 & -100 & 100 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

What happens with

$$C = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad X_0 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

The problem of determining when square roots of matrices exist and procedures for finding them are thoroughly investigated in Higham [91] (Chapter 6).

Problem 41.4. (1) Show that Newton's method applied to the function

$$f(x) = \alpha - \frac{1}{x}$$

with $\alpha \neq 0$ and $x \in \mathbb{R} - \{0\}$ yields the sequence (x_k) with

$$x_{k+1} = x_k(2 - \alpha x_k), \quad k \geq 0.$$

(2) If we let $r_k = 1 - \alpha x_k$, prove that $r_{k+1} = r_k^2$ for all $k \geq 0$. Deduce that Newton's method converges to $1/\alpha$ if $0 < \alpha x_0 < 2$.

Problem 41.5. (1) Show that Newton's method applied to the matrix function

$$f(X) = A - X^{-1},$$

with A and X invertible $n \times n$ matrices and started with any $n \times n$ matrix X_0 yields the sequence (X_k) with

$$X_{k+1} = X_k(2I - AX_k), \quad k \geq 0.$$

(2) If we let $R_k = I - AX_k$, prove that

$$R_{k+1} = I - (I - R_k)(I + R_k) = R_k^2$$

for all $k \geq 0$. Deduce that Newton's method converges to A^{-1} iff the spectral radius of $I - AX_0$ is strictly smaller than 1, that is, $\rho(I - AX_0) < 1$.

(3) Assume that A is symmetric positive definite and let $X_0 = \mu I$. Prove that the condition $\rho(I - AX_0) < 1$ is equivalent to

$$0 < \mu < \frac{2}{\rho(A)}.$$

(4) Write a `Matlab` program implementing Newton's method specified in (1). Test your program with the $n \times n$ matrix

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix},$$

and with $X_0 = \mu I_n$, for various values of n, including $n = 8, 10, 20$, and various values of μ such that $0 < \mu \leq 1/2$. Find some $\mu > 1/2$ causing divergence.

Problem 41.6. A method for computing the nth root $x^{1/n}$ of a positive real number $x \in \mathbb{R}$, with $n \in \mathbb{N}$ a positive integer $n \geq 2$, proceeds as follows: define the sequence (x_k), where x_0 is any chosen positive real, and

$$x_{k+1} = \frac{1}{n}\left((n-1)x_k + \frac{x}{x_k^{n-1}}\right), \quad k \geq 0.$$

(1) Implement the above method in `Matlab` and test it for various input values of x, x_0, and of $n \geq 2$, by running successively your program for $m = 2, 3, \ldots, 100$ iterations. Have your program plot the points (i, x_i) to watch how quickly the sequence converges.

Experiment with various choices of x_0. One of these choices should be $x_0 = x$. Compare your answers with the result of applying the of `Matlab` function $x \mapsto x^{1/n}$.

In some case, when x_0 is small, the number of iterations has to be at least 1000. Exhibit this behavior.

Problem 41.7. Refer to Problem 41.6 for the definition of the sequence (x_k).

(1) Define the *relative error* ϵ_k as

$$\epsilon_k = \frac{x_k}{x^{1/n}} - 1, \quad k \geq 0.$$

Prove that

$$\epsilon_{k+1} = \frac{x^{(1-1/n)}}{n x_k^{n-1}}\left(\frac{(n-1)x_k^n}{x} - \frac{n x_k^{n-1}}{x^{(1-1/n)}} + 1\right),$$

and then that

$$\epsilon_{k+1} = \frac{1}{n(\epsilon_k+1)^{n-1}}\left(\epsilon_k(\epsilon_k+1)^{n-2}((n-1)\epsilon_k + (n-2)) + 1 - (\epsilon_k+1)^{n-2}\right),$$

for all $k \geq 0$.

(2) Since

$$\epsilon_k + 1 = \frac{x_k}{x^{1/n}},$$

and since we assumed $x_0, x > 0$, we have $\epsilon_0 + 1 > 0$. We would like to prove that

$$\epsilon_k \geq 0, \quad \text{for all} \quad k \geq 1.$$

For this consider the variations of the function f given by

$$f(u) = (n-1)u^n - n x^{1/n} u^{n-1} + x,$$

for $u \in \mathbb{R}$.

Use the above to prove that $f(u) \geq 0$ for all $u \geq 0$. Conclude that

$$\epsilon_k \geq 0, \quad \text{for all} \quad k \geq 1.$$

(3) Prove that if $n = 2$, then

$$0 \leq \epsilon_{k+1} = \frac{\epsilon_k^2}{2(\epsilon_k + 1)}, \quad \text{for all} \quad k \geq 0,$$

else if $n \geq 3$, then

$$0 \leq \epsilon_{k+1} \leq \frac{(n-1)}{n}\epsilon_k, \quad \text{for all} \quad k \geq 1.$$

Prove that the sequence (x_k) converges to $x^{1/n}$ for every initial value $x_0 > 0$.

(4) When $n = 2$, we saw in Problem 41.7(3) that

$$0 \leq \epsilon_{k+1} = \frac{\epsilon_k^2}{2(\epsilon_k + 1)}, \quad \text{for all} \quad k \geq 0.$$

For $n = 3$, prove that

$$\epsilon_{k+1} = \frac{2\epsilon_k^2(3/2 + \epsilon_k)}{3(\epsilon_k + 1)^2}, \quad \text{for all} \quad k \geq 0,$$

and for $n = 4$, prove that

$$\epsilon_{k+1} = \frac{3\epsilon_k^2}{4(\epsilon_k + 1)^3}\left(2 + (8/3)\epsilon_k + \epsilon_k^2\right), \quad \text{for all} \quad k \geq 0.$$

Let μ_3 and μ_4 be the functions given by

$$\begin{aligned} \mu_3(a) &= \frac{3}{2} + a \\ \mu_4(a) &= 2 + \frac{8}{3}a + a^2, \end{aligned}$$

so that if $n = 3$, then

$$\epsilon_{k+1} = \frac{2\epsilon_k^2\mu_3(\epsilon_k)}{3(\epsilon_k + 1)^2}, \quad \text{for all} \quad k \geq 0,$$

and if $n = 4$, then

$$\epsilon_{k+1} = \frac{3\epsilon_k^2\mu_4(\epsilon_k)}{4(\epsilon_k + 1)^3}, \quad \text{for all} \quad k \geq 0.$$

Prove that

$$a\mu_3(a) \leq (a+1)^2 - 1, \quad \text{for all} \quad a \geq 0,$$

and

$$a\mu_4(a) \leq (a+1)^3 - 1, \quad \text{for all} \quad a \geq 0.$$

Let $\eta_{3,k} = \mu_3(\epsilon_1)\epsilon_k$ when $n = 3$, and $\eta_{4,k} = \mu_4(\epsilon_1)\epsilon_k$ when $n = 4$. Prove that

$$\eta_{3,k+1} \leq \frac{2}{3}\eta_{3,k}^2, \quad \text{for all} \quad k \geq 1,$$

and

$$\eta_{4,k+1} \leq \frac{3}{4}\eta_{4,k}^2, \quad \text{for all} \quad k \geq 1.$$

Deduce from the above that the rate of convergence of $\eta_{i,k}$ is very fast, for $i = 3, 4$ (and $k \geq 1$).

Remark: If we let $\mu_2(a) = a$ for all a and $\eta_{2,k} = \epsilon_k$, we then proved that

$$\eta_{2,k+1} \leq \frac{1}{2}\eta_{2,k}^2, \quad \text{for all} \quad k \geq 1.$$

Problem 41.8. This is a continuation of Problem 41.7.

(1) Prove that for all $n \geq 2$, we have

$$\epsilon_{k+1} = \left(\frac{n-1}{n}\right) \frac{\epsilon_k^2 \mu_n(\epsilon_k)}{(\epsilon_k + 1)^{n-1}}, \quad \text{for all} \quad k \geq 0,$$

where μ_n is given by

$$\mu_n(a) = \frac{1}{2}n + \sum_{j=1}^{n-4} \frac{1}{n-1}\left((n-1)\binom{n-2}{j} + (n-2)\binom{n-2}{j+1} - \binom{n-2}{j+2}\right)a^j + \frac{n(n-2)}{n-1}a^{n-3} + a^{n-2}.$$

Furthermore, prove that μ_n can be expressed as

$$\mu_n(a) = \frac{1}{2}n + \frac{n(n-2)}{3}a + \sum_{j=2}^{n-4} \frac{(j+1)n}{(j+2)(n-1)}\binom{n-1}{j+1}a^j + \frac{n(n-2)}{n-1}a^{n-3} + a^{n-2}.$$

(2) Prove that for every j, with $1 \leq j \leq n-1$, the coefficient of a^j in $a\mu_n(a)$ is less than or equal to the coefficient of a^j in $(a+1)^{n-1} - 1$, and thus

$$a\mu_n(a) \leq (a+1)^{n-1} - 1, \quad \text{for all} \quad a \geq 0,$$

with strict inequality if $n \geq 3$. In fact, prove that if $n \geq 3$, then for every j, with $3 \leq j \leq n-2$, the coefficient of a^j in $a\mu_n(a)$ is strictly less than the coefficient of a^j in $(a+1)^{n-1} - 1$, and if $n \geq 4$, this also holds for $j = 2$.

Let $\eta_{n,k} = \mu_n(\epsilon_1)\epsilon_k$ $(n \geq 2)$. Prove that

$$\eta_{n,k+1} \leq \left(\frac{n-1}{n}\right)\eta_{n,k}^2, \quad \text{for all} \quad k \geq 1.$$

Chapter 42

Quadratic Optimization Problems

In this chapter we consider two classes of quadratic optimization problems that appear frequently in engineering and in computer science (especially in computer vision):

1. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over all $x \in \mathbb{R}^n$, or subject to linear or affine constraints.

2. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over the unit sphere.

In both cases, A is a symmetric matrix. We also seek necessary and sufficient conditions for Q to have a global minimum.

42.1 Quadratic Optimization: The Positive Definite Case

Many problems in physics and engineering can be stated as the minimization of some energy function, with or without constraints. Indeed, it is a fundamental principle of mechanics that nature acts so as to minimize energy. Furthermore, if a physical system is in a stable state of equilibrium, then the energy in that state should be minimal. For example, a small ball placed on top of a sphere is in an unstable equilibrium position. A small motion causes the ball to roll down. On the other hand, a ball placed inside and at the bottom of a sphere is in a stable equilibrium position because the potential energy is minimal.

The simplest kind of energy function is a quadratic function. Such functions can be conveniently defined in the form

$$Q(x) = x^\top Ax - x^\top b,$$

where A is a symmetric $n \times n$ matrix and x, b, are vectors in $\mathbb{R}^n$, viewed as column vectors. Actually, for reasons that will be clear shortly, it is preferable to put a factor $\frac{1}{2}$ in front of the quadratic term, so that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b.$$

The question is, under what conditions (on A) does $Q(x)$ have a global minimum, preferably unique?

We give a complete answer to the above question in two stages:

1. In this section we show that if A is symmetric positive definite, then $Q(x)$ has a unique global minimum precisely when
$$Ax = b.$$

2. In Section 42.2 we give necessary and sufficient conditions in the general case, in terms of the pseudo-inverse of A.

We begin with the matrix version of Definition 22.2.

Definition 42.1. A symmetric *positive definite matrix* is a matrix whose eigenvalues are strictly positive, and a symmetric *positive semidefinite matrix* is a matrix whose eigenvalues are nonnegative.

Equivalent criteria are given in the following proposition.

Proposition 42.1. *Given any Euclidean space E of dimension n, the following properties hold:*

(1) Every self-adjoint linear map $f\colon E \to E$ is positive definite iff

$$\langle f(x), x\rangle > 0$$

for all $x \in E$ with $x \neq 0$.

(2) Every self-adjoint linear map $f\colon E \to E$ is positive semidefinite iff

$$\langle f(x), x\rangle \geq 0$$

for all $x \in E$.

Proof. (1) First assume that f is positive definite. Recall that every self-adjoint linear map has an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. With respect to this basis, for every $x = x_1e_1 + \cdots + x_ne_n \neq 0$, we have

$$\langle f(x), x\rangle = \Big\langle f\Big(\sum_{i=1}^n x_ie_i\Big), \sum_{i=1}^n x_ie_i\Big\rangle = \Big\langle \sum_{i=1}^n \lambda_ix_ie_i, \sum_{i=1}^n x_ie_i\Big\rangle = \sum_{i=1}^n \lambda_ix_i^2,$$

which is strictly positive, since $\lambda_i > 0$ for $i = 1, \ldots, n$, and $x_i^2 > 0$ for some i, since $x \neq 0$.

Conversely, assume that

$$\langle f(x), x\rangle > 0$$

for all $x \neq 0$. Then for $x = e_i$, we get

$$\langle f(e_i), e_i\rangle = \langle \lambda_i e_i, e_i\rangle = \lambda_i,$$

and thus $\lambda_i > 0$ for all $i = 1, \ldots, n$.

(2) As in (1), we have

$$\langle f(x), x\rangle = \sum_{i=1}^{n} \lambda_i x_i^2,$$

and since $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because f is positive semidefinite, we have $\langle f(x), x\rangle \geq 0$, as claimed. The converse is as in (1) except that we get only $\lambda_i \geq 0$ since $\langle f(e_i), e_i\rangle \geq 0$. $\square$

Some special notation is customary (especially in the field of convex optimization) to express that a symmetric matrix is positive definite or positive semidefinite.

Definition 42.2. Given any $n \times n$ symmetric matrix A we write $A \succeq 0$ if A is positive semidefinite and we write $A \succ 0$ if A is positive definite.

Remark: It should be noted that we can define the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite. It is easy to check that this relation is actually a partial order on matrices, called the *positive semidefinite cone ordering*; for details, see Boyd and Vandenberghe [29], Section 2.4.

If A is symmetric positive definite, it is easily checked that A^{-1} is also symmetric positive definite. Also, if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is injective), then $A^\top CA$ is symmetric positive definite.

We can now prove that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a global minimum when A is symmetric positive definite.

Proposition 42.2. *Given a quadratic function*

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

if A is symmetric positive definite, then $Q(x)$ has a unique global minimum for the solution $x_0 = A^{-1}b$ of the linear system $Ax = b$. The minimum value of $Q(x)$ is

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

Proof. Since A is positive definite, it is invertible since its eigenvalues are all strictly positive. Let $x_0 = A^{-1}b$, and compute $Q(y) - Q(x_0)$ for any $y \in \mathbb{R}^n$. Since $Ax_0 = b$, we get

$$\begin{aligned} Q(y) - Q(x_0) &= \frac{1}{2}y^\top Ay - y^\top b - \frac{1}{2}x_0^\top Ax_0 + x_0^\top b \\ &= \frac{1}{2}y^\top Ay - y^\top Ax_0 + \frac{1}{2}x_0^\top Ax_0 \\ &= \frac{1}{2}(y - x_0)^\top A(y - x_0). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative, and thus

$$Q(y) \geq Q(x_0)$$

for all $y \in \mathbb{R}^n$, which proves that $x_0 = A^{-1}b$ is a global minimum of $Q(x)$. A simple computation yields

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

$\square$

Remarks:

(1) The quadratic function $Q(x)$ is also given by

$$Q(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

but the definition using $x^\top b$ is more convenient for the proof of Proposition 42.2.

(2) If $Q(x)$ contains a constant term $c \in \mathbb{R}$, so that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b + c,$$

the proof of Proposition 42.2 still shows that $Q(x)$ has a unique global minimum for $x = A^{-1}b$, but the minimal value is

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b + c.$$

Thus when the energy function $Q(x)$ of a system is given by a quadratic function

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

where A is symmetric positive definite, finding the global minimum of $Q(x)$ is equivalent to solving the linear system $Ax = b$. Sometimes, it is useful to recast a linear problem $Ax = b$ as a variational problem (finding the minimum of some energy function). However, very often, a minimization problem comes with extra constraints that must be satisfied for all admissible solutions.

Example 42.1. For instance, we may want to minimize the quadratic function

$$Q(x_1, x_2) = \frac{1}{2}\big(x_1^2 + x_2^2\big)$$

subject to the constraint

$$2x_1 - x_2 = 5.$$

The solution for which $Q(x_1, x_2)$ is minimum is no longer $(x_1, x_2) = (0, 0)$, but instead, $(x_1, x_2) = (2, -1)$, as will be shown later.

Geometrically, the graph of the function defined by $z = Q(x_1, x_2)$ in $\mathbb{R}^3$ is a paraboloid of revolution P with axis of revolution Oz. The constraint

$$2x_1 - x_2 = 5$$

corresponds to the vertical plane H parallel to the z-axis and containing the line of equation $2x_1 - x_2 = 5$ in the xy-plane. Thus, as illustrated by Figure 42.1, the constrained minimum of Q is located on the parabola that is the intersection of the paraboloid P with the plane H.

A nice way to solve constrained minimization problems of the above kind is to use the method of *Lagrange multipliers* discussed in Section 40.1. But first let us define precisely what kind of minimization problems we intend to solve.

Definition 42.3. The *quadratic constrained minimization problem* consists in minimizing a quadratic function

$$Q(x) = \frac{1}{2}x^\top A^{-1}x - b^\top x$$

subject to the linear constraints

$$B^\top x = f,$$

where A^{-1} is an $m \times m$ symmetric positive definite matrix, B is an $m \times n$ matrix of rank n (so that $m \geq n$), and where $b, x \in \mathbb{R}^m$ (viewed as column vectors), and $f \in \mathbb{R}^n$ (viewed as a column vector).

The reason for using A^{-1} instead of A is that the constrained minimization problem has an interpretation as a set of equilibrium equations in which the matrix that arises naturally is A (see Strang [169]). Since A and A^{-1} are both symmetric positive definite, this doesn't make any difference, but it seems preferable to stick to Strang's notation.

In Example 42.1 we have $m = 2$, $n = 1$,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad B = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad f = 5.$$

As explained in Section 40.1, the method of Lagrange multipliers consists in incorporating the n constraints $B^\top x = f$ into the quadratic function $Q(x)$, by introducing extra variables

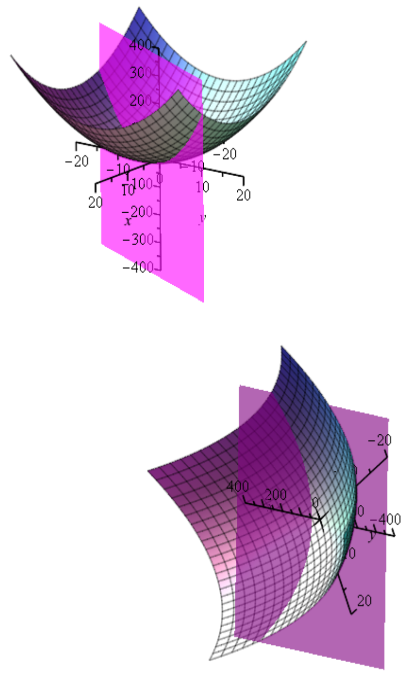


Figure 42.1: Two views of the constrained optimization problem $Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ subject to the constraint $2x_1 - x_2 = 5$. The minimum $(x_1, x_2) = (2, -1)$ is the the vertex of the parabolic curve formed the intersection of the magenta planar constraint with the bowl shaped surface.

$\lambda = (\lambda_1, \ldots, \lambda_n)$ called *Lagrange multipliers*, one for each constraint. We form the *Lagrangian*

$$L(x, \lambda) = Q(x) + \lambda^\top (B^\top x - f) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f.$$

We know from Theorem 40.2 that a necessary condition for our constrained optimization problem to have a solution is that $\nabla L(x, \lambda) = 0$. Since

$$\begin{aligned}\frac{\partial L}{\partial x}(x, \lambda) &= A^{-1}x - (b - B\lambda)\\ \frac{\partial L}{\partial \lambda}(x, \lambda) &= B^\top x - f,\end{aligned}$$

we obtain the system of linear equations

$$\begin{aligned}A^{-1}x + B\lambda &= b,\\ B^\top x &= f,\end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

We shall prove in Proposition 42.3 below that our constrained minimization problem has a unique solution actually given by the above system.

Note that the matrix of this system is symmetric. We solve it as follows. Eliminating x from the first equation

$$A^{-1}x + B\lambda = b,$$

we get

$$x = A(b - B\lambda),$$

and substituting into the second equation, we get

$$B^\top A(b - B\lambda) = f,$$

that is,

$$B^\top AB\lambda = B^\top Ab - f.$$

However, by a previous remark, since A is symmetric positive definite and the columns of B are linearly independent, $B^\top AB$ is symmetric positive definite, and thus invertible. Thus we obtain the solution

$$\lambda = (B^\top AB)^{-1}(B^\top Ab - f), \qquad x = A(b - B\lambda).$$

Note that this way of solving the system requires solving for the Lagrange multipliers first.

Letting $e = b - B\lambda$, we also note that the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

is equivalent to the system

$$\begin{aligned} e &= b - B\lambda, \\ x &= Ae, \\ B^\top x &= f. \end{aligned}$$

The latter system is called the *equilibrium equations* by Strang [169]. Indeed, Strang shows that the equilibrium equations of many physical systems can be put in the above form. This includes spring-mass systems, electrical networks and trusses, which are structures built from elastic bars. In each case, x, e, b, A, λ, f, and $K = B^\top AB$ have a physical interpretation. The matrix $K = B^\top AB$ is usually called the *stiffness matrix*. Again, the reader is referred to Strang [169].

In order to prove that our constrained minimization problem has a unique solution, we proceed to prove that the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of another function $-G(\lambda)$. We get $G(\lambda)$ by minimizing the Lagrangian $L(x, \lambda)$ treated as a function of x alone. The function $-G(\lambda)$ is the *dual function* of the Lagrangian $L(x, \lambda)$. Here we are encountering a special case of the notion of dual function defined in Section 50.7.

Since A^{-1} is symmetric positive definite and

$$L(x, \lambda) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f,$$

by Proposition 42.2 the global minimum (with respect to x) of $L(x, \lambda)$ is obtained for the solution x of

$$A^{-1} x = b - B\lambda,$$

that is, when

$$x = A(b - B\lambda),$$

and the minimum of $L(x, \lambda)$ is

$$\min_x L(x, \lambda) = -\frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) - \lambda^\top f.$$

Letting

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we will show in Proposition 42.3 that the solution of the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of $-G(\lambda)$. This is a special case of the duality discussed in Section 50.7.

Of course, since we minimized $L(x, \lambda)$ with respect to x, we have

$$L(x, \lambda) \geq -G(\lambda)$$

for all x and all λ. However, when the constraint $B^\top x = f$ holds, $L(x, \lambda) = Q(x)$, and thus for any admissible x, which means that $B^\top x = f$, we have

$$\min_x Q(x) \geq \max_\lambda -G(\lambda).$$

In order to prove that the unique minimum of the constrained problem $Q(x)$ subject to $B^\top x = f$ is the unique maximum of $-G(\lambda)$, we compute $Q(x) + G(\lambda)$.

Proposition 42.3. *The quadratic constrained minimization problem of Definition 42.3 has a unique solution (x, λ) given by the system*

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Furthermore, the component λ of the above solution is the unique value for which $-G(\lambda)$ is maximum.

Proof. As we suggested earlier, let us compute $Q(x) + G(\lambda)$, assuming that the constraint $B^\top x = f$ holds. Eliminating f, since $b^\top x = x^\top b$ and $\lambda^\top B^\top x = x^\top B\lambda$, we get

$$\begin{aligned} Q(x) + G(\lambda) &= \frac{1}{2} x^\top A^{-1} x - b^\top x + \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f \\ &= \frac{1}{2}(A^{-1}x + B\lambda - b)^\top A(A^{-1}x + B\lambda - b). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative. In fact, it is null iff

$$A^{-1}x + B\lambda - b = 0,$$

that is,

$$A^{-1}x + B\lambda = b.$$

But then the unique constrained minimum of $Q(x)$ subject to $B^\top x = f$ is equal to the unique maximum of $-G(\lambda)$ exactly when $B^\top x = f$ and $A^{-1}x + B\lambda = b$, which proves the proposition. $\square$

We can confirm that the maximum of $-G(\lambda)$, or equivalently the minimum of

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

corresponds to value of λ obtained by solving the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Indeed, since

$$G(\lambda) = \frac{1}{2}\lambda^\top B^\top AB\lambda - \lambda^\top B^\top Ab + \lambda^\top f + \frac{1}{2} b^\top b,$$

and $B^\top AB$ is symmetric positive definite, by Proposition 42.2, the global minimum of $G(\lambda)$ is obtained when

$$B^\top AB\lambda - B^\top Ab + f = 0,$$

that is, $\lambda = (B^\top AB)^{-1}(B^\top Ab - f)$, as we found earlier.

Remarks:

(1) There is a form of duality going on in this situation. The constrained minimization of $Q(x)$ subject to $B^\top x = f$ is called the *primal problem*, and the unconstrained maximization of $-G(\lambda)$ is called the *dual problem*. Duality is the fact stated slightly loosely as

$$\min_x Q(x) = \max_\lambda -G(\lambda).$$

A general treatment of duality in constrained minimization problems is given in Section 50.7.

Recalling that $e = b - B\lambda$, since

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we can also write

$$G(\lambda) = \frac{1}{2}e^\top Ae + \lambda^\top f.$$

This expression often represents the total potential energy of a system. Again, the optimal solution is the one that minimizes the potential energy (and thus maximizes $-G(\lambda)$).

(2) It is immediately verified that the equations of Proposition 42.3 are equivalent to the equations stating that the partial derivatives of the Lagrangian $L(x, \lambda)$ are null:

$$\begin{aligned} \frac{\partial L}{\partial x_i} &= 0, \quad i = 1, \ldots, m, \\ \frac{\partial L}{\partial \lambda_j} &= 0, \quad j = 1, \ldots, n. \end{aligned}$$

Thus, the constrained minimum of $Q(x)$ subject to $B^\top x = f$ is an extremum of the Lagrangian $L(x, \lambda)$. As we showed in Proposition 42.3, this extremum corresponds to simultaneously minimizing $L(x, \lambda)$ with respect to x and maximizing $L(x, \lambda)$ with respect to λ. Geometrically, such a point is a *saddle point* for $L(x, \lambda)$. Saddle points are discussed in Section 50.7.

(3) The Lagrange multipliers sometimes have a natural physical meaning. For example, in the spring-mass system they correspond to node displacements. In some general sense, Lagrange multipliers are correction terms needed to satisfy equilibrium equations and the price paid for the constraints. For more details, see Strang [169].

Going back to the constrained minimization of $Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ subject to

$$2x_1 - x_2 = 5,$$

the Lagrangian is

$$L(x_1, x_2, \lambda) = \frac{1}{2}\big(x_1^2 + x_2^2\big) + \lambda(2x_1 - x_2 - 5),$$

and the equations stating that the Lagrangian has a saddle point are

$$\begin{aligned} x_1 + 2\lambda &= 0, \\ x_2 - \lambda &= 0, \\ 2x_1 - x_2 - 5 &= 0. \end{aligned}$$

We obtain the solution $(x_1, x_2, \lambda) = (2, -1, -1)$.

The use of Lagrange multipliers in optimization and variational problems is discussed extensively in Chapter 50.

Least squares methods and Lagrange multipliers are used to tackle many problems in computer graphics and computer vision; see Trucco and Verri [178], Metaxas [124], Jain, Katsuri, and Schunck [100], Faugeras [59], and Foley, van Dam, Feiner, and Hughes [63].

42.2 Quadratic Optimization: The General Case

In this section we complete the study initiated in Section 42.1 and give necessary and sufficient conditions for the quadratic function $\frac{1}{2}x^\top Ax - x^\top b$ to have a global minimum. We begin with the following simple fact:

Proposition 42.4. *If A is an invertible symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$, in which case this optimal value is obtained for a unique value of x, namely $x^* = A^{-1}b$, and with

$$f(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

Proof. Observe that

$$\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) = \frac{1}{2}x^\top Ax - x^\top b + \frac{1}{2}b^\top A^{-1}b.$$

Thus,

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b = \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b.$$

If A has some negative eigenvalue, say $-\lambda$ (with $\lambda > 0$), if we pick any eigenvector u of A associated with λ, then for any $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, if we let $x = \alpha u + A^{-1}b$, then since $Au = -\lambda u$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b \\ &= \frac{1}{2}\alpha u^\top A\alpha u - \frac{1}{2}b^\top A^{-1}b \\ &= -\frac{1}{2}\alpha^2\lambda \left\|u\right\|_2^2 - \frac{1}{2}b^\top A^{-1}b, \end{aligned}$$

and since α can be made as large as we want and $\lambda > 0$, we see that f has no minimum. Consequently, in order for f to have a minimum, we must have $A \succeq 0$. If $A \succeq 0$, since A is invertible, it is positive definite, so $(x - A^{-1}b)^\top A(x - A^{-1}b) > 0$ iff $x - A^{-1}b \neq 0$, and it is clear that the minimum value of f is achieved when $x - A^{-1}b = 0$, that is, $x = A^{-1}b$. $\square$

Let us now consider the case of an arbitrary symmetric matrix A.

Proposition 42.5. *If A is an $n \times n$ symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$ and $(I - AA^+)b = 0$, in which case this minimum value is

$$p^* = -\frac{1}{2}b^\top A^+ b.$$

Furthermore, if A is diagonalized as $A = U^\top \Sigma U$ (with U orthogonal), then the optimal value is achieved by all $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, where r is the rank of A.

Proof. The case that A is invertible is taken care of by Proposition 42.4, so we may assume that A is singular. If A has rank $r < n$, then we can diagonalize A as

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U,$$

where U is an orthogonal matrix and where Σ_r is an $r \times r$ diagonal invertible matrix. Then we have

$$\begin{aligned} f(x) &= \frac{1}{2}x^\top U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - x^\top U^\top Ub \\ &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub. \end{aligned}$$

If we write

$$Ux = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

with $y, c \in \mathbb{R}^r$ and $z, d \in \mathbb{R}^{n-r}$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub \\ &= \frac{1}{2}(y^\top \; z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} - (y^\top \; z^\top) \begin{pmatrix} c \\ d \end{pmatrix} \\ &= \frac{1}{2}y^\top \Sigma_r y - y^\top c - z^\top d. \end{aligned}$$

For $y = 0$, we get

$$f(x) = -z^\top d,$$

so if $d \neq 0$, the function f has no minimum. Therefore, if f has a minimum, then $d = 0$. However, $d = 0$ means that

$$Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

and we know from Proposition 23.5 that b is in the range of A (here, U is $V^\top$), which is equivalent to $(I - AA^+)b = 0$. If $d = 0$, then

$$f(x) = \frac{1}{2} y^\top \Sigma_r y - y^\top c.$$

Consider the function $g\colon \mathbb{R}^r \to \mathbb{R}$ given by

$$g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c, \quad y \in \mathbb{R}^r.$$

Since

$$\begin{pmatrix} y \\ z \end{pmatrix} = U^\top x$$

and $U^\top$ is invertible (with inverse U), when x ranges over $\mathbb{R}^n$, y ranges over the whole of $\mathbb{R}^r$, and since $f(x) = g(y)$, the function f has a minimum iff g has a minimum. Since Σ_r is invertible, by Proposition 42.4, the function g has a minimum iff $\Sigma_r \succeq 0$, which is equivalent to $A \succeq 0$.

Therefore, we have proven that if f has a minimum, then $(I - AA^+)b = 0$ and $A \succeq 0$. Conversely, if $(I - AA^+)b = 0$, then

$$\begin{aligned} \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U \right) b &= \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U \right) b \\ &= U^\top \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} Ub = 0, \end{aligned}$$

which implies that if

$$Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

then $d = 0$, so as above

$$f(x) = g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

and because $A \succeq 0$, we also have $\Sigma_r \succeq 0$, so g and f have a minimum.

When the above conditions hold, since

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U$$

is positive semidefinite, the pseudo-inverse A^+ of A is given by

$$A^+ = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U,$$

and since

$$f(x) = g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

by Proposition 42.4 the minimum of g is achieved iff $y^* = \Sigma_r^{-1} c$. Since $f(x)$ is independent of z, we can choose $z = 0$, and since $d = 0$, for x^* given by

$$Ux^* = \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

we deduce that

$$x^* = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Ub = A^+ b, \tag{$*$}$$

and the minimum value of f is

$$f(x^*) = \frac{1}{2}(A^+ b)^\top A A^+ b - b^\top A^+ b = \frac{1}{2} b^\top A^+ A A^+ b - b^\top A^+ b = -\frac{1}{2} b^\top A^+ b,$$

since A^+ is symmetric and $A^+ A A^+ = A^+$. For any $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix}, \quad z \in \mathbb{R}^{n-r},$$

since

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ z \end{pmatrix},$$

and since $f(x)$ is independent of z (because $f(x) = g(y)$), we have

$$f(x) = f(x^*) = -\frac{1}{2} b^\top A^+ b. \qquad \square$$

The problem of minimizing the function

$$f(x) = \frac{1}{2} x^\top A x - x^\top b$$

in the case where we add either linear constraints of the form $C^\top x = 0$ or affine constraints of the form $C^\top x = t$ (where $t \in \mathbb{R}^m$ and $t \neq 0$) where C is an $n \times m$ matrix can be reduced to the unconstrained case using a QR-decomposition of C. Let us show how to do this for linear constraints of the form $C^\top x = 0$.

If we use a QR decomposition of C, by permuting the columns of C to make sure that the first r columns of C are linearly independent (where $r = \text{rank}(C)$), we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an $n \times n$ orthogonal matrix, R is an $r \times r$ invertible upper triangular matrix, S is an $r \times (m - r)$ matrix, and Π is a permutation matrix (C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$C^\top x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}(y^\top \; z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} + (y^\top \; z^\top) Qb \\ \text{subject to} \quad & y = 0, \; y \in \mathbb{R}^r, \; z \in \mathbb{R}^{n-r}. \end{aligned}$$

Thus, the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} is an $r \times r$ matrix and G_{22} is an $(n - r) \times (n - r)$ matrix and

$$Qb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad b_1 \in \mathbb{R}^r, \; b_2 \in \mathbb{R}^{n-r},$$

our problem becomes

$$\text{minimize} \; \frac{1}{2} z^\top G_{22} z + z^\top b_2, \quad z \in \mathbb{R}^{n-r},$$

the problem solved in Proposition 42.5.

Constraints of the form $C^\top x = t$ (where $t \neq 0$) can be handled in a similar fashion. In this case, we may assume that C is an $n \times m$ matrix with full rank (so that $m \leq n$) and $t \in \mathbb{R}^m$. Then we use a QR-decomposition of the form

$$C = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix. If we write

$$x = P \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n-m}$, the equation $C^\top x = t$ becomes

$$(R^\top \; 0) P^\top x = t,$$

that is,

$$(R^\top \; 0) \begin{pmatrix} y \\ z \end{pmatrix} = t,$$

which yields

$$R^\top y = t.$$

Since R is invertible, we get $y = (R^\top)^{-1} t$, and then it is easy to see that our original problem reduces to an unconstrained problem in terms of the matrix $P^\top AP$; the details are left as an exercise.

42.3 Maximizing a Quadratic Function on the Unit Sphere

In this section we discuss various quadratic optimization problems mostly arising from computer vision (image segmentation and contour grouping). These problems can be reduced to the following basic optimization problem: given an $n \times n$ real symmetric matrix A

$$\begin{aligned} &\text{maximize} && x^\top Ax \\ &\text{subject to} && x^\top x = 1, \; x \in \mathbb{R}^n. \end{aligned}$$

In view of Proposition 23.10, the maximum value of $x^\top Ax$ on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1. Similarly, the minimum value of $x^\top Ax$ on the unit sphere is equal to the smallest eigenvalue λ_n of the matrix A, and it is achieved for any unit eigenvector u_n associated with λ_n.

A variant of the above problem often encountered in computer vision consists in minimizing $x^\top Ax$ on the ellipsoid given by an equation of the form

$$x^\top Bx = 1,$$

where B is a symmetric positive definite matrix. Since B is positive definite, it can be diagonalized as

$$B = QDQ^\top,$$

where Q is an orthogonal matrix and D is a diagonal matrix,

$$D = \mathrm{diag}(d_1, \ldots, d_n),$$

with $d_i > 0$, for $i = 1, \ldots, n$. If we define the matrices $B^{1/2}$ and $B^{-1/2}$ by

$$B^{1/2} = Q \,\mathrm{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right) Q^\top$$

and

$$B^{-1/2} = Q \,\mathrm{diag}\left(1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n}\right) Q^\top,$$

it is clear that these matrices are symmetric, that $B^{-1/2}BB^{-1/2} = I$, and that $B^{1/2}$ and $B^{-1/2}$ are mutual inverses. Then if we make the change of variable

$$x = B^{-1/2}y,$$

the equation $x^\top Bx = 1$ becomes $y^\top y = 1$, and the optimization problem

$$\begin{aligned} &\text{minimize} && x^\top Ax \\ &\text{subject to} && x^\top Bx = 1,\ x \in \mathbb{R}^n, \end{aligned}$$

is equivalent to the problem

$$\begin{aligned} &\text{minimize} && y^\top B^{-1/2}AB^{-1/2}y \\ &\text{subject to} && y^\top y = 1,\ y \in \mathbb{R}^n, \end{aligned}$$

where $y = B^{1/2}x$ and $B^{-1/2}AB^{-1/2}$ are symmetric.

The complex version of our basic optimization problem in which A is a Hermitian matrix also arises in computer vision. Namely, given an $n \times n$ complex Hermitian matrix A,

$$\begin{aligned} &\text{maximize} && x^*Ax \\ &\text{subject to} && x^*x = 1,\ x \in \mathbb{C}^n. \end{aligned}$$

Again by Proposition 23.10, the maximum value of x^*Ax on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1.

Remark: It is worth pointing out that if A is a *skew-Hermitian* matrix, that is, if $A^* = -A$, then x^*Ax is *pure imaginary or zero*.

Indeed, since $z = x^*Ax$ is a scalar, we have $z^* = \overline{z}$ (the conjugate of z), so we have

$$\overline{x^*Ax} = (x^*Ax)^* = x^*A^*x = -x^*Ax,$$

so $\overline{x^*Ax} + x^*Ax = 2\mathrm{Re}(x^*Ax) = 0$, which means that x^*Ax is pure imaginary or zero.

In particular, if A is a real matrix and if A is *skew-symmetric*, then

$$x^\top A x = 0.$$

Thus, for any real matrix (symmetric or not),

$$x^\top A x = x^\top H(A) x,$$

where $H(A) = (A + A^\top)/2$, the symmetric part of A.

There are situations in which it is necessary to add linear constraints to the problem of maximizing a quadratic function on the sphere. This problem was completely solved by Golub [78] (1973). The problem is the following: given an $n \times n$ real symmetric matrix A and an $n \times p$ matrix C,

$$\begin{aligned} &\text{minimize} && x^\top A x \\ &\text{subject to} && x^\top x = 1,\; C^\top x = 0,\; x \in \mathbb{R}^n. \end{aligned}$$

As in Section 42.2, Golub shows that the linear constraint $C^\top x = 0$ can be eliminated as follows: if we use a QR decomposition of C, by permuting the columns, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an orthogonal $n \times n$ matrix, R is an $r \times r$ invertible upper triangular matrix, and S is an $r \times (p - r)$ matrix (assuming C has rank r). If we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$\Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Q x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{aligned} &\text{minimize} && (y^\top\; z^\top) Q A Q^\top \begin{pmatrix} y \\ z \end{pmatrix} \\ &\text{subject to} && z^\top z = 1,\; z \in \mathbb{R}^{n-r}, \\ & && y = 0,\; y \in \mathbb{R}^r. \end{aligned}$$

Thus the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22} \end{pmatrix},$$

our problem becomes

$$\begin{aligned} &\text{minimize} && z^\top G_{22} z \\ &\text{subject to} && z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \end{aligned}$$

a standard eigenvalue problem.

Remark: There is a way of finding the eigenvalues of G_{22} which does not require the QR-factorization of C. Observe that if we let

$$J = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

then

$$JQAQ^\top J = \begin{pmatrix} 0 & 0 \\ 0 & G_{22} \end{pmatrix},$$

and if we set

$$P = Q^\top JQ,$$

then

$$PAP = Q^\top JQAQ^\top JQ.$$

Now, $Q^\top JQAQ^\top JQ$ and $JQAQ^\top J$ have the same eigenvalues, so PAP and $JQAQ^\top J$ also have the same eigenvalues. It follows that the solutions of our optimization problem are among the eigenvalues of $K = PAP$, and at least r of those are 0. Using the fact that CC^+ is the projection onto the range of C, where C^+ is the pseudo-inverse of C, it can also be shown that

$$P = I - CC^+,$$

the projection onto the kernel of $C^\top$. So P can be computed directly in terms of C. In particular, when $n \geq p$ and C has full rank (the columns of C are linearly independent), then we know that $C^+ = (C^\top C)^{-1} C^\top$ and

$$P = I - C(C^\top C)^{-1} C^\top.$$

This fact is used by Cour and Shi [42] and implicitly by Yu and Shi [192].

The problem of adding affine constraints of the form $N^\top x = t$, where $t \neq 0$, also comes up in practice. At first glance, this problem may not seem harder than the linear problem in which $t = 0$, but it is. This problem was extensively studied in a paper by Gander, Golub, and von Matt [75] (1989).

Gander, Golub, and von Matt consider the following problem: Given an $(n+m)\times(n+m)$ real symmetric matrix A (with $n > 0$), an $(n+m)\times m$ matrix N with full rank, and a nonzero vector $t \in \mathbb{R}^m$ with $\left\|(N^\top)^+ t\right\| < 1$ (where $(N^\top)^+$ denotes the pseudo-inverse of $N^\top$),

$$\begin{aligned} &\text{minimize} && x^\top Ax \\ &\text{subject to} && x^\top x = 1,\ N^\top x = t,\ x \in \mathbb{R}^{n+m}. \end{aligned}$$

The condition $\left\|(N^\top)^+ t\right\| < 1$ ensures that the problem has a solution and is not trivial. The authors begin by proving that the affine constraint $N^\top x = t$ can be eliminated. One way to do so is to use a QR decomposition of N. If

$$N = P\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $(n+m)\times(n+m)$ matrix and R is an $m \times m$ invertible upper triangular matrix, then if we observe that

$$\begin{aligned} x^\top Ax &= x^\top PP^\top APP^\top x, \\ N^\top x &= (R^\top\ 0)P^\top x = t, \\ x^\top x &= x^\top PP^\top x = 1, \end{aligned}$$

and if we write

$$P^\top AP = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix},$$

where B is an $m \times m$ symmetric matrix, C is an $n \times n$ symmetric matrix, Γ is an $m \times n$ matrix, and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

with $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$, we then get

$$\begin{aligned} x^\top Ax &= y^\top By + 2z^\top \Gamma y + z^\top Cz, \\ R^\top y &= t, \\ y^\top y + z^\top z &= 1. \end{aligned}$$

Thus

$$y = (R^\top)^{-1} t,$$

and if we write

$$s^2 = 1 - y^\top y > 0$$

and

$$b = \Gamma y,$$

we get the simplified problem

$$\begin{array}{ll} \text{minimize} & z^\top C z + 2z^\top b \\ \text{subject to} & z^\top z = s^2,\ z \in \mathbb{R}^m. \end{array}$$

Unfortunately, if $b \neq 0$, Proposition 23.10 is no longer applicable. It is still possible to find the minimum of the function $z^\top C z + 2z^\top b$ using Lagrange multipliers, but such a solution is too involved to be presented here. Interested readers will find a thorough discussion in Gander, Golub, and von Matt [75].

42.4 Summary

The main concepts and results of this chapter are listed below:

- Quadratic optimization problems; *quadratic functions*.
- Symmetric *positive definite* and *positive semidefinite* matrices.
- The *positive semidefinite cone ordering*.
- Existence of a global minimum when A is symmetric positive definite.
- Constrained quadratic optimization problems.
- *Lagrange multipliers*; *Lagrangian*.
- *Primal* and *dual* problems.
- Quadratic optimization problems: the case of a symmetric invertible matrix A.
- Quadratic optimization problems: the general case of a symmetric matrix A.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $C^\top x = t$, with $t \neq 0$.
- Maximizing a quadratic function over the unit sphere.
- Maximizing a quadratic function over an ellipsoid.
- Maximizing a Hermitian quadratic form.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $N^\top x = t$, with $t \neq 0$.

42.5 Problems

Problem 42.1. Prove that the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite is indeed a partial order.

Problem 42.2. (1) Prove that if A is symmetric positive definite, then so is A^{-1}.

(2) Prove that if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is injective), then $A^\top CA$ is symmetric positive definite.

Problem 42.3. Find the minimum of the function

$$Q(x_1, x_2) = \frac{1}{2}(2x_1^2 + x_2^2)$$

subject to the constraint

$$x_1 - x_2 = 3.$$

Problem 42.4. Consider the problem of minimizing the function

$$f(x) = \frac{1}{2} x^\top Ax - x^\top b$$

in the case where we add an affine constraint of the form $C^\top x = t$, with $t \in \mathbb{R}^m$ and $t \neq 0$, and where C is an $n \times m$ matrix of rank $m \leq n$. As in Section 42.2, use a QR-decomposition

$$C = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix, and write

$$x = P \begin{pmatrix} y \\ z \end{pmatrix},$$

to deduce that

$$R^\top y = t.$$

Give the details of the reduction of this constrained minimization problem to an unconstrained minimization problem involving the matrix $P^\top AP$.

Problem 42.5. Find the maximum and the minimum of the function

$$Q(x, y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

on the unit circle $x^2 + y^2 = 1$.

Chapter 43

Schur Complements and Applications

Schur complements arise naturally in the process of inverting block matrices of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and in characterizing when symmetric versions of these matrices are positive definite or positive semidefinite. These characterizations come up in various quadratic optimization problems; see Boyd and Vandenberghe [29], especially Appendix B. In the most general case, pseudo-inverses are also needed.

In this chapter we introduce Schur complements and describe several interesting ways in which they are used. Along the way we provide some details and proofs of some results from Appendix A.5 (especially Section A.5.5) of Boyd and Vandenberghe [29].

43.1 Schur Complements

Let M be an $n \times n$ matrix written as a 2×2 block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix). We can try to solve the linear system

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix},$$

that is,

$$\begin{aligned} Ax + By &= c, \\ Cx + Dy &= d, \end{aligned}$$

by mimicking Gaussian elimination. If we assume that D is invertible, then we first solve for y, getting

$$y = D^{-1}(d - Cx),$$

and after substituting this expression for y in the first equation, we get

$$Ax + B(D^{-1}(d - Cx)) = c,$$

that is,

$$(A - BD^{-1}C)x = c - BD^{-1}d.$$

If the matrix $A - BD^{-1}C$ is invertible, then we obtain the solution to our system

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}(c - BD^{-1}d), \\ y &= D^{-1}(d - C(A - BD^{-1}C)^{-1}(c - BD^{-1}d)). \end{aligned}$$

If A is invertible, then by eliminating x first using the first equation, we obtain analogous formulas involving the matrix $D - CA^{-1}B$. The above formulas suggest that the matrices $A - BD^{-1}C$ and $D - CA^{-1}B$ play a special role and suggest the following definition:

Definition 43.1. Given any $n \times n$ block matrix of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix), if D is invertible, then the matrix $A - BD^{-1}C$ is called the *Schur complement* of D in M. If A is invertible, then the matrix $D - CA^{-1}B$ is called the *Schur complement* of A in M.

The above equations written as

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}c - (A - BD^{-1}C)^{-1}BD^{-1}d, \\ y &= -D^{-1}C(A - BD^{-1}C)^{-1}c \\ &\quad + (D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1})d, \end{aligned}$$

yield a formula for the inverse of M in terms of the Schur complement of D in M, namely

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

A moment of reflection reveals that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix},$$

and then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}.$$

By taking inverses, we obtain the following result.

Proposition 43.1. *If the matrix D is invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}.$$

The above expression can be checked directly and has the advantage of requiring only the invertibility of D.

Remark: If A is invertible, then we can use the Schur complement $D - CA^{-1}B$ of A to obtain the following factorization of M:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

If $D - CA^{-1}B$ is invertible, we can invert all three matrices above, and we get another formula for the inverse of M in terms of $(D - CA^{-1}B)$, namely,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, by comparing the two expressions for M^{-1}, we get the (non-obvious) formula

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Using this formula, we obtain another expression for the inverse of M involving the Schur complements of A and D (see Horn and Johnson [95]):

Proposition 43.2. *If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If we set $D = I$ and change B to $-B$, we get

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I - CA^{-1}B)^{-1}CA^{-1},$$

a formula known as the *matrix inversion lemma* (see Boyd and Vandenberghe [29], Appendix C.4, especially C.4.3).

43.2 Symmetric Positive Definite Matrices and Schur Complements

If we assume that our block matrix M is symmetric, so that A, D are symmetric and $C = B^\top$, then we see by Proposition 43.1 that M is expressed as

$$M = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}B^\top & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix}^\top,$$

which shows that M is similar to a block diagonal matrix (obviously, the Schur complement, $A - BD^{-1}B^\top$, is symmetric). As a consequence, we have the following version of "Schur's trick" to check whether $M \succ 0$ for a symmetric matrix.

Proposition 43.3. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if C is invertible, then the following properties hold:

(1) $M \succ 0$ iff $C \succ 0$ and $A - BC^{-1}B^\top \succ 0$.

(2) If $C \succ 0$, then $M \succeq 0$ iff $A - BC^{-1}B^\top \succeq 0$.

Proof. (1) Since C is invertible, we have

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^{-1}B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^\top. \quad (*)$$

Observe that

$$\begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

so $(*)$ yields

$$\begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix}^\top = \begin{pmatrix} A - BC^{-1}B^\top & 0 \\ 0 & C \end{pmatrix},$$

and we know that for any symmetric matrix T, here $T = M$, and any invertible matrix N, here

$$N = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

the matrix T is positive definite ($T \succ 0$) iff $NTN^\top$ (which is obviously symmetric) is positive definite ($NTN^\top \succ 0$). But a block diagonal matrix is positive definite iff each diagonal block is positive definite, which concludes the proof.

(2) This is because for any symmetric matrix T and any invertible matrix N, we have $T \succeq 0$ iff $NTN^\top \succeq 0$. ∎

Another version of Proposition 43.3 using the Schur complement of A instead of the Schur complement of C also holds. The proof uses the factorization of M using the Schur complement of A (see Section 43.1).

Proposition 43.4. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if A is invertible then the following properties hold:

(1) $M \succ 0$ iff $A \succ 0$ and $C - B^\top A^{-1} B \succ 0$.

(2) If $A \succ 0$, then $M \succeq 0$ iff $C - B^\top A^{-1} B \succeq 0$.

Here is an illustration of Proposition 43.4(2). Consider the nonlinear quadratic constraint

$$(Ax + b)^\top (Ax + b) \leq c^\top x + d,$$

were $A \in \mathrm{M}_n(\mathbb{R}), x, b, c \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Since obviously $I = I_n$ is invertible and $I \succ 0$, we have

$$\begin{pmatrix} I & Ax + b \\ (Ax + b)^\top & c^\top x + d \end{pmatrix} \succeq 0$$

iff $c^\top x + d - (Ax + b)^\top (Ax + b) \succeq 0$ iff $(Ax + b)^\top (Ax + b) \leq c^\top x + d$, since the matrix (a scalar) $c^\top x + d - (Ax + b)^\top (Ax + b)$ is the Schur complement of I in the above matrix.

The trick of using Schur complements to convert nonlinear inequality constraints into linear constraints on symmetric matrices involving the semidefinite ordering $\succeq$ is used extensively to convert nonlinear problems into semidefinite programs; see Boyd and Vandenberghe [29].

When C is singular (or A is singular), it is still possible to characterize when a symmetric matrix M as above is positive semidefinite, but this requires using a version of the Schur complement involving the pseudo-inverse of C, namely $A - BC^+ B^\top$ (or the Schur complement, $C - B^\top A^+ B$, of A). We use the criterion of Proposition 42.5, which tells us when a quadratic function of the form $\frac{1}{2} x^\top P x - x^\top b$ has a minimum and what this optimum value is (where P is a symmetric matrix).

43.3 Symmetric Positive Semidefinite Matrices and Schur Complements

We now return to our original problem, characterizing when a symmetric matrix

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

is positive semidefinite. Thus, we want to know when the function

$$f(x,y) = (x^\top \;\; y^\top) \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^\top A x + 2x^\top B y + y^\top C y$$

has a minimum with respect to both x and y. If we hold y constant, Proposition 42.5 implies that $f(x,y)$ has a minimum iff $A \succeq 0$ and $(I - AA^+)By = 0$, and then the minimum value is

$$f(x^*, y) = -y^\top B^\top A^+ B y + y^\top C y = y^\top (C - B^\top A^+ B) y.$$

Since we want $f(x,y)$ to be uniformly bounded from below for all x, y, we must have $(I - AA^+)B = 0$. Now $f(x^*, y)$ has a minimum iff $C - B^\top A^+ B \succeq 0$. Therefore, we have established that $f(x,y)$ has a minimum over all x, y iff

$$A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$$

Similar reasoning applies if we first minimize with respect to y and then with respect to x, but this time, the Schur complement $A - BC^+B^\top$ of C is involved. Putting all these facts together, we get our main result:

Theorem 43.5. *Given any symmetric matrix*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

the following conditions are equivalent:

(1) $M \succeq 0$ (M is positive semidefinite).

(2) $A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$

(3) $C \succeq 0, \quad (I - CC^+)B^\top = 0, \quad A - BC^+B^\top \succeq 0.$

If $M \succeq 0$ as in Theorem 43.5, then it is easy to check that we have the following factorizations (using the fact that $A^+AA^+ = A^+$ and $C^+CC^+ = C^+$):

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^+ \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^+B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & 0 \\ C^+B^\top & I \end{pmatrix}$$

and

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^\top A^+ & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & C - B^\top A^+ B \end{pmatrix} \begin{pmatrix} I & A^+B \\ 0 & I \end{pmatrix}.$$

43.4 Summary

The main concepts and results of this chapter are listed below:

- Schur complements.
- The matrix inversion lemma.
- Symmetric positive definite matrices and Schur complements.
- Symmetric positive semidefinite matrices and Schur complements.

43.5 Problems

Problem 43.1. Prove that maximizing the function $g(\lambda)$ given by

$$g(\lambda) = c_0 + \lambda c_1 - (b_0 + \lambda b_1)^\top (A_0 + \lambda A_1)^+ (b_0 + \lambda b_1),$$

subject to

$$A_0 + \lambda A_1 \succeq 0, \quad b_0 + \lambda b_1 \in \mathrm{range}(A_0 + \lambda A_1),$$

with A_0, A_1 some $n \times n$ symmetric positive semidefinite matrices, $b_0, b_1 \in \mathbb{R}^n$, and $c_0, c_1 \in \mathbb{R}$, is equivalent to maximizing γ subject to the constraints

$$\begin{aligned} &\lambda \geq 0 \\ &\begin{pmatrix} A_0 + \lambda A_1 & b_0 + \lambda b_1 \\ (b_0 + \lambda b_1)^\top & c_0 + \lambda c_1 - \gamma \end{pmatrix} \succeq 0. \end{aligned}$$

Problem 43.2. Let $a_1, \ldots, a_m$ be m vectors in $\mathbb{R}^n$ and assume that they span $\mathbb{R}^n$.

(1) Prove that the matrix

$$\sum_{k=1}^m a_k a_k^\top$$

is symmetric positive definite.

(2) Define the matrix X by

$$X = \left(\sum_{k=1}^m a_k a_k^\top \right)^{-1}.$$

Prove that

$$\begin{pmatrix} \sum_{k=1}^m a_k a_k^\top & a_i \\ a_i^\top & 1 \end{pmatrix} \succeq 0, \quad i = 1, \ldots, m.$$

Deduce that

$$a_i^\top X a_i \leq 1, \quad 1 \leq i \leq m.$$

Problem 43.3. Consider the function g of Example 39.11 defined by

$$g(a, b, c) = \log(ac - b^2),$$

where $ac - b^2 > 0$. We found that the Hessian matrix of g is given by

$$Hg(a, b, c) = \frac{1}{(ac - b^2)^2} \begin{pmatrix} -c^2 & 2bc & -b^2 \\ 2bc & -2(b^2 + ac) & 2ab \\ -b^2 & 2ab & -a^2 \end{pmatrix}.$$

Use the Schur complement (of a^2) to prove that the matrix $-Hg(a, b, c)$ is symmetric positive definite if $ac - b^2 > 0$ and $a, c > 0$.

Part VII

Linear Optimization

Chapter 44

Convex Sets, Cones, $\mathcal{H}$-Polyhedra

44.1 What is Linear Programming?

What is *linear programming*? At first glance, one might think that this is some style of computer programming. After all, there is imperative programming, functional programming, object-oriented programming, *etc.* The term linear programming is somewhat misleading, because it really refers to a method for *planning* with linear constraints, or more accurately, an *optimization method* where both the objective function and the constraints are linear.[1]

Linear programming was created in the late 1940's, one of the key players being George Dantzing, who invented the simplex algorithm. Kantorovitch also did some pioneering work on linear programming as early as 1939. The term *linear programming* has a military connotation because in the early 1950's it was used as a synonym for plans or schedules for training troops, logistical supply, resource allocation, *etc.* Unfortunately the term linear programming is well established and we are stuck with it.

Interestingly, even though originally most applications of linear programming were in the field of economics and industrial engineering, linear programming has become an important tool in theoretical computer science and in the theory of algorithms. Indeed, linear programming is often an effective tool for designing approximation algorithms to solve hard problems (typically NP-hard problems). Linear programming is also the "baby version" of convex programming, a very effective methodology which has received much attention in recent years.

Our goal is to present the mathematical underpinnings of linear programming, in particular the existence of an optimal solution if a linear program is feasible and bounded, and the duality theorem in linear programming, one of the deepest results in this field. The duality theorem in linear programming also has significant algorithmic implications but we do not discuss this here. We present the simplex algorithm, the dual simplex algorithm, and the primal dual algorithm. We also describe the tableau formalism for running the simplex

[1] Again, we witness another unfortunate abuse of terminology; the constraints are in fact *affine*.

algorithm and its variants. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

However, we do not discuss other methods such as the ellipsoid method or interior points methods. For these more algorithmic issues, we refer the reader to standard texts on linear programming. In our opinion, one of the clearest (and among the most concise!) is Matousek and Gardner [123]; Chvatal [40] and Schrijver [148] are classics. Papadimitriou and Steiglitz [134] offers a very crisp presentation in the broader context of combinatorial optimization, and Bertsimas and Tsitsiklis [21] and Vanderbei [181] are very complete.

Linear programming has to do with maximizing a linear cost function $c_1x_1 + \cdots + c_nx_n$ with respect to m "linear" inequalities of the form

$$a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i.$$

These constraints can be put together into an $m \times n$ matrix $A = (a_{ij})$, and written more concisely as

$$Ax \leq b.$$

For technical reasons that will appear clearer later on, it is often preferable to add the nonnegativity constaints $x_i \geq 0$ for $i = 1, \ldots, n$. We write $x \geq 0$. It is easy to show that every linear program is equivalent to another one satisfying the constraints $x \geq 0$, at the expense of adding new variables that are also constrained to be nonnegative. Let $\mathcal{P}(A, b)$ be the set of *feasible solutions* of our linear program given by

$$\mathcal{P}(A, b) = \{x \in \mathbb{R}^n \mid Ax \leq b,\ x \geq 0\}.$$

Then there are two basic questions:

(1) Is $\mathcal{P}(A, b)$ nonempty, that is, does our linear program have a chance to have a solution?

(2) Does the objective function $c_1x_1 + \cdots + c_nx_n$ have a maximum value on $\mathcal{P}(A, b)$?

The answer to both questions can be **no**. But if $\mathcal{P}(A, b)$ is nonempty and if the objective function is bounded above (on $\mathcal{P}(A, b)$), then it can be shown that the maximum of $c_1x_1 + \cdots + c_nx_n$ is achieved by some $x \in \mathcal{P}(A, b)$. Such a solution is called an *optimal solution*. Perhaps surprisingly, this result is not so easy to prove (unless one has the simplex method at his disposal). We will prove this result in full detail (see Proposition 45.1).

The reason why linear constraints are so important is that the domain of potential optimal solutions $\mathcal{P}(A, b)$ is *convex*. In fact, $\mathcal{P}(A, b)$ is a convex polyhedron which is the intersection of half-spaces cut out by affine hyperplanes. The objective function being linear is convex, and this is also a crucial fact. Thus, we are led to study convex sets, in particular those that arise from solutions of inequalities defined by affine forms, but also convex cones.

We give a brief introduction to these topics. As a reward, we provide several criteria for testing whether a system of inequalities

$$Ax \leq b,\ x \geq 0$$

has a solution or not in terms of versions of the *Farkas lemma* (see Proposition 50.3 and Proposition 47.4). Then we give a complete proof of the strong duality theorem for linear programming (see Theorem 47.7). We also discuss the complementary slackness conditions and show that they can be exploited to design an algorithm for solving a linear program that uses both the primal problem and its dual. This algorithm known as the *primal dual algorithm*, although not used much nowadays, has been the source of inspiration for a whole class of approximation algorithms also known as primal dual algorithms.

We hope that this chapter and the next three will be a motivation for learning more about linear programming, convex optimization, but also convex geometry. The "bible" in convex optimization is Boyd and Vandenberghe [29], and one of the best sources for convex geometry is Ziegler [195]. This is a rather advanced text, so the reader may want to begin with Gallier [73].

44.2 Affine Subsets, Convex Sets, Affine Hyperplanes, Half-Spaces

We view $\mathbb{R}^n$ as consisting of *column vectors* ($n \times 1$ matrices). As usual, row vectors represent *linear forms*, that is linear maps $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, in the sense that the row vector y (a $1 \times n$ matrix) represents the linear form φ if $\varphi(x) = yx$ for all $x \in \mathbb{R}^n$. We denote the space of linear forms (row vectors) by $(\mathbb{R}^n)^*$.

Recall that a *linear combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are *arbitrary* scalars in $\mathbb{R}$. Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all linear combinations of the vectors in S is the smallest (linear) subspace containing S called the *linear span* of S, and denoted $\mathrm{span}(S)$. A *linear subspace* of $\mathbb{R}^n$ is any nonempty subset of $\mathbb{R}^n$ closed under linear combinations.

Definition 44.1. An *affine combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are scalars in $\mathbb{R}$ *satisfying the condition*

$$\lambda_1 + \cdots + \lambda_m = 1.$$

Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all affine combinations of the vectors in S is the smallest affine subspace containing S called the *affine hull* of S and denoted $\mathrm{aff}(S)$.

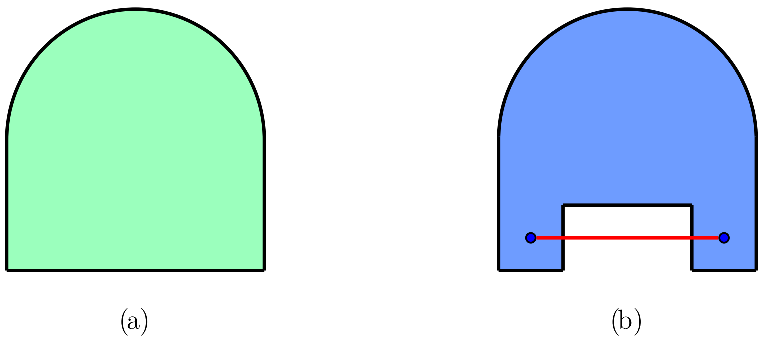


Figure 44.1: (a) A convex set; (b) A nonconvex set

Definition 44.2. An *affine subspace* A of $\mathbb{R}^n$ is any subset of $\mathbb{R}^n$ closed under affine combinations.

If A is a nonempty affine subspace of $\mathbb{R}^n$, then it can be shown that $V_A = \{a - b \mid a, b \in A\}$ is a linear subspace of $\mathbb{R}^n$ and that

$$A = a + V_A = \{a + v \mid v \in V_A\}$$

for *any* $a \in A$; see Gallier [72] (Section 2.5).

Definition 44.3. Given an affine subspace A, the linear space $V_A = \{a - b \mid a, b \in A\}$ is called the *direction of* A. The *dimension* of the nonempty affine subspace A is the dimension of its direction V_A.

Definition 44.4. *Convex combinations* are affine combinations $\lambda_1 x_1 + \cdots + \lambda_m x_m$ satisfying the extra condition that $\lambda_i \geq 0$ for $i = 1, \ldots, m$.

A convex set is defined as follows.

Definition 44.5. A subset V of $\mathbb{R}^n$ is *convex* if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$). Given any two points a, b, the notation $[a, b]$ is often used to denote the line segment between a and b, that is,

$$[a, b] = \{c \in \mathbb{R}^n \mid c = (1 - \lambda)a + \lambda b,\ 0 \leq \lambda \leq 1\},$$

and thus a set V is convex if $[a, b] \subseteq V$ for any two points $a, b \in V$ ($a = b$ is allowed). The *dimension* of a convex set V is the dimension of its affine hull $\mathrm{aff}(A)$.

The empty set is trivially convex, every one-point set $\{a\}$ is convex, and the entire affine space $\mathbb{R}^n$ is convex.

It is obvious that the intersection of any family (finite or infinite) of convex sets is convex.

Definition 44.6. Given any (nonempty) subset S of $\mathbb{R}^n$, the smallest convex set containing S is denoted by $\mathrm{conv}(S)$ and called the *convex hull of S* (it is the intersection of all convex sets containing S).

It is essential not only to have a good understanding of $\mathrm{conv}(S)$, but to also have good methods for computing it. We have the following simple but crucial result.

Proposition 44.1. *For any family $S = (a_i)_{i\in I}$ of points in $\mathbb{R}^n$, the set V of convex combinations $\sum_{i\in I} \lambda_i a_i$ (where $\sum_{i\in I} \lambda_i = 1$ and $\lambda_i \geq 0$) is the convex hull $\mathrm{conv}(S)$ of $S = (a_i)_{i\in I}$.*

It is natural to wonder whether Proposition 44.1 can be sharpened in two directions: (1) Is it possible to have a fixed bound on the number of points involved in the convex combinations? (2) Is it necessary to consider convex combinations of all points, or is it possible to consider only a subset with special properties?

The answer is yes in both cases. In Case 1, Carathéodory's theorem asserts that it is enough to consider convex combinations of $n+1$ points. For example, in the plane $\mathbb{R}^2$, the convex hull of a set S of points is the union of all triangles (interior points included) with vertices in S. In Case 2, the theorem of Krein and Milman asserts that a convex set that is also compact is the convex hull of its extremal points (given a convex set S, a point $a \in S$ is *extremal* if $S - \{a\}$ is also convex).

We will not prove these theorems here, but we invite the reader to consult Gallier [73] or Berger [12].

Convex sets also arise as half-spaces cut out by affine hyperplanes.

Definition 44.7. An *affine form* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is defined by some linear form $c \in (\mathbb{R}^n)^*$ and some scalar $\beta \in \mathbb{R}$ so that

$$\varphi(x) = cx + \beta \quad \text{for all } x \in \mathbb{R}^n.$$

If $c \neq 0$, the affine form φ specified by (c, β) defines the *affine hyperplane* (for short *hyperplane*) $H(\varphi)$ given by

$$H(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) = 0\} = \{x \in \mathbb{R}^n \mid cx + \beta = 0\},$$

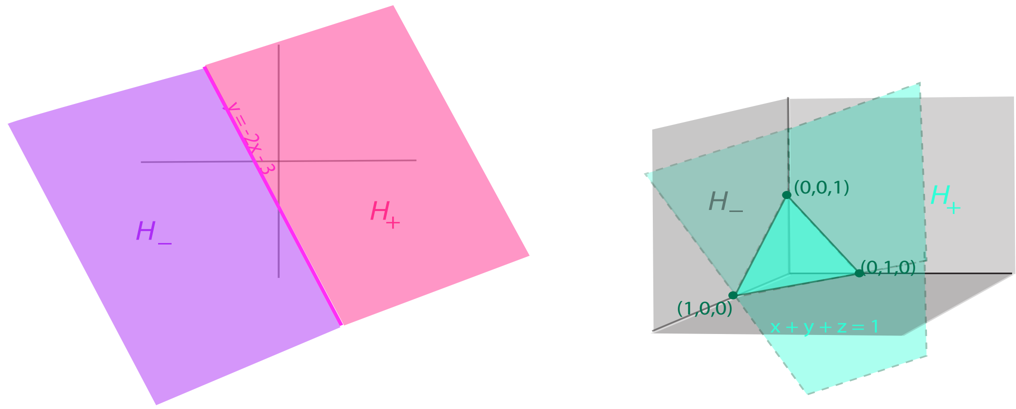
and the two *(closed) half-spaces*

$$\begin{aligned} H_+(\varphi) &= \{x \in \mathbb{R}^n \mid \varphi(x) \geq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \geq 0\}, \\ H_-(\varphi) &= \{x \in \mathbb{R}^n \mid \varphi(x) \leq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \leq 0\}. \end{aligned}$$

When $\beta = 0$, we call H a *linear hyperplane*.

Both $H_+(\varphi)$ and $H_-(\varphi)$ are convex and $H = H_+(\varphi) \cap H_-(\varphi)$.

For example, $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\varphi(x, y) = 2x + y + 3$ is an affine form defining the line given by the equation $y = -2x - 3$. Another example of an affine form is $\varphi\colon \mathbb{R}^3 \to \mathbb{R}$



i. ii.

Figure 44.2: Figure i. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y) = 2x + y + 3$, while Figure ii. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y, z) = x + y + z - 1$.

with $\varphi(x, y, z) = x + y + z - 1$; this affine form defines the plane given by the equation $x + y + z = 1$, which is the plane through the points $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$. Both of these hyperplanes are illustrated in Figure 44.2.

Definition 44.8. For any two vector $x, y \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ we write $x \leq y$ iff $x_i \leq y_i$ for $i = 1, \ldots, n$, and $x \geq y$ iff $y \leq x$. In particular $x \geq 0$ iff $x_i \geq 0$ for $i = 1, \ldots, n$.

Certain special types of convex sets called cones and $\mathcal{H}$-polyhedra play an important role. The set of feasible solutions of a linear program is an $\mathcal{H}$-polyhedron, and cones play a crucial role in the proof of Proposition 45.1 and in the Farkas–Minkowski proposition (Proposition 47.2).

44.3 Cones, Polyhedral Cones, and $\mathcal{H}$-Polyhedra

Cones and polyhedral cones are defined as follows.

Definition 44.9. Given a nonempty subset $S \subseteq \mathbb{R}^n$, the *cone* $C = \mathrm{cone}(S)$ spanned by S is the convex set

$$\mathrm{cone}(S) = \left\{ \sum_{i=1}^{k} \lambda_i u_i,\ u_i \in S, \lambda_i \in \mathbb{R},\ \lambda_i \geq 0 \right\},$$

of positive combinations of vectors from S. If S consists of a finite set of vectors, the cone $C = \mathrm{cone}(S)$ is called a *polyhedral cone*. Figure 44.3 illustrates a polyhedral cone.

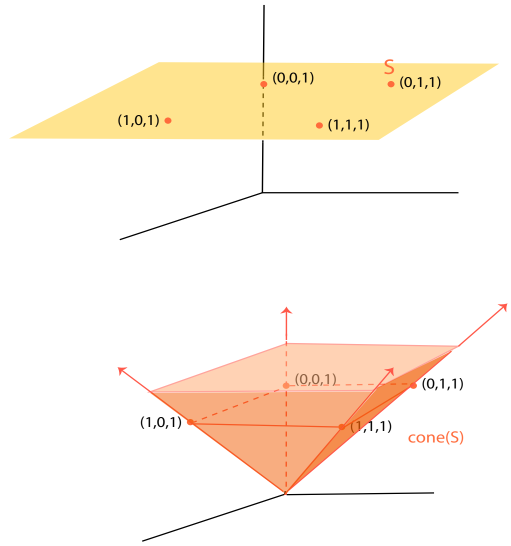


Figure 44.3: Let $S = \{(0,0,1),(1,0,1),(1,1,1),(0,1,1)\}$. The polyhedral cone, cone(S), is the solid "pyramid" with apex at the origin and square cross sections.

Note that if some nonzero vector u belongs to a cone C, then $\lambda u \in C$ for all $\lambda \geq 0$, that is, the *ray* $\{\lambda u \mid \lambda \geq 0\}$ belongs to C.

Remark: The cones (and polyhedral cones) of Definition 44.9 are *always convex*. For this reason, we use the simpler terminology cone instead of convex cone. However, there are more general kinds of cones (see Definition 50.1) that are not convex (for example, a union of polyhedral cones or the linear cone generated by the curve in Figure 44.4), and if we were dealing with those we would refer to the cones of Definition 44.9 as convex cones.

Definition 44.10. An $\mathcal{H}$-*polyhedron*, for short a *polyhedron*, is any subset $\mathcal{P} = \bigcap_{i=1}^{s} C_i$ of $\mathbb{R}^n$ defined as the intersection of a finite number s of closed half-spaces C_i. An example of an $\mathcal{H}$-polyhedron is shown in Figure 44.6. An $\mathcal{H}$-*polytope* is a bounded $\mathcal{H}$-polyhedron, which means that there is a closed ball $B_r(x)$ of center x and radius $r > 0$ such that $\mathcal{P} \subseteq B_r(x)$. An example of a $\mathcal{H}$-polytope is shown in Figure 44.5.

By convention, we agree that $\mathbb{R}^n$ itself is an $\mathcal{H}$-polyhedron.

Remark: The $\mathcal{H}$-polyhedra of Definition 44.10 are always convex. For this reason, as in the case of cones we use the simpler terminology $\mathcal{H}$-polyhedron instead of convex $\mathcal{H}$-polyhedron. In algebraic topology, there are more general polyhedra that are not convex.

It can be shown that an $\mathcal{H}$-polytope $\mathcal{P}$ is equal to the convex hull of finitely many points (the extreme points of $\mathcal{P}$). This is a nontrivial result whose proof takes a significant amount of work; see Gallier [73] and Ziegler [195].

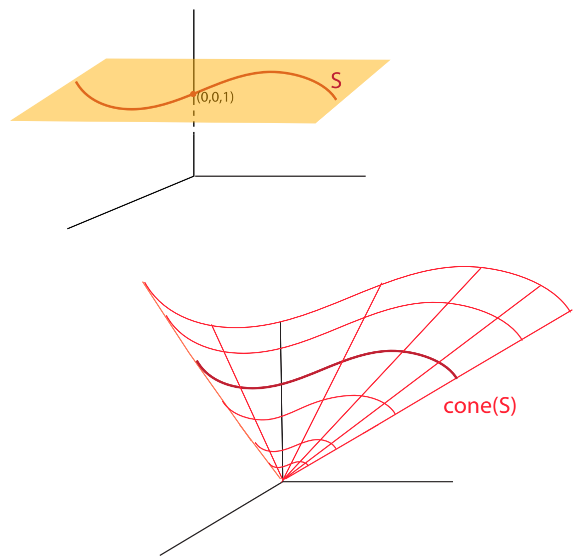


Figure 44.4: Let S be a planar curve in $z = 1$. The linear cone of S, consisting of all half rays connecting S to the origin, is not convex.

An unbounded $\mathcal{H}$-polyhedron is not equal to the convex hull of finite set of points. To obtain an equivalent notion we introduce the notion of a $\mathcal{V}$-polyhedron.

Definition 44.11. A *$\mathcal{V}$-polyhedron* is any convex subset $A \subseteq \mathbb{R}^n$ of the form

$$A = \mathrm{conv}(Y) + \mathrm{cone}(V) = \{a + v \mid a \in \mathrm{conv}(Y),\ v \in \mathrm{cone}(V)\},$$

where $Y \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ are *finite* (possibly empty).

When $V = \emptyset$ we simply have a *polytope*, and when $Y = \emptyset$ or $Y = \{0\}$, we simply have a cone.

It can be shown that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron and conversely. This is one of the major theorems in the theory of polyhedra, and its proof is nontrivial. For a complete proof, see Gallier [73] and Ziegler [195].

Every polyhedral cone is closed. This is an important fact that is used in the proof of several other key results such as Proposition 45.1 and the Farkas–Minkowski proposition (Proposition 47.2).

Although it seems obvious that a polyhedral cone should be closed, a rigorous proof is not entirely trivial.

Indeed, the fact that a polyhedral cone is closed relies crucially on the fact that C is spanned by a *finite* number of vectors, because the cone generated by an infinite set may not be closed. For example, consider the closed disk $D \subseteq \mathbb{R}^2$ of center $(0, 1)$ and radius 1,

Figure 44.5: An icosahedron is an example of an $\mathcal{H}$-polytope.

which is tangent to the x-axis at the origin. Then the cone(D) consists of the open upper half-plane *plus* the origin $(0, 0)$, but this set is not closed.

Proposition 44.2. *Every polyhedral cone C is closed.*

Proof. This is proven by showing that

1. Every primitive cone is closed, where a *primitive cone* is a polyhedral cone spanned by linearly independent vectors.

2. A polyhedral cone C is the union of finitely many primitive cones.

Assume that $(a_1, \ldots, a_m)$ are linearly independent vectors in $\mathbb{R}^n$, and consider any sequence $(x^{(k)})_{k\geq 0}$

$$x^{(k)} = \sum_{i=1}^{m} \lambda_i^{(k)} a_i$$

of vectors in the primitive cone cone$(\{a_1, \ldots, a_m\})$, which means that $\lambda_j^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$. The vectors $x^{(k)}$ belong to the subspace U spanned by $(a_1, \ldots, a_m)$, and U is closed. Assume that the sequence $(x^{(k)})_{k\geq 0}$ converges to a limit $x \in \mathbb{R}^n$. Since U is closed and $x^{(k)} \in U$ for all $k \geq 0$, we have $x \in U$. If we write $x = x_1 a_1 + \cdots + x_m a_m$, we would like to prove that $x_i \geq 0$ for $i = 1, \ldots, m$. The sequence the $(x^{(k)})_{k\geq 0}$ converges to x iff

$$\lim_{k \mapsto \infty} \left\| x^{(k)} - x \right\| = 0,$$

iff

$$\lim_{k \mapsto \infty} \left(\sum_{i=1}^{m} |\lambda_i^{(k)} - x_i|^2 \right)^{1/2} = 0$$

iff

$$\lim_{k \mapsto \infty} \lambda_i^{(k)} = x_i, \quad i = 1, \ldots, m.$$

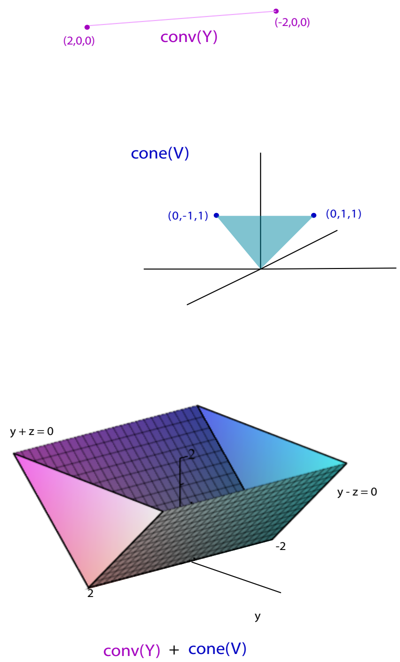


Figure 44.6: The "triangular trough" determined by the inequalities $y - z \leq 0$, $y + z \geq 0$, and $-2 \leq x \leq 2$ is an $\mathcal{H}$-polyhedron and an $\mathcal{V}$-polyhedron, where $Y = \{(2, 0, 0), (-2, 0, 0)\}$ and $V = \{(0, 1, 1), (0, -1, 1)\}$.

Since $\lambda_i^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$, we have $x_i \geq 0$ for $i = 1, \ldots, m$, so $x \in \mathrm{cone}(\{a_1, \ldots, a_m\})$.

Next, assume that x belongs to the polyhedral cone C. Consider a positive combination

$$x = \lambda_1 a_1 + \cdots + \lambda_k a_k, \tag{$*_1$}$$

for some nonzero $a_1, \ldots, a_k \in C$, with $\lambda_i \geq 0$ and with k *minimal*. Since k is minimal, we must have $\lambda_i > 0$ for $i = 1, \ldots, k$. We claim that $(a_1, \ldots, a_k)$ are linearly independent.

If not, there is some nontrivial linear combination

$$\mu_1 a_1 + \cdots + \mu_k a_k = 0, \tag{$*_2$}$$

and since the a_i are nonzero, $\mu_j \neq 0$ for some at least some j. We may assume that $\mu_j < 0$ for some j (otherwise, we consider the family $(-\mu_i)_{1 \leq i \leq k}$), so let

$$J = \{j \in \{1, \ldots, k\} \mid \mu_j < 0\}.$$

For any $t \in \mathbb{R}$, since $x = \lambda_1 a_1 + \cdots + \lambda_k a_k$, using $(*_2)$ we get

$$x = (\lambda_1 + t\mu_1) a_1 + \cdots + (\lambda_k + t\mu_k) a_k, \tag{$*_3$}$$

and if we pick

$$t = \min_{j \in J} \left(-\frac{\lambda_j}{\mu_j} \right) \geq 0,$$

we have $(\lambda_i + t\mu_i) \geq 0$ for $i = 1, \ldots, k$, but $\lambda_j + t\mu_j = 0$ for some $j \in J$, so $(*_3)$ is an expression of x with less that k nonzero coefficients, contradicting the minimality of k in $(*_1)$. Therefore, $(a_1, \ldots, a_k)$ are linearly independent.

Since a polyhedral cone C is spanned by finitely many vectors, there are finitely many primitive cones (corresponding to linearly independent subfamilies), and since every $x \in C$, belongs to some primitive cone, C is the union of a finite number of primitive cones. Since every primitive cone is closed, as a union of finitely many closed sets, C itself is closed.

The above facts are also proven in Matousek and Gardner [123] (Chapter 6, Section 5, Lemma 6.5.3, 6.5.4, and 6.5.5). $\square$

Another way to prove that a polyhedral cone C is closed is to show that C is also a $\mathcal{H}$-polyhedron. This takes even more work; see Gallier [73] (Chapter 4, Section 4, Proposition 4.16). Yet another proof is given in Lax [113] (Chapter 13, Theorem 1).

44.4 Summary

The main concepts and results of this chapter are listed below:

- Affine combination.
- Affine hull.
- Affine subspace; direction of an affine subspace, dimension of an affine subspace.
- Convex combination.
- Convex set, dimension of a convex set.
- Convex hull.
- Affine form.
- Affine hyperplane, half-spaces.
- Cone, polyhedral cone.
- $\mathcal{H}$-polyhedron, $\mathcal{H}$-polytope.
- $\mathcal{V}$-polyhedron, polytope.
- Primitive cone.

44.5 Problems

Problem 44.1. Prove Proposition 44.1.

Problem 44.2. Describe an icosahedron both as an $\mathcal{H}$-polytope and as a $\mathcal{V}$-polytope. Do the same thing for a dodecahedron. What do you observe?

Chapter 45

Linear Programs

In this chapter we introduce linear programs and the basic notions relating to this concept. We define the $\mathcal{H}$-polyhedron $\mathcal{P}(A, b)$ of feasible solutions. Then we define bounded and unbounded linear programs and the notion of optimal solution. We define slack variables and the important notion of *linear program in standard form*.

We show that if a linear program in standard form has a feasible solution and is bounded above, then it has an optimal solution. This is not an obvious result and the proof relies on the fact that a polyhedral cone is closed (this result was shown in the previous chapter).

Next we show that in order to find optimal solutions it suffices to consider solutions of a special form called *basic feasible solutions*. We prove that if a linear program in standard form has a feasible solution and is bounded above, then some basic feasible solution is an optimal solution (Theorem 45.4).

Geometrically, a basic feasible solution corresponds to a *vertex*. In Theorem 45.6 we prove that a basic feasible solution of a linear program in standard form is a vertex of the polyhedron $\mathcal{P}(A, b)$. Finally, we prove that if a linear program in standard form has some feasible solution, then it has a basic feasible solution (see Theorem 45.7). This fact allows the simplex algorithm described in the next chapter to get started.

45.1 Linear Programs, Feasible Solutions, Optimal Solutions

The purpose of linear programming is to solve the following type of optimization problem.

Definition 45.1. *A Linear Program* (P) is the following kind of optimization problem:

$$\begin{aligned}
&\text{maximize} \quad cx\\
&\text{subject to}\\
&\quad a_1 x \leq b_1\\
&\quad \ldots\\
&\quad a_m x \leq b_m\\
&\quad x \geq 0,
\end{aligned}$$

where $x \in \mathbb{R}^n$, $c, a_1, \ldots, a_m \in (\mathbb{R}^n)^*$, $b_1, \ldots, b_m \in \mathbb{R}$.

The linear form c defines the *objective function* $x \mapsto cx$ of the Linear Program (P) (from $\mathbb{R}^n$ to $\mathbb{R}$), and the inequalities $a_i x \leq b_i$ and $x_j \geq 0$ are called the *constraints* of the Linear Program (P).

If we define the $m \times n$ matrix

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

whose rows are the row vectors $a_1, \ldots, a_m$ and b as the column vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

the m inequality constraints $a_i x \leq b_i$ can be written in matrix form as

$$Ax \leq b.$$

Thus the Linear Program (P) can also be stated as the Linear Program (P):

$$\begin{aligned}
&\text{maximize} \quad cx\\
&\text{subject to} \quad Ax \leq b \ \text{and} \ x \geq 0.
\end{aligned}$$

We should note that in many applications, the natural primal optimization problem is actually the *minimization* of some objective function $cx = c_1 x_1 + \cdots + c_n x_n$, rather its maximization. For example, many of the optimization problems considered in Papadimitriou and Steiglitz [134] are minimization problems.

Of course, minimizing cx is equivalent to maximizing $-cx$, so our presentation covers minimization too.

Here is an explicit example of a linear program of Type (P):

Example 45.1.

$$
\begin{aligned}
\text{maximize} \quad & x_1 + x_2 \\
\text{subject to} \quad & \\
& x_2 - x_1 \leq 1 \\
& x_1 + 6x_2 \leq 15 \\
& 4x_1 - x_2 \leq 10 \\
& x_1 \geq 0,\; x_2 \geq 0,
\end{aligned}
$$

and in matrix form

$$
\begin{aligned}
\text{maximize} \quad & \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
\text{subject to} \quad & \\
& \begin{pmatrix} -1 & 1 \\ 1 & 6 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 15 \\ 10 \end{pmatrix} \\
& x_1 \geq 0,\; x_2 \geq 0.
\end{aligned}
$$

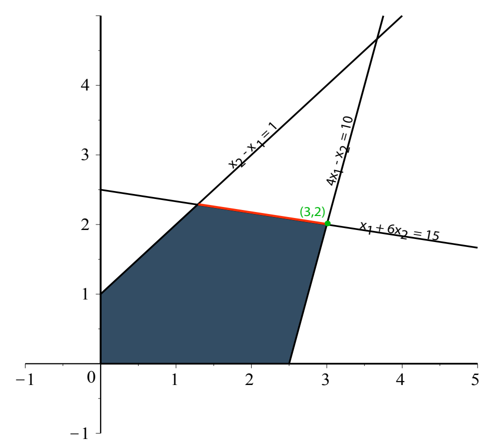


Figure 45.1: The $\mathcal{H}$-polyhedron associated with Example 45.1. The green point $(3, 2)$ is the unique optimal solution.

It turns out that $x_1 = 3, x_2 = 2$ yields the maximum of the objective function $x_1 + x_2$, which is 5. This is illustrated in Figure 45.1. Observe that the set of points that satisfy the above constraints is a convex region cut out by half planes determined by the lines of equations

$$\begin{aligned} x_2 - x_1 &= 1 \\ x_1 + 6x_2 &= 15 \\ 4x_1 - x_2 &= 10 \\ x_1 &= 0 \\ x_2 &= 0. \end{aligned}$$

In general, each constraint $a_i x \leq b_i$ corresponds to the affine form φ_i given by $\varphi_i(x) = a_i x - b_i$ and defines the half-space $H_-(\varphi_i)$, and each inequality $x_j \geq 0$ defines the half-space $H_+(x_j)$. The intersection of these half-spaces is the set of solutions of all these constraints. It is a (possibly empty) $\mathcal{H}$-polyhedron denoted $\mathcal{P}(A, b)$.

Definition 45.2. If $\mathcal{P}(A, b) = \emptyset$, we say that the Linear Program (P) has *no feasible solution*, and otherwise any $x \in \mathcal{P}(A, b)$ is called a *feasible solution* of (P).

The linear program shown in Example 45.2 obtained by reversing the direction of the inequalities $x_2 - x_1 \leq 1$ and $4x_1 - x_2 \leq 10$ in the linear program of Example 45.1 has no feasible solution; see Figure 45.2.

Example 45.2.

$$\begin{aligned} \text{maximize} \quad & x_1 + x_2 \\ \text{subject to} \quad & \\ & x_1 - x_2 \leq -1 \\ & x_1 + 6x_2 \leq 15 \\ & x_2 - 4x_1 \leq -10 \\ & x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Assume $\mathcal{P}(A, b) \neq \emptyset$, so that the Linear Program (P) has a feasible solution. In this case, consider the image $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ of $\mathcal{P}(A, b)$ under the objective function $x \mapsto cx$.

Definition 45.3. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is unbounded above, then we say that the Linear Program (P) is *unbounded*.

The linear program shown in Example 45.3 obtained from the linear program of Example 45.1 by deleting the constraints $4x_1 - x_2 \leq 10$ and $x_1 + 6x_2 \leq 15$ is unbounded.

Example 45.3.

$$\begin{aligned} \text{maximize} \quad & x_1 + x_2 \\ \text{subject to} \quad & \\ & x_2 - x_1 \leq 1 \\ & x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

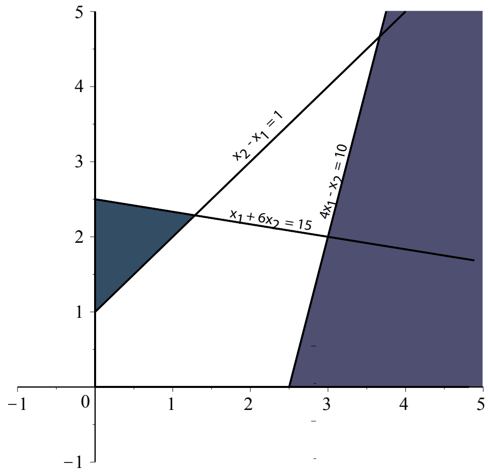


Figure 45.2: There is no $\mathcal{H}$-polyhedron associated with Example 45.2 since the blue and purple regions do not overlap.

Otherwise, we will prove shortly that if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some $p \in \mathcal{P}(A, b)$.

Definition 45.4. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, any point $p \in \mathcal{P}(A, b)$ such that $cp = \max\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is called an *optimal solution* (or *optimum*) of (P). Optimal solutions are often denoted by an upper $*$; for example, p^*.

The linear program of Example 45.1 has a unique optimal solution $(3, 2)$, but observe that the linear program of Example 45.4 in which the objective function is $(1/6)x_1 + x_2$ has infinitely many optimal solutions; the maximum of the objective function is $15/6$ which occurs along the points of orange boundary line in Figure 45.1.

Example 45.4.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && 4x_1 - x_2 \leq 10 \\ & && x_1 \geq 0,\; x_2 \geq 0. \end{aligned}$$

The proof that if the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A,b)\}$ is nonempty and bounded above, then there is an optimal solution $p \in \mathcal{P}(A,b)$, is not as trivial as it might seem. It relies on the fact that a polyhedral cone is closed, a fact that was shown in Section 44.3.

We also use a trick that makes the proof simpler, which is that a Linear Program (P) with inequality constraints $Ax \leq b$

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \ \text{ and } \ x \geq 0, \end{aligned}$$

is equivalent to the Linear Program (P_2) with equality constraints

$$\begin{aligned} &\text{maximize} \quad \widehat{c}\,\widehat{x} \\ &\text{subject to} \quad \widehat{A}\widehat{x} = b \ \text{ and } \ \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ is an $m \times (n+m)$ matrix, $\widehat{c}$ is a linear form in $(\mathbb{R}^{n+m})^*$, and $\widehat{x} \in \mathbb{R}^{n+m}$, given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad \text{and} \quad \widehat{x} = \begin{pmatrix} x \\ z \end{pmatrix},$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$.

Indeed, $\widehat{A}\widehat{x} = b$ and $\widehat{x} \geq 0$ iff

$$Ax + z = b, \quad x \geq 0, \ z \geq 0,$$

iff

$$Ax \leq b, \quad x \geq 0,$$

and $\widehat{c}\,\widehat{x} = cx$.

Definition 45.5. The variables z are called *slack variables*, and a linear program of the form (P_2) is called a linear program in *standard form*.

The result of converting the linear program of Example 45.4 to standard form is the program shown in Example 45.5.

Example 45.5.

$$\begin{aligned} &\text{maximize} \quad \frac{1}{6}x_1 + x_2 \\ &\text{subject to} \\ &\qquad x_2 - x_1 + z_1 = 1 \\ &\qquad x_1 + 6x_2 + z_2 = 15 \\ &\qquad 4x_1 - x_2 + z_3 = 10 \\ &\qquad x_1 \geq 0, \ x_2 \geq 0, \ z_1 \geq 0, \ z_2 \geq 0, \ z_3 \geq 0. \end{aligned}$$

We can now prove that if a linear program has a feasible solution and is bounded, then it has an optimal solution.

Proposition 45.1. *Let (P_2) be a linear program in standard form, with equality constraint $Ax = b$. If $\mathcal{P}(A, b)$ is nonempty and bounded above, and if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that*

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some optimum solution $p \in \mathcal{P}(A, b)$.

Proof. Since $\mu = \sup\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, there is a sequence $(x^{(k)})_{k\geq 0}$ of vectors $x^{(k)} \in \mathcal{P}(A, b)$ such that $\lim_{k\mapsto\infty} cx^{(k)} = \mu$. In particular, if we write $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ we have $x_j^{(k)} \geq 0$ for $j = 1, \ldots, n$ and for all $k \geq 0$. Let $\widetilde{A}$ be the $(m + 1) \times n$ matrix

$$\widetilde{A} = \begin{pmatrix} c \\ A \end{pmatrix},$$

and consider the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ of vectors $\widetilde{A}x^{(k)} \in \mathbb{R}^{m+1}$. We have

$$\widetilde{A}x^{(k)} = \begin{pmatrix} c \\ A \end{pmatrix} x^{(k)} = \begin{pmatrix} cx^{(k)} \\ Ax^{(k)} \end{pmatrix} = \begin{pmatrix} cx^{(k)} \\ b \end{pmatrix},$$

since by hypothesis $x^{(k)} \in \mathcal{P}(A, b)$, and the constraints are $Ax = b$ and $x \geq 0$. Since by hypothesis $\lim_{k\mapsto\infty} cx^{(k)} = \mu$, the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ converges to the vector $\begin{pmatrix} \mu \\ b \end{pmatrix}$. Now, observe that each vector $\widetilde{A}x^{(k)}$ can be written as the convex combination

$$\widetilde{A}x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} \widetilde{A}^j,$$

with $x_j^{(k)} \geq 0$ and where $\widetilde{A}^j \in \mathbb{R}^{m+1}$ is the jth column of $\widetilde{A}$. Therefore, $\widetilde{A}x^{(k)}$ belongs to the polyheral cone

$$C = \mathrm{cone}(\widetilde{A}^1, \ldots, \widetilde{A}^n) = \{\widetilde{A}x \mid x \in \mathbb{R}^n,\ x \geq 0\},$$

and since by Proposition 44.2 this cone is closed, $\lim_{k\geq\infty} \widetilde{A}x^{(k)} \in C$, which means that there is some $u \in \mathbb{R}^n$ with $u \geq 0$ such that

$$\begin{pmatrix} \mu \\ b \end{pmatrix} = \lim_{k\geq\infty} \widetilde{A}x^{(k)} = \widetilde{A}u = \begin{pmatrix} cu \\ Au \end{pmatrix},$$

that is, $cu = \mu$ and $Au = b$. Hence, u is an optimal solution of (P_2). ∎

The next question is, how do we find such an optimal solution? It turns out that for linear programs in standard form where the constraints are of the form $Ax = b$ and $x \geq 0$, there are always optimal solutions of a special type called *basic feasible solutions*.

45.2 Basic Feasible Solutions and Vertices

If the system $Ax = b$ has a solution and if some row of A is a linear combination of other rows, then the corresponding equation is redundant, so we may assume that the rows of A are linearly independent; that is, we may assume that A has rank m, so $m \leq n$.

Definition 45.6. If A is an $m \times n$ matrix, for any nonempty subset K of $\{1, \ldots, n\}$, let A_K be the submatrix of A consisting of the columns of A whose indices belong to K. We denote the jth column of the matrix A by A^j.

Definition 45.7. Given a Linear Program (P_2)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A has rank m, a vector $x \in \mathbb{R}^n$ is a *basic feasible solution* of (P) if $x \in \mathcal{P}(A, b) \neq \emptyset$, and if there is some subset K of $\{1, \ldots, n\}$ of size m such that

(1) The matrix A_K is invertible (that is, the columns of A_K are linearly independent).

(2) $x_j = 0$ for all $j \notin K$.

The subset K is called a *basis* of x. Every index $k \in K$ is called *basic*, and every index $j \notin K$ is called *nonbasic*. Similarly, the columns A^k corresponding to indices $k \in K$ are called *basic*, and the columns A^j corresponding to indices $j \notin K$ are called *nonbasic*. The variables corresponding to basic indices $k \in K$ are called *basic variables*, and the variables corresponding to indices $j \notin K$ are called *nonbasic*.

For example, the linear program

$$\begin{aligned} &\text{maximize} \quad x_1 + x_2 \\ &\text{subject to} \quad x_1 + x_2 + x_3 = 1 \text{ and } x_1 \geq 0, \ x_2 \geq 0, \ x_3 \geq 0, \end{aligned} \qquad (*)$$

has three basic feasible solutions; the basic feasible solution $K = \{1\}$ corresponds to the point $(1, 0, 0)$; the basic feasible solution $K = \{2\}$ corresponds to the point $(0, 1, 0)$; the basic feasible solution $K = \{3\}$ corresponds to the point $(0, 0, 1)$. Each of these points corresponds to the vertices of the slanted purple triangle illustrated in Figure 45.3. The vertices $(1, 0, 0)$ and $(0, 1, 0)$ optimize the objective function with a value of 1.

We now show that if the Standard Linear Program (P_2) as in Definition 45.7 has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. We follow Matousek and Gardner [123] (Chapter 4, Section 2, Theorem 4.2.3).

First we obtain a more convenient characterization of a basic feasible solution.

Proposition 45.2. *Given any Standard Linear Program (P_2) where $Ax = b$ and A is an $m \times n$ matrix of rank m, for any feasible solution x, if $J_> = \{j \in \{1, \ldots, n\} \mid x_j > 0\}$, then x is a basic feasible solution iff the columns of the matrix $A_{J_>}$ are linearly independent.*

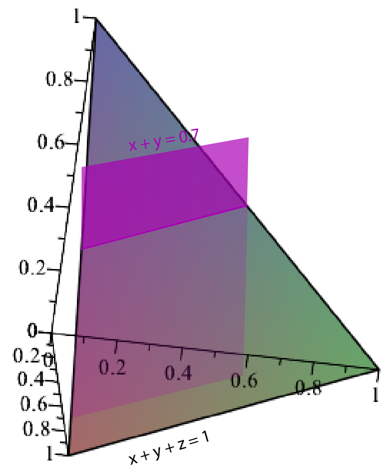


Figure 45.3: The $\mathcal{H}$-polytope associated with Linear Program $(*)$. The objective function (with $x_1 \to x$ and $x_2 \to y$) is represented by vertical planes parallel to the purple plane $x + y = 0.7$, and reaches it maximal value when $x + y = 1$.

Proof. If x is a basic feasible solution, then there is some subset $K \subseteq \{1, \ldots, n\}$ of size m such that the columns of A_K are linearly independent and $x_j = 0$ for all $j \notin K$, so by definition, $J_> \subseteq K$, which implies that the columns of the matrix $A_{J_>}$ are linearly independent.

Conversely, assume that x is a feasible solution such that the columns of the matrix $A_{J_>}$ are linearly independent. If $|J_>| = m$, we are done since we can pick $K = J_>$ and then x is a basic feasible solution. If $|J_>| < m$, we can extend $J_>$ to an m-element subset K by adding $m - |J_>|$ column indices so that the columns of A_K are linearly independent, which is possible since A has rank m. $\square$

Next we prove that if a linear program in standard form has any feasible solution x_0 and is bounded above, then is has some basic feasible solution $\widetilde{x}$ which is as good as x_0, in the sense that $c\widetilde{x} \geq cx_0$.

Proposition 45.3. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) is bounded above and if x_0 is some feasible solution of (P_2), then there is some basic feasible solution $\widetilde{x}$ such that $c\widetilde{x} \geq cx_0$.*

Proof. Among the feasible solutions x such that $cx \geq cx_0$ (x_0 is one of them) pick one with the *maximum* number of coordinates x_j equal to 0, say $\widetilde{x}$. Let

$$K = J_> = \{j \in \{1, \ldots, n\} \mid \widetilde{x}_j > 0\}$$

and let $s = |K|$. We claim that $\widetilde{x}$ is a basic feasible solution, and by construction $c\widetilde{x} \geq cx_0$.

If the columns of A_K are linearly independent, then by Proposition 45.2 we know that $\widetilde{x}$ is a basic feasible solution and we are done.

Otherwise, the columns of A_K are linearly dependent, so there is some nonzero vector $v = (v_1, \ldots, v_s)$ such that $A_K\, v = 0$. Let $w \in \mathbb{R}^n$ be the vector obtained by extending v by setting $w_j = 0$ for all $j \notin K$. By construction,

$$Aw = A_K\, v = 0.$$

We will derive a contradiction by exhibiting a feasible solution $x(t_0)$ such that $cx(t_0) \geq cx_0$ with more zero coordinates than $\widetilde{x}$.

For this we claim that we may assume that w satisfies the following two conditions:

(1) $cw \geq 0$.

(2) There is some $j \in K$ such that $w_j < 0$.

If $cw = 0$ and if Condition (2) fails, since $w \neq 0$, we have $w_j > 0$ for some $j \in K$, in which case we can use $-w$, for which $w_j < 0$.

If $cw < 0$, then $c(-w) > 0$, so we may assume that $cw > 0$. If $w_j > 0$ for all $j \in K$, since $\widetilde{x}$ is feasible, $\widetilde{x} \geq 0$, and so $x(t) = \widetilde{x} + tw \geq 0$ for all $t \geq 0$. Furthermore, since $Aw = 0$ and $\widetilde{x}$ is feasible, we have

$$Ax(t) = A\widetilde{x} + tAw = b,$$

and thus $x(t)$ is feasible for all $t \geq 0$. We also have

$$cx(t) = c\widetilde{x} + tcw.$$

Since $cw > 0$, as $t > 0$ goes to infinity the objective function $cx(t)$ also tends to infinity, contradicting the fact that is is bounded above. Therefore, some w satisfying Conditions (1) and (2) above must exist.

We show that there is some $t_0 > 0$ such that $cx(t_0) \geq cx_0$ and $x(t_0) = \widetilde{x} + t_0 w$ is feasible, yet $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction.

Since $x(t) = \widetilde{x} + tw$, we have

$$x(t)_i = \widetilde{x}_i + tw_i,$$

so if we let $I = \{i \in \{1, \ldots, n\} \mid w_i < 0\} \subseteq K$, which is nonempty since w satisfies Condition (2) above, if we pick

$$t_0 = \min_{i \in I} \left\{ \frac{-\widetilde{x}_i}{w_i} \right\},$$

then $t_0 > 0$, because $w_i < 0$ for all $i \in I$, and by definition of K we have $\widetilde{x}_i > 0$ for all $i \in K$. By the definition of $t_0 > 0$ and since $\widetilde{x} \geq 0$, we have

$$x(t_0)_j = \widetilde{x}_j + t_0 w_j \geq 0 \quad \text{for all } j \in K,$$

so $x(t_0) \geq 0$, and $x(t_0)_i = 0$ for some $i \in I$. Since $Ax(t_0) = b$ (for any t), $x(t_0)$ is a feasible solution,

$$cx(t_0) = c\widetilde{x} + t_0 cw \geq cx_0 + t_0 cw \geq cx_0,$$

and $x(t_0)_i = 0$ for some $i \in I$, we see that $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction. $\square$

Proposition 45.3 implies the following important result.

Theorem 45.4. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) has some feasible solution and if it is bounded above, then some basic feasible solution $\widetilde{x}$ is an optimal solution of (P_2).*

Proof. By Proposition 45.3, for any feasible solution x there is some basic feasible solution $\widetilde{x}$ such that $cx \leq c\widetilde{x}$. But there are only finitely many basic feasible solutions, so one of them has to yield the maximum of the objective function. $\square$

Geometrically, basic solutions are exactly the vertices of the polyhedron $\mathcal{P}(A, b)$, a notion that we now define.

Definition 45.8. Given an $\mathcal{H}$-polyhedron $\mathcal{P} \subseteq \mathbb{R}^n$, a *vertex* of $\mathcal{P}$ is a point $v \in \mathcal{P}$ with property that there is some nonzero linear form $c \in (\mathbb{R}^n)^*$ and some $\mu \in \mathbb{R}$, such that v is the unique point of $\mathcal{P}$ for which the map $x \mapsto cx$ has the maximum value μ ; that is, $cy < cv = \mu$ for all $y \in \mathcal{P} - \{v\}$. Geometrically, this means that the hyperplane of equation $cy = \mu$ touches $\mathcal{P}$ exactly at v. More generally, a convex subset F of $\mathcal{P}$ is a *k-dimensional face* of $\mathcal{P}$ if F has dimension k and if there is some affine form $\varphi(x) = cx - \mu$ such that $cy = \mu$ for all $y \in F$, and $cy < \mu$ for all $y \in \mathcal{P} - F$. A 1-dimensional face is called an *edge*.

The concept of a vertex is illustrated in Figure 45.4, while the concept of an edge is illustrated in Figure 45.5.

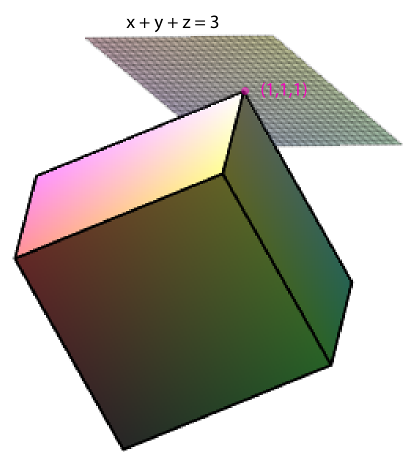


Figure 45.4: The cube centered at the origin with diagonal through $(-1, -1, -1)$ and $(1, 1, 1)$ has eight vertices. The vertex $(1, 1, 1)$ is associated with the linear form $x + y + z = 3$.

Since a k-dimensional face F of $\mathcal{P}$ is equal to the intersection of the hyperplane $H(\varphi)$ of equation $cx = \mu$ with $\mathcal{P}$, it is indeed convex and the notion of dimension makes sense.

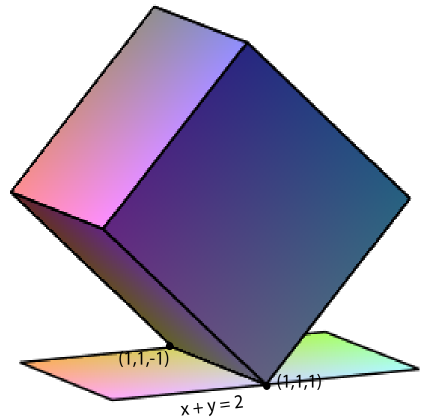


Figure 45.5: The cube centered at the origin with diagonal through $(-1,-1,-1)$ and $(1,1,1)$ has twelve edges. The edge from $(1,1,-1)$ to $(1,1,1)$ is associated with the linear form $x+y=2$.

Observe that a 0-dimensional face of $\mathcal{P}$ is a vertex. If $\mathcal{P}$ has dimension d, then the $(d-1)$-dimensional faces of $\mathcal{P}$ are called its *facets*.

If (P) is a linear program in standard form, then its basic feasible solutions are exactly the vertices of the polyhedron $\mathcal{P}(A,b)$. To prove this fact we need the following simple proposition

Proposition 45.5. *Let $Ax=b$ be a linear system where A is an $m\times n$ matrix of rank m. For any subset $K\subseteq\{1,\ldots,n\}$ of size m, if A_K is invertible, then there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$ (of course, $x\geq 0$)*

Proof. In order for x to be feasible we must have $Ax=b$. Write $N=\{1,\ldots,n\}-K$, x_K for the vector consisting of the coordinates of x with indices in K, and x_N for the vector consisting of the coordinates of x with indices in N. Then

$$Ax = A_Kx_K + A_Nx_N = b.$$

In order for x to be a basic feasible solution we must have $x_N=0$, so

$$A_Kx_K = b.$$

Since by hypothesis A_K is invertible, $x_K=A_K^{-1}b$ is uniquely determined. If $x_K\geq 0$ then x is a basic feasible solution, otherwise it is not. This proves that there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$. $\square$

Theorem 45.6. *Let (P) be a linear program in standard form, where $Ax=b$ and A is an $m\times n$ matrix of rank m. For every $v\in\mathcal{P}(A,b)$, the following conditions are equivalent:*

(1) v is a vertex of the Polyhedron $\mathcal{P}(A, b)$.

(2) v is a basic feasible solution of the Linear Program (P).

Proof. First, assume that v is a vertex of $\mathcal{P}(A, b)$, and let $\varphi(x) = cx - \mu$ be a linear form such that $cy < \mu$ for all $y \in \mathcal{P}(A, b)$ and $cv = \mu$. This means that v is the unique point of $\mathcal{P}(A, b)$ for which the objective function $x \mapsto cx$ has the maximum value μ on $\mathcal{P}(A, b)$, so by Theorem 45.4, since this maximum is achieved by some basic feasible solution, by uniqueness v must be a basic feasible solution.

Conversely, suppose v is a basic feasible solution of (P) corresponding to a subset $K \subseteq \{1, \ldots, n\}$ of size m. Let $\widehat{c} \in (\mathbb{R}^n)^*$ be the linear form defined by

$$\widehat{c}_j = \begin{cases} 0 & \text{if } j \in K \\ -1 & \text{if } j \notin K. \end{cases}$$

By construction $\widehat{c}\,v = 0$ and $\widehat{c}\,x \leq 0$ for any $x \geq 0$, hence the function $x \mapsto \widehat{c}\,x$ on $\mathcal{P}(A, b)$ has a maximum at v. Furthermore, $\widehat{c}\,x < 0$ for any $x \geq 0$ such that $x_j > 0$ for some $j \notin K$. However, by Proposition 45.5, the vector v is the only basic feasible solution such that $v_j = 0$ for all $j \notin K$, and therefore v is the only point of $\mathcal{P}(A, b)$ maximizing the function $x \mapsto \widehat{c}\,x$, so it is a vertex. □

In theory, to find an optimal solution we try all $\binom{n}{m}$ possible m-elements subsets K of $\{1, \ldots, n\}$ and solve for the corresponding unique solution x_K of $A_K x = b$. Then we check whether such a solution satisfies $x_K \geq 0$, compute cx_K, and return some feasible x_K for which the objective function is maximum. This is a totally impractical algorithm.

A practical algorithm is the *simplex algorithm*. Basically, the simplex algorithm tries to "climb" in the polyhderon $\mathcal{P}(A, b)$ from vertex to vertex along edges (using basic feasible solutions), trying to maximize the objective function. We present the simplex algorithm in the next chapter. The reader may also consult texts on linear programming. In particular, we recommend Matousek and Gardner [123], Chvatal [40], Papadimitriou and Steiglitz [134], Bertsimas and Tsitsiklis [21], Ciarlet [41], Schrijver [148], and Vanderbei [181].

Observe that Theorem 45.4 asserts that if a Linear Program (P) in standard form (where $Ax = b$ and A is an $m \times n$ matrix of rank m) has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. By Theorem 45.6, the polyhedron $\mathcal{P}(A, b)$ must have some vertex.

But suppose we only know that $\mathcal{P}(A, b)$ is nonempty; that is, we don't know that the objective function cx is bounded above. Does $\mathcal{P}(A, b)$ have some vertex?

The answer to the above question is *yes*, and this is important because the simplex algorithm needs an initial basic feasible solution to get started. Here we prove that if $\mathcal{P}(A, b)$ is nonempty, then it must contain a vertex. This proof still doesn't constructively yield a vertex, but we will see in the next chapter that the simplex algorithm always finds a vertex if there is one (provided that we use a pivot rule that prevents cycling).

Theorem 45.7. *Let (P) be a linear program in standard form, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If $\mathcal{P}(A, b)$ is nonempty (there is a feasible solution), then $\mathcal{P}(A, b)$ has some vertex; equivalently, (P) has some basic feasible solution.*

Proof. The proof relies on a trick, which is to add slack variables $x_{n+1}, \ldots, x_{n+m}$ and use the new objective function $-(x_{n+1} + \cdots + x_{n+m})$.

If we let $\widehat{A}$ be the $m \times (m + n)$-matrix, and x, $\overline{x}$, and $\widehat{x}$ be the vectors given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix} \in \mathbb{R}^m, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix} \in \mathbb{R}^{n+m},$$

then consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \ \text{and} \ \widehat{x} \geq 0. \end{aligned}$$

Since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0. The system $\widehat{A}\,\widehat{x} = b$ is equivalent to the system

$$Ax + \overline{x} = b,$$

so for every feasible solution $u \in \mathcal{P}(A, b)$, since $Au = b$, the vector $(u, 0_m)$ is also a feasible solution of $(\widehat{P})$, in fact an optimal solution since the value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ for $\overline{x} = 0$ is 0. By Proposition 45.3, the linear program $(\widehat{P})$ has some basic feasible solution (u^*, w^*) for which the value of the objective function is greater than or equal to the value of the objective function for $(u, 0_m)$, and since $(u, 0_m)$ is an optimal solution, (u^*, w^*) is also an optimal solution of $(\widehat{P})$. This implies that $w^* = 0$, since otherwise the objective function $-(x_{n+1} + \cdots + x_{n+m})$ would have a strictly negative value.

Therefore, $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$, and thus the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K associated with u^*, and u^* is indeed a basic feasible solution of (P). $\square$

The definition of a basic feasible solution can be adapted to linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$; see Matousek and Gardner [123] (Chapter 4, Section 4, Definition 4.4.2).

The most general type of linear program allows constraints of the form $a_i x \geq b_i$ or $a_i x = b_i$ besides constraints of the form $a_i x \leq b_i$. The variables x_i may also take negative values. It is always possible to convert such programs to the type considered in Definition 45.1. We proceed as follows.

Every constraint $a_i x \geq b_i$ is replaced by the constraint $-a_i x \leq -b_i$. Every equality constraint $a_i x = b_i$ is replaced by the two constraints $a_i x \leq b_i$ and $-a_i x \leq -b_i$.

If there are n variables x_i, we create n new variables y_i and n new variables z_i and replace every variable x_i by $y_i - z_i$. We also add the $2n$ constraints $y_i \geq 0$ and $z_i \geq 0$. If the constraints are given by the inequalities $Ax \leq b$, we now have constraints given by

$$\begin{pmatrix} A & -A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \leq b, \quad y \geq 0,\ z \geq 0.$$

We replace the objective function cx by $cy - cz$.

Remark: We also showed that we can replace the inequality constraints $Ax \leq b$ by equality constraints $Ax = b$, by adding slack variables constrained to be nonnegative.

45.3 Summary

The main concepts and results of this chapter are listed below:

- Linear program.
- Objective function, constraints.
- Feasible solution.
- Bounded and unbounded linear programs.
- Optimal solution, optimum.
- Slack variables, linear program in standard form.
- Basic feasible solution.
- Basis of a variable.
- Basic, nonbasic index, basic, nonbasic variable.
- Vertex, face, edge, facet.

45.4 Problems

Problem 45.1. Convert the following program to standard form:

$$\begin{aligned}
\text{maximize} \quad & x_1 + x_2 \\
\text{subject to} \quad & \\
& x_2 - x_1 \leq 1 \\
& x_1 + 6x_2 \leq 15 \\
& -4x_1 + x_2 \geq 10.
\end{aligned}$$

Problem 45.2. Convert the following program to standard form:

$$\begin{aligned} \text{maximize} \quad & 3x_1 - 2x_2 \\ \text{subject to} \quad & \\ & 2x_1 - x_2 \leq 4 \\ & x_1 + 3x_2 \geq 5 \\ & x_2 \geq 0. \end{aligned}$$

Problem 45.3. The notion of basic feasible solution for linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$ is defined as follows. A basic feasible solution of a (general) linear program with n variables is a feasible solution for which some n linearly independent constraints hold with equality.

Prove that the definition of a basic feasible solution for linear programs in standard form is a special case of the above definition.

Problem 45.4. Consider the linear program

$$\begin{aligned} \text{maximize} \quad & x_1 + x_2 \\ \text{subject to} \quad & \\ & x_1 + x_2 \leq 1. \end{aligned}$$

Show that none of the optimal solutions are basic.

Problem 45.5. The *standard n-simplex* is the subset Δ^n of $\mathbb{R}^{n+1}$ given by

$$\Delta^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1 + \cdots + x_{n+1} = 1,\ x_i \geq 0,\ 1 \leq i \leq n+1\}.$$

(1) Prove that Δ^n is convex and that it is the convex hull of the $n+1$ vectors $e_1, \ldots e_{n+1}$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n+1$.

(2) Prove that Δ^n is the intersection of $n+1$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume under the standard simplex Δ^n is $1/(n+1)!$.

Problem 45.6. The n-dimensional *cross-polytope* is the subset XP_n of $\mathbb{R}^n$ given by

$$XP_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_1| + \cdots + |x_n| \leq 1\}.$$

(1) Prove that XP_n is convex and that it is the convex hull of the $2n$ vectors $\pm e_i$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n$.

(2) Prove that XP_n is the intersection of 2^n half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of XP_n is $2^n/n!$.

Problem 45.7. The n-dimensional *hypercube* is the subset C_n of $\mathbb{R}^n$ given by

$$C_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_i| \leq 1, 1 \leq i \leq n\}.$$

(1) Prove that C_n is convex and that it is the convex hull of the 2^n vectors $(\pm 1, \ldots, \pm 1)$, $i = 1, \ldots, n$.

(2) Prove that C_n is the intersection of $2n$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of C_n is 2^n.

Chapter 46

The Simplex Algorithm

46.1 The Idea Behind the Simplex Algorithm

The simplex algorithm, due to Dantzig, applies to a linear program (P) in standard form, where the constraints are given by $Ax = b$ and $x \geq 0$, with A a $m \times n$ matrix of rank m, and with an objective function $x \mapsto cx$. This algorithm either reports that (P) has no feasible solution, or that (P) is unbounded, or yields an optimal solution. Geometrically, the algorithm climbs from vertex to vertex in the polyhedron $\mathcal{P}(A, b)$, trying to improve the value of the objective function. Since vertices correspond to basic feasible solutions, the simplex algorithm actually works with basic feasible solutions.

Recall that a basic feasible solution x is a feasible solution for which there is a subset $K \subseteq \{1, \ldots, n\}$ of size m such that the matrix A_K consisting of the columns of A whose indices belong to K are linearly independent, and that $x_j = 0$ for all $j \notin K$. We also let $J_>(x)$ be the set of indices

$$J_>(x) = \{j \in \{1, \ldots, n\} \mid x_j > 0\},$$

so for a basic feasible solution x associated with K, we have $J_>(x) \subseteq K$. In fact, by Proposition 45.2, a feasible solution x is a basic feasible solution iff the columns of $A_{J_>(x)}$ are linearly independent.

If $J_>(x)$ had cardinality m for all basic feasible solutions x, then the simplex algorithm would make progress at every step, in the sense that it would strictly increase the value of the objective function. Unfortunately, it is possible that $|J_>(x)| < m$ for certain basic feasible solutions, and in this case a step of the simplex algorithm may not increase the value of the objective function. Worse, in rare cases, it is possible that the algorithm enters an infinite loop. This phenomenon called *cycling* can be detected, but in this case the algorithm fails to give a conclusive answer.

Fortunately, there are ways of preventing the simplex algorithm from cycling (for example, Bland's rule discussed later), although proving that these rules work correctly is quite involved.

The potential "bad" behavior of a basic feasible solution is recorded in the following definition.

Definition 46.1. Given a Linear Program (P) in standard form where the constraints are given by $Ax = b$ and $x \geq 0$, with A an $m \times n$ matrix of rank m, a basic feasible solution x is *degenerate* if $|J_>(x)| < m$, otherwise it is *nondegenerate*.

The origin 0_n, if it is a basic feasible solution, is degenerate. For a less trivial example, $x = (0, 0, 0, 2)$ is a degenerate basic feasible solution of the following linear program in which $m = 2$ and $n = 4$.

Example 46.1.

$$
\begin{aligned}
&\text{maximize} && x_2\\
&\text{subject to} && \\
& && -x_1 + x_2 + x_3 = 0\\
& && x_1 + x_4 = 2\\
& && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0.
\end{aligned}
$$

The matrix A and the vector b are given by

$$
A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix},
$$

and if $x = (0, 0, 0, 2)$, then $J_>(x) = \{4\}$. There are two ways of forming a set of two linearly independent columns of A containing the fourth column.

Given a basic feasible solution x associated with a subset K of size m, since the columns of the matrix A_K are linearly independent, by abuse of language we call the columns of A_K a *basis* of x.

If u is a vertex of (P), that is, a basic feasible solution of (P) associated with a basis K (of size m), in "normal mode," the simplex algorithm tries to move along an edge from the vertex u to an adjacent vertex v (with $u, v \in \mathcal{P}(A, b) \subseteq \mathbb{R}^n$) corresponding to a basic feasible solution whose basis is obtained by replacing one of the basic vectors A^k with $k \in K$ by another nonbasic vector A^j for some $j \notin K$, in such a way that the value of the objective function is increased.

Let us demonstrate this process on an example.

Example 46.2. Let (P) be the following linear program in standard form.

$$
\begin{aligned}
\text{maximize} \quad & x_1 + x_2 \\
\text{subject to} \quad & \\
& -x_1 + x_2 + x_3 = 1 \\
& x_1 + x_4 = 3 \\
& x_2 + x_5 = 2 \\
& x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0,\ x_5 \geq 0.
\end{aligned}
$$

The matrix A and the vector b are given by

$$
A = \begin{pmatrix} -1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.
$$

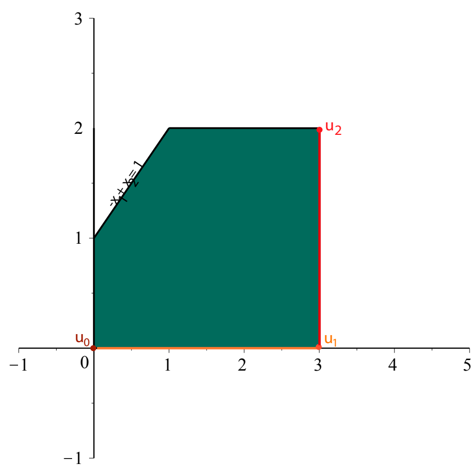


Figure 46.1: The planar $\mathcal{H}$-polyhedron associated with Example 46.2. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal orange line to feasible solution at vertex u_1. It then moves along the vertical red line to obtain the optimal feasible solution u_2.

The vector $u_0 = (0, 0, 1, 3, 2)$ corresponding to the basis $K = \{3, 4, 5\}$ is a basic feasible solution, and the corresponding value of the objective function is $0 + 0 = 0$. Since the columns (A^3, A^4, A^5) corresponding to $K = \{3, 4, 5\}$ are linearly independent we can express A^1 and A^2 as

$$
\begin{aligned}
A^1 &= -A^3 + A^4 \\
A^2 &= A^3 + A^5.
\end{aligned}
$$

Since

$$1A^3 + 3A^4 + 2A^5 = Au_0 = b,$$

for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^1 + \theta A^1 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + (1+\theta)A^3 + (3-\theta)A^4 + 2A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= \theta A^2 + (1-\theta)A^3 + 3A^4 + (2-\theta)A^5. \end{aligned}$$

In the first case, the vector $(\theta, 0, 1+\theta, 3-\theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is θ.

In the second case, the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$ is a feasible solution iff $0 \leq \theta \leq 1$, and the new value of the objective function is also θ.

Consider the first case. It is natural to ask whether we can get another vertex and increase the objective function by setting to zero one of the coordinates of $(\theta, 0, 1+\theta, 3-\theta, 2)$, in this case the fouth one, by picking $\theta = 3$. This yields the feasible solution $(3, 0, 4, 0, 2)$, which corresponds to the basis (A^1, A^3, A^5), and so is indeed a basic feasible solution, with an improved value of the objective function equal to 3. Note that A^4 *left* the basis (A^3, A^4, A^5) and A^1 *entered* the new basis (A^1, A^3, A^5).

We can now express A^2 and A^4 in terms of the basis (A^1, A^3, A^5), which is easy to do since we already have A^1 and A^2 in term of (A^3, A^4, A^5), and A^1 and A^4 are swapped. Such a step is called a *pivoting step*. We obtain

$$\begin{aligned} A^2 &= A^3 + A^5 \\ A^4 &= A^1 + A^3. \end{aligned}$$

Then we repeat the process with $u_1 = (3, 0, 4, 0, 2)$ and the basis (A^1, A^3, A^5). We have

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= 3A^1 + \theta A^2 + (4-\theta)A^3 + (2-\theta)A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + (4-\theta)A^3 + \theta A^4 + 2A^5. \end{aligned}$$

In the first case, the point $(3, \theta, 4-\theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the new value of the objective function is $3+\theta$. In the second case, the point $(3-\theta, 0, 4-\theta, \theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is $3-\theta$. To increase the objective function, we must choose the first case and we pick $\theta = 2$. Then we get the feasible solution $u_2 = (3, 2, 2, 0, 0)$, which corresponds to the basis (A^1, A^2, A^3), and thus is a basic feasible solution. The new value of the objective function is 5.

Next we express A^4 and A^5 in terms of the basis (A^1, A^2, A^3). Again this is easy to do since we just swapped A^5 and A^2 (a pivoting step), and we get

$$\begin{aligned} A^5 &= A^2 - A^3 \\ A^4 &= A^1 + A^3. \end{aligned}$$

We repeat the process with $u_2 = (3, 2, 2, 0, 0)$ and the basis (A^1, A^2, A^3). We have

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + 2A^2 + (2-\theta)A^3 + \theta A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^5 + \theta A^5 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^2 - A^3) + \theta A^5 \\ &= 3A^1 + (2-\theta)A^2 + (2+\theta)A^3 + \theta A^5. \end{aligned}$$

In the first case, the point $(3-\theta, 2, 2-\theta, \theta, 0)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is $5-\theta$. In the second case, the point $(3, 2-\theta, 2+\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is also $5-\theta$. Since we must have $\theta \geq 0$ to have a feasible solution, there is no way to increase the objective function. In this situation, it turns out that we have reached an optimal solution, in our case $u_2 = (3, 2, 2, 0, 0)$, with the maximum of the objective function equal to 5.

We could also have applied the simplex algorithm to the vertex $u_0 = (0, 0, 1, 3, 2)$ and to the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$, which is a feasible solution iff $0 \leq \theta \leq 1$, with new value of the objective function θ. By picking $\theta = 1$, we obtain the feasible solution $(0, 1, 0, 3, 1)$, corresponding to the basis (A^2, A^4, A^5), which is indeed a vertex. The new value of the objective function is 1. Then we express A^1 and A^3 in terms the basis (A^2, A^4, A^5) obtaining

$$\begin{aligned} A^1 &= A^4 - A^3 \\ A^3 &= A^2 - A^5, \end{aligned}$$

and repeat the process with $(0, 1, 0, 3, 1)$ and the basis (A^2, A^4, A^5). After three more steps we will reach the optimal solution $u_2 = (3, 2, 2, 0, 0)$.

Let us go back to the linear program of Example 46.1 with objective function x_2 and where the matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Recall that $u_0 = (0, 0, 0, 2)$ is a degenerate basic feasible solution, and the objective function has the value 0. See Figure 46.2 for a planar picture of the $\mathcal{H}$-polyhedron associated with Example 46.1.

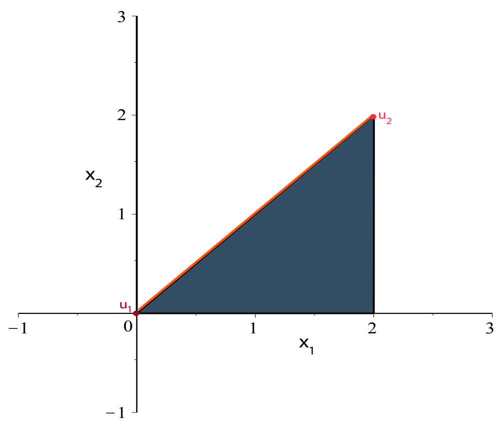


Figure 46.2: The planar $\mathcal{H}$-polyhedron associated with Example 46.1. The initial basic feasible solution is the origin. The simplex algorithm moves along the slanted orange line to the apex of the triangle.

Pick the basis (A^3, A^4). Then we have

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3, \end{aligned}$$

and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^3 + (2 - \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^2 + \theta A^2 \\ &= 2A^4 - \theta A^3 + \theta A^2 \\ &= \theta A^2 - \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, \theta, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is 0, and in the second case the point $(0, \theta, -\theta, 2)$ is a feasible solution iff $\theta = 0$, and the value of the objective function is θ. However, since we must have $\theta = 0$ in the second case, there is no way to increase the objective function either.

It turns out that in order to make the cases considered by the simplex algorithm as mutually exclusive as possible, since in the second case the coefficient of θ in the value of the objective function is nonzero, namely 1, we should choose the second case. We must pick $\theta = 0$, but we can swap the vectors A^3 and A^2 (because A^2 is coming in and A^3 has the coefficient $-\theta$, which is the reason why θ must be zero), and we obtain the basic feasible solution $u_1 = (0, 0, 0, 2)$ with the new basis (A^2, A^4). Note that this basic feasible solution corresponds to the same vertex $(0, 0, 0, 2)$ as before, but the basis has changed. The vectors A^1 and A^3 can be expressed in terms of the basis (A^2, A^4) as

$$\begin{aligned} A^1 &= -A^2 + A^4 \\ A^3 &= A^2. \end{aligned}$$

We now repeat the procedure with $u_1 = (0, 0, 0, 2)$ and the basis (A^2, A^4), and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^2 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^2 + (2-\theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^3 + \theta A^3 \\ &= 2A^4 - \theta A^2 + \theta A^3 \\ &= -\theta A^2 + \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, \theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is θ, and in the second case the point $(0, -\theta, \theta, 2)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is θ. In order to increase the objective function we must choose the first case and pick $\theta = 2$. We obtain the feasible solution $u_2 = (2, 2, 0, 0)$ whose corresponding basis is (A^1, A^2) and the value of the objective function is 2.

The vectors A^3 and A^4 are expressed in terms of the basis (A^1, A^2) as

$$\begin{aligned} A^3 &= A^2 \\ A^4 &= A^1 + A^3, \end{aligned}$$

and we repeat the procedure with $u_2 = (2, 2, 0, 0)$ and the basis (A^1, A^2). We get

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^3 + \theta A^3 \\ &= 2A^1 + 2A^2 - \theta A^2 + \theta A^3 \\ &= 2A^1 + (2-\theta)A^2 + \theta A^3, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^4 + \theta A^4 \\ &= 2A^1 + 2A^2 - \theta(A^1 + A^3) + \theta A^4 \\ &= (2-\theta)A^1 + 2A^2 - \theta A^3 + \theta A^4. \end{aligned}$$

In the first case, the point $(2, 2-\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is $2-\theta$, and in the second case, the point $(2-\theta, 2, -\theta, \theta)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is 2. This time there is no way to improve the objective function and we have reached an optimal solution $u_2 = (2, 2, 0, 0)$ with the maximum of the objective function equal to 2.

Let us now consider an example of an unbounded linear program.

Example 46.3. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} \quad x_1 \\ &\text{subject to} \\ &\qquad\qquad x_1 - x_2 + x_3 = 1 \\ &\qquad\quad -x_1 + x_2 + x_4 = 2 \\ &\qquad\qquad x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

The vector $u_0 = (0, 0, 1, 2)$ corresponding to the basis $K = \{3, 4\}$ is a basic feasible solution, and the corresponding value of the objective function is 0. The vectors A^1 and A^2 are expressed in terms of the basis (A^3, A^4) by

$$\begin{aligned} A^1 &= A^3 - A^4 \\ A^2 &= -A^3 + A^4. \end{aligned}$$

Starting with $u_0 = (0, 0, 1, 2)$, we get

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^1 + \theta A^1 \\ &= A^3 + 2A^4 - \theta(A^3 - A^4) + \theta A^1 \\ &= \theta A^1 + (1-\theta)A^3 + (2+\theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^2 + \theta A^2 \\ &= A^3 + 2A^4 - \theta(-A^3 + A^4) + \theta A^2 \\ &= \theta A^2 + (1+\theta)A^3 + (2-\theta)A^4. \end{aligned}$$

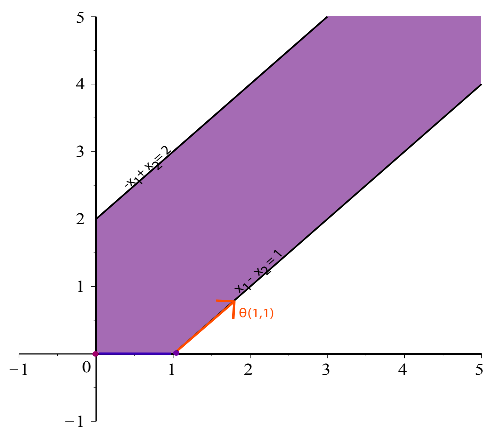


Figure 46.3: The planar $\mathcal{H}$-polyhedron associated with Example 46.3. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal indigo line to basic feasible solution at vertex $(1, 0)$. Any optimal feasible solution occurs by moving along the boundary line parameterized by the orange arrow $\theta(1, 1)$.

In the first case, the point $(\theta, 0, 1-\theta, 2+\theta)$ is a feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is θ, and in the second case, the point $(0, \theta, 1+\theta, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is 0. In order to increase the objective function we must choose the first case, and we pick $\theta = 1$. We get the feasible solution $u_1 = (1, 0, 0, 3)$ corresponding to the basis (A^1, A^4), so it is a basic feasible solution, and the value of the objective function is 1.

The vectors A^2 and A^3 are given in terms of the basis (A^1, A^4) by

$$\begin{aligned} A^2 &= -A^1 \\ A^3 &= A^1 + A^4. \end{aligned}$$

Repeating the process with $u_1 = (1, 0, 0, 3)$, we get

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^2 + \theta A^2 \\ &= A^1 + 3A^4 - \theta(-A^1) + \theta A^2 \\ &= (1+\theta)A^1 + \theta A^2 + 3A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^3 + \theta A^3 \\ &= A^1 + 3A^4 - \theta(A^1 + A^4) + \theta A^3 \\ &= (1-\theta)A^1 + \theta A^3 + (3-\theta)A^4. \end{aligned}$$

In the first case, the point $(1+\theta, \theta, 0, 3)$ is a feasible solution for all $\theta \geq 0$ and the value of the objective function if $1+\theta$, and in the second case, the point $(1-\theta, 0, \theta, 3-\theta)$ is a feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is $1-\theta$. This time, we are in the situation where the points

$$(1+\theta, \theta, 0, 3) = (1, 0, 0, 3) + \theta(1, 1, 0, 0), \quad \theta \geq 0$$

form an infinite ray in the set of feasible solutions, and the objective function $1+\theta$ is unbounded from above on this ray. This indicates that our linear program, although feasible, is unbounded.

Let us now describe a step of the simplex algorithm in general.

46.2 The Simplex Algorithm in General

We assume that we already have an initial vertex u_0 to start from. This vertex corresponds to a basic feasible solution with basis K_0. We will show later that it is always possible to find a basic feasible solution of a Linear Program (P) is standard form, or to detect that (P) has no feasible solution.

The idea behind the simplex algorithm is this: Given a pair (u, K) consisting of a basic feasible solution u and a basis K for u, find another pair (u^+, K^+) consisting of another basic feasible solution u^+ and a basis K^+ for u^+, such that K^+ is obtained from K by deleting some basic index $k^- \in K$ and adding some nonbasic index $j^+ \notin K$, in such a way that the value of the objective function increases (preferably strictly). The step which consists in swapping the vectors A^{k^-} and A^{j^+} is called a *pivoting step*.

Let u be a given vertex corresponds to a basic feasible solution with basis K. Since the m vectors A^k corresponding to indices $k \in K$ are linearly independent, they form a basis, so for every nonbasic $j \notin K$, we write

$$A^j = \sum_{k \in K} \gamma_k^j A^k. \tag{$*$}$$

We let $\gamma_K^j \in \mathbb{R}^m$ be the vector given by $\gamma_K^j = (\gamma_k^j)_{k \in K}$. Actually, since the vector γ_K^j depends on K, to be very precise we should denote its components by $(\gamma_K^j)_k$, but to simplify notation we usually write γ_k^j instead of $(\gamma_K^j)_k$ (unless confusion arises). We will explain later how the coefficients γ_k^j can be computed efficiently.

Since u is a feasible solution we have $u \geq 0$ and $Au = b$, that is,

$$\sum_{k \in K} u_k A^k = b. \tag{$**$}$$

For every nonbasic $j \notin K$, a candidate for entering the basis K, we try to find a new vertex $u(\theta)$ that improves the objective function, and for this we add $-\theta A^j + \theta A^j = 0$ to b in

Equation $(**)$ and then replace the occurrence of A^j in $-\theta A^j$ by the right hand side of Equation $(*)$ to obtain

$$\begin{aligned} b &= \sum_{k\in K} u_k A^k - \theta A^j + \theta A^j \\ &= \sum_{k\in K} u_k A^k - \theta\Big(\sum_{k\in K} \gamma^j_k A^k\Big) + \theta A^j \\ &= \sum_{k\in K}\Big(u_k - \theta\gamma^j_k\Big) A^k + \theta A^j. \end{aligned}$$

Consequently, the vector $u(\theta)$ appearing on the right-hand side of the above equation given by

$$u(\theta)_i = \begin{cases} u_i - \theta\gamma^j_i & \text{if } i \in K \\ \theta & \text{if } i = j \\ 0 & \text{if } i \notin K \cup \{j\} \end{cases}$$

automatically satisfies the constraints $Au(\theta) = b$, and this vector is a feasible solution iff

$$\theta \geq 0 \quad \text{and} \quad u_k \geq \theta\gamma^j_k \quad \text{for all } k \in K.$$

Obviously $\theta = 0$ is a solution, and if

$$\theta^j = \min\left\{ \frac{u_k}{\gamma^j_k} \;\middle|\; \gamma^j_k > 0,\ k \in K \right\} > 0,$$

then we have a range of feasible solutions for $0 \leq \theta \leq \theta^j$. The value of the objective function for $u(\theta)$ is

$$cu(\theta) = \sum_{k\in K} c_k(u_k - \theta\gamma^j_k) + \theta c_j = cu + \theta\Big(c_j - \sum_{k\in K}\gamma^j_k c_k\Big).$$

Since the potential change in the objective function is

$$\theta\Big(c_j - \sum_{k\in K}\gamma^j_k c_k\Big)$$

and $\theta \geq 0$, if $c_j - \sum_{k\in K}\gamma^j_k c_k \leq 0$, then the objective function can't be increased.

However, if $c_{j^+} - \sum_{k\in K}\gamma^{j^+}_k c_k > 0$ for some $j^+ \notin K$, and if $\theta^{j^+} > 0$, then the objective function can be strictly increased by choosing any $\theta > 0$ such that $\theta \leq \theta^{j^+}$, so it is natural to zero at least one coefficient of $u(\theta)$ by picking $\theta = \theta^{j^+}$, which also maximizes the increase of the objective function. In this case (Case below (B2)), we obtain a new feasible solution $u^+ = u(\theta^{j^+})$.

Now, if $\theta^{j^+} > 0$, then there is some index $k \in K$ such $u_k > 0$, $\gamma^{j^+}_k > 0$, and $\theta^{j^+} = u_k/\gamma^{j^+}_k$, so we can pick such an index k^- for the vector A^{k^-} leaving the basis K. We claim that

$K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis. This is because the coefficient $\gamma_{k^-}^{j^+}$ associated with the column A^{k^-} is nonzero (in fact, $\gamma_{k^-}^{j^+} > 0$), so Equation ($*$), namely

$$A^{j^+} = \gamma_{k^-}^{j^+} A^{k^-} + \sum_{k \in K - \{k^-\}} \gamma_k^{j^+} A^k,$$

yields the equation

$$A^{k^-} = (\gamma_{k^-}^{j^+})^{-1} A^{j^+} - \sum_{k \in K - \{k^-\}} (\gamma_{k^-}^{j^+})^{-1} \gamma_k^{j^+} A^k,$$

and these equations imply that the subspaces spanned by the vectors $(A^k)_{k\in K}$ and the vectors $(A^k)_{k\in K^+}$ are identical. However, K is a basis of dimension m so this subspace has dimension m, and since K^+ also has m elements, it must be a basis. Therefore, $u^+ = u(\theta^{j^+})$ is a basic feasible solution.

The above case is the most common one, but other situations may arise. In what follows, we discuss all eventualities.

Case (A).

We have $c_j - \sum_{k\in K} \gamma_k^j c_k \leq 0$ for all $j \notin K$. Then it turns out that u is an *optimal solution*. Otherwise, we are in Case (B).

Case (B).

We have $c_j - \sum_{k\in K} \gamma_k^j c_k > 0$ for some $j \notin K$ (not necessarily unique). There are three subcases.

Case (B1).

If for some $j \notin K$ as above we also have $\gamma_k^j \leq 0$ for all $k \in K$, since $u_k \geq 0$ for all $k \in K$, this places no restriction on θ, and the objective function is *unbounded above*. This is demonstrated by Example 46.3 with $K = \{3, 4\}$ and $j = 2$ since $\gamma_{\{3,4\}}^2 = (-1, 0)$.

Case (B2).

There is some index $j^+ \notin K$ such that simultaneously

(1) $c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^{j^+} > 0$, and for every $k \in K$, if $\gamma_k^{j^+} > 0$ then $u_k > 0$, which implies that $\theta^{j^+} > 0$.

If we pick $\theta = \theta^{j^+}$ where

$$\theta^{j^+} = \min\left\{ \frac{u_k}{\gamma_k^{j^+}} \;\middle|\; \gamma_k^{j^+} > 0,\ k \in K \right\} > 0,$$

then the feasible solution u^+ given by

$$u_i^+ = \begin{cases} u_i - \theta^{j^+}\gamma_i^{j^+} & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}$$

is a vertex of $\mathcal{P}(A, b)$. If we pick any index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$, then $K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis for u^+. The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. The objective function increases strictly. This is demonstrated by Example 46.2 with $K = \{3, 4, 5\}$, $j = 1$, and $k = 4$, Then $\gamma_{\{3,4,5\}}^1 = (-1, 1, 0)$, with $\gamma_4^1 = 1$. Since $u = (0, 0, 1, 3, 2)$, $\theta^1 = \frac{u_4}{\gamma_4^1} = 3$, and the new optimal solutions becomes $u^+ = (3, 0, 1 - 3(-1), 3 - 3(1), 2 - 3(0)) = (3, 0, 4, 0, 2)$.

Case (B3).

There is some index $j \notin K$ such that $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$, and for each of the indices $j \notin K$ satisfying the above property we have simultaneously

(1) $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^j > 0$, and $u_k = 0$, which implies that $\theta^j = 0$.

Consequently, the objective function *does not change*. In this case, u is a degenerate basic feasible solution.

We can associate to $u^+ = u$ a new basis K^+ as follows: Pick any index $j^+ \notin K$ such that

$$c_{j^+} - \sum_{k \in K} \gamma_k^{j^+} c_k > 0,$$

and any index $k^- \in K$ such that

$$\gamma_{k^-}^{j^+} > 0,$$

and let $K^+ = (K - \{k^-\}) \cup \{j^+\}$. As in Case (B2), The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. However, the objective function *does not change* since $\theta^{j+} = 0$. This is demonstrated by Example 46.1 with $K = \{3, 4\}$, $j = 2$, and $k = 3$.

It is easy to prove that in Case (A) the basic feasible solution u is an optimal solution, and that in Case (B1) the linear program is unbounded. We already proved that in Case (B2) the vector u^+ and its basis K^+ constitutes a basic feasible solution, and the proof in Case (B3) is similar. For details, see Ciarlet [41] (Chapter 10).

It is convenient to reinterpret the various cases considered by introducing the following sets:

$$B_1 = \Big\{ j \notin K \mid c_j - \sum_{k\in K} \gamma^j_k c_k > 0,\ \max_{k\in K} \gamma^j_k \leq 0 \Big\}$$
$$B_2 = \Big\{ j \notin K \mid c_j - \sum_{k\in K} \gamma^j_k c_k > 0,\ \max_{k\in K} \gamma^j_k > 0,\ \min\Big\{ \frac{u_k}{\gamma^j_k} \;\Big|\; k \in K, \gamma^j_k > 0 \Big\} > 0 \Big\}$$
$$B_3 = \Big\{ j \notin K \mid c_j - \sum_{k\in K} \gamma^j_k c_k > 0,\ \max_{k\in K} \gamma^j_k > 0,\ \min\Big\{ \frac{u_k}{\gamma^j_k} \;\Big|\; k \in K, \gamma^j_k > 0 \Big\} = 0 \Big\},$$

and

$$B = B_1 \cup B_2 \cup B_3 = \Big\{ j \notin K \mid c_j - \sum_{k\in K} \gamma^j_k c_k > 0 \Big\}.$$

Then it is easy to see that the following equivalences hold:

$$\begin{aligned} &\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset \\ &\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset \\ &\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset \\ &\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset. \end{aligned}$$

Furthermore, Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

If Case (B1) and Case (B2) arise simultaneously, we opt for Case (B1) which says that the Linear Program (P) is unbounded and terminate the algorithm.

Here are a few remarks about the method.

In Case (B2), which is the path followed by the algorithm most frequently, various choices have to be made for the index $j^+ \notin K$ for which $\theta^{j^+} > 0$ (the new index in K^+). Similarly, various choices have to be made for the index $k^- \in K$ leaving K, but such choices are typically less important.

Similarly in Case (B3), various choices have to be made for the new index $j^+ \notin K$ going into K^+. In Cases (B2) and (B3), criteria for making such choices are called *pivot rules*.

Case (B3) only arises when u is a degenerate vertex. But even if u is degenerate, Case (B2) may arise if $u_k > 0$ whenever $\gamma^j_k > 0$. It may also happen that u is nondegenerate but as a result of Case (B2), the new vertex u^+ is degenerate because at least two components $u_{k_1} - \theta^{j^+}\gamma^{j^+}_{k_1}$ and $u_{k_2} - \theta^{j^+}\gamma^{j^+}_{k_2}$ vanish for some distinct $k_1, k_2 \in K$.

Cases (A) and (B1) correspond to situations where the algorithm terminates, and Case (B2) can only arise a finite number of times during execution of the simplex algorithm, since the objective function is strictly increased from vertex to vertex and there are only finitely many vertices. Therefore, if the simplex algorithm is started on any initial basic feasible solution u_0, then one of three mutually exclusive situations may arise:

(1) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (A). Then the last vertex produced by the algorithm is an optimal solution. This is what occurred in Examples 46.1 and 46.2.

(2) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (B1). We conclude that the problem is unbounded, and thus has no solution. This is what occurred in Example 46.3.

(3) There is a finite sequence of occurrences of Case (B2) and/or Case (B3), followed by an infinite sequence of Case (B3). If this occurs, the algorithm visits the some basis twice. This a phenomenon known as *cycling*. In this eventually the algorithm fails to come to a conclusion.

There are examples for which cycling occur, although this is rare in practice. Such an example is given in Chvatal [40]; see Chapter 3, pages 31-32, for an example with seven variables and three equations that cycles after six iterations under a certain pivot rule.

The third possibility can be avoided by the choice of a suitable pivot rule. Two of these rules are *Bland's rule* and the *lexicographic rule*; see Chvatal [40] (Chapter 3, pages 34-38).

Bland's rule says: choose the smallest of the eligible incoming indices $j^+ \notin K$, and similarly choose the smallest of the eligible outgoing indices $k^- \in K$.

It can be proven that cycling cannot occur if Bland's rule is chosen as the pivot rule. The proof is very technical; see Chvatal [40] (Chapter 3, pages 37-38), Matousek and Gardner [123] (Chapter 5, Theorem 5.8.1), and Papadimitriou and Steiglitz [134] (Section 2.7). Therefore, assuming that some initial basic feasible solution is provided, and using a suitable pivot rule (such as Bland's rule), the simplex algorithm always terminates and either yields an optimal solution or reports that the linear program is unbounded. Unfortunately, Bland's rules is one of the slowest pivot rules.

The choice of a pivot rule affects greatly the number of pivoting steps that the simplex algorithms goes through. It is not our intention here to explain the various pivot rules. We simply mention the following rules, referring the reader to Matousek and Gardner [123] (Chapter 5, Section 5.7) or to the texts cited in Section 44.1.

1. Largest coefficient, or Dantzig's rule.

2. Largest increase.

3. Steepest edge.

4. Bland's Rule.

5. Random edge.

The steepest edge rule is one of the most popular. The idea is to maximize the ratio

$$\frac{c(u^+ - u)}{\|u^+ - u\|}.$$

The random edge rule picks the index $j^+ \notin K$ of the entering basis vector uniformly at random among all eligible indices.

Let us now return to the issue of the initialization of the simplex algorithm. We use the Linear Program $(\widehat{P})$ introduced during the proof of Theorem 45.7.

Consider a Linear Program $(P2)$

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form where A is an $m \times n$ matrix of rank m.

First, observe that since the constraints are equations, we can ensure that $b \geq 0$, because every equation $a_i x = b_i$ where $b_i < 0$ can be replaced by $-a_i x = -b_i$. The next step is to introduce the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} \quad -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} \quad \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ and $\widehat{x}$ are given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n+m} \end{pmatrix}.$$

Since we assumed that $b \geq 0$, the vector $\widehat{x} = (0_n, b)$ is a feasible solution of $(\widehat{P})$, in fact a basic feasible solutions since the matrix associated with the indices $n+1, \ldots, n+m$ is the identity matrix I_m. Furthermore, since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0.

If we execute the simplex algorithm with a pivot rule that prevents cycling, starting with the basic feasible solution $(0_n, d)$, since the objective function is bounded by 0, the simplex algorithm terminates with an optimal solution given by some basic feasible solution, say (u^*, w^*), with $u^* \in \mathbb{R}^n$ and $w^* \in \mathbb{R}^m$.

As in the proof of Theorem 45.7, for every feasible solution $u \in \mathcal{P}(A, b)$, the vector $(u, 0_m)$ is an optimal solution of $(\widehat{P})$. Therefore, if $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, since otherwise for every feasible solution $u \in \mathcal{P}(A, b)$ the vector $(u, 0_m)$ would yield a value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ equal to 0, but (u^*, w^*) yields a strictly negative value since $w^* \neq 0$.

Otherwise, $w^* = 0$, and u^* is a feasible solution of $(P2)$. Since $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$ the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K^* associated with u^*, and u^* is indeed a basic feasible solution of $(P2)$.

Running the simplex algorithm on the Linear Program $\widehat{P}$ to obtain an initial feasible solution (u_0, K_0) of the linear program $(P2)$ is called *Phase I* of the simplex algorithm. Running the simplex algorithm on the Linear Program $(P2)$ with some initial feasible solution (u_0, K_0) is called *Phase II* of the simplex algorithm. If a feasible solution of the Linear Program $(P2)$ is readily available then Phase I is skipped. Sometimes, at the end of Phase I, an optimal solution of $(P2)$ is already obtained.

In summary, we proved the following fact worth recording.

Proposition 46.1. *For any Linear Program $(P2)$*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \ \textit{and} \ x \geq 0, \end{aligned}$$

in standard form, where A is an $m \times n$ matrix of rank m and $b \geq 0$, consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} \quad -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} \quad \widehat{A}\,\widehat{x} = b \ \textit{and} \ \widehat{x} \geq 0. \end{aligned}$$

The simplex algorithm with a pivot rule that prevents cycling started on the basic feasible solution $\widehat{x} = (0_n, b)$ of $(\widehat{P})$ terminates with an optimal solution (u^*, w^*).

(1) If $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, that is, the Linear Program $(P2)$ has no feasible solution.

(2) If $w^* = 0$, then $\mathcal{P}(A, b) \neq \emptyset$, and u^* is a basic feasible solution of $(P2)$ associated with some basis K.

Proposition 46.1 shows that determining whether the polyhedron $\mathcal{P}(A, b)$ defined by a system of equations $Ax = b$ and inequalities $x \geq 0$ is nonempty is decidable. This decision procedure uses a fail-safe version of the simplex algorithm (that prevents cycling), and the proof that it always terminates and returns an answer is nontrivial.

46.3 How to Perform a Pivoting Step Efficiently

We now discuss briefly how to perform the computation of (u^+, K^+) from a basic feasible solution (u, K).

In order to avoid applying permutation matrices it is preferable to allow a basis K to be a sequence of indices, possibly out of order. Thus, for any $m \times n$ matrix A (with $m \leq n$)

and any sequence $K = (k_1, k_2, \cdots, k_m)$ of m elements with $k_i \in \{1, \ldots, n\}$, the matrix A_K denotes the $m \times m$ matrix whose ith column is the k_ith column of A, and similarly for any vector $u \in \mathbb{R}^n$ (resp. any linear form $c \in (\mathbb{R}^n)^*$), the vector $u_K \in \mathbb{R}^m$ (the linear form $c_K \in (\mathbb{R}^m)^*$) is the vector whose ith entry is the k_ith entry in u (resp. the linear whose ith entry is the k_ith entry in c).

For each nonbasic $j \notin K$, we have

$$A^j = \gamma^j_{k_1} A^{k_1} + \cdots + \gamma^j_{k_m} A^{k_m} = A_K \gamma^j_K,$$

so the vector γ^j_K is given by $\gamma^j_K = A_K^{-1} A^j$, that is, by solving the system

$$A_K \gamma^j_K = A^j. \tag{$*_\gamma$}$$

To be very precise, since the vector γ^j_K depends on K its components should be denoted by $(\gamma^j_K)_{k_i}$, but as we said before, to simplify notation we write $\gamma^j_{k_i}$ instead of $(\gamma^j_K)_{k_i}$.

In order to decide which case applies ((A), (B1), (B2), (B3)), we need to compute the numbers $c_j - \sum_{k \in K} \gamma^j_k c_k$ for all $j \notin K$. For this, observe that

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - c_K \gamma^j_K = c_j - c_K A_K^{-1} A^j.$$

If we write $\beta_K = c_K A_K^{-1}$, then

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - \beta_K A^j,$$

and we see that $\beta_K^\top \in \mathbb{R}^m$ is the solution of the system $\beta_K^\top = (A_K^{-1})^\top c_k^\top$, which means that $\beta_K^\top$ is the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top. \tag{$*_\beta$}$$

Remark: Observe that since u is a basis feasible solution of (P), we have $u_j = 0$ for all $j \notin K$, so u is the solution of the equation $A_K u_K = b$. As a consequence, the value of the objective function for u is $cu = c_K u_K = c_K A_K^{-1} b$. This fact will play a crucial role in Section 47.2 to show that when the simplex algorithm terminates with an optimal solution of the Linear Program (P), then it also produces an optimal solution of the Dual Linear Program (D).

Assume that we have a basic feasible solution u, a basis K for u, and that we also have the matrix A_K as well its inverse A_K^{-1} (perhaps implicitly) and also the inverse $(A_K^\top)^{-1}$ of $A_K^\top$ (perhaps implicitly). Here is a description of an iteration step of the simplex algorithm, following almost exactly Chvatal (Chvatal [40], Chapter 7, Box 7.1).

An Iteration Step of the (Revised) Simplex Method

Step 1. Compute the numbers $c_j - \sum_{k\in K} \gamma_k^j c_k = c_j - \beta_K A^j$ for all $j \notin K$, and for this, compute $\beta_K^\top$ as the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top.$$

If $c_j - \beta_K A^j \leq 0$ for all $j \notin K$, stop and return the optimal solution u (Case (A)).

Step 2. If Case (B) arises, use a pivot rule to determine which index $j^+ \notin K$ should enter the new basis K^+ (the condition $c_{j^+} - \beta_K A^{j^+} > 0$ should hold).

Step 3. Compute $\max_{k\in K} \gamma_k^{j^+}$. For this, solve the linear system

$$A_K \gamma_K^{j^+} = A^{j^+}.$$

Step 4. If $\max_{k\in K} \gamma_k^{j^+} \leq 0$, then stop and report that Linear Program (P) is unbounded (Case (B1)).

Step 5. If $\max_{k\in K} \gamma_k^{j^+} > 0$, use the ratios $u_k/\gamma_k^{j^+}$ for all $k \in K$ such that $\gamma_k^{j^+} > 0$ to compute θ^{j^+}, and use a pivot rule to determine which index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$ should leave K (Case (B2)).

If $\max_{k\in K} \gamma_k^{j^+} = 0$, then use a pivot rule to determine which index k^- for which $\gamma_{k^-}^{j^+} > 0$ should leave the basis K (Case (B3)).

Step 6. Update u, K, and A_K, to u^+ and K^+, and A_{K^+}. During this step, given the basis K specified by the sequence $K = (k_1, \ldots, k_\ell, \ldots, k_m)$, with $k^- = k_\ell$, then K^+ is the sequence obtained by replacing k_ℓ by the incoming index j^+, so $K^+ = (k_1, \ldots, j^+, \ldots, k_m)$ with j^+ in the ℓth slot.

The vector u is easily updated. To compute A_{K^+} from A_K we take advantage of the fact that A_K and A_{K^+} only differ by a *single column*, namely the ℓth column A^{j^+}, which is given by the linear combination

$$A^{j^+} = A_K \gamma_K^{j^+}.$$

To simplify notation, denote $\gamma_K^{j^+}$ by γ, and recall that $k^- = k_\ell$. If $K = (k_1, \ldots, k_m)$, then $A_K = [A^{k_1} \cdots A^{k^-} \cdots A^{i_m}]$, and since A_{K^+} is the result of replacing the ℓth column A^{k^-} of A_K by the column A^{j^+}, we have

$$A_{K^+} = [A^{k_1} \cdots A^{j^+} \cdots A^{i_m}] = [A^{k_1} \cdots A_K\gamma \cdots A^{i_m}] = A_K E(\gamma),$$

where $E(\gamma)$ is the following invertible matrix obtained from the identity matrix I_m by replacing its ℓth column by γ:

$$E(\gamma) = \begin{pmatrix} 1 & & & \gamma_1 & & & \\ & \ddots & & \vdots & & & \\ & & 1 & \gamma_{\ell-1} & & & \\ & & & \gamma_\ell & & & \\ & & & \gamma_{\ell+1} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & \gamma_m & & & 1 \end{pmatrix}.$$

Since $\gamma_\ell = \gamma_{k^-}^{j^+} > 0$, the matrix $E(\gamma)$ is invertible, and it is easy to check that its inverse is given by

$$E(\gamma)^{-1} = \begin{pmatrix} 1 & & & -\gamma_\ell^{-1}\gamma_1 & & & \\ & \ddots & & \vdots & & & \\ & & 1 & -\gamma_\ell^{-1}\gamma_{\ell-1} & & & \\ & & & \gamma_\ell^{-1} & & & \\ & & & -\gamma_\ell^{-1}\gamma_{\ell+1} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -\gamma_\ell^{-1}\gamma_m & & & 1 \end{pmatrix},$$

which is very cheap to compute. We also have

$$A_{K^+}^{-1} = E(\gamma)^{-1} A_K^{-1}.$$

Consequently, if A_K and A_K^{-1} are available, then A_{K^+} and $A_{K^+}^{-1}$ can be computed cheaply in terms of A_K and A_K^{-1} and matrices of the form $E(\gamma)$. Then the systems $(*_\gamma)$ to find the vectors γ_K^j can be solved cheaply.

Since

$$A_{K^+}^\top = E(\gamma)^\top A_K^\top$$

and

$$(A_{K^+}^\top)^{-1} = (A_K^\top)^{-1}(E(\gamma)^\top)^{-1},$$

the matrices $A_{K^+}^\top$ and $(A_{K^+}^\top)^{-1}$ can also be computed cheaply from $A_K^\top$, $(A_K^\top)^{-1}$, and matrices of the form $E(\gamma)^\top$. Thus the systems $(*_\beta)$ to find the linear forms β_K can also be solved cheaply.

A matrix of the form $E(\gamma)$ is called an *eta matrix*; see Chvatal [40] (Chapter 7). We showed that the matrix A_{K^s} obtained after s steps of the simplex algorithm can be written as

$$A_{K^s} = A_{K^{s-1}} E_s$$

for some eta matrix E_s, so A_{k^s} can be written as the product

$$A_{K^s} = E_1 E_2 \cdots E_s$$

of s eta matrices. Such a factorization is called an *eta factorization*. The eta factorization can be used to either invert A_{K^s} or to solve a system of the form $A_{K_s}\gamma = A^{j^+}$ iteratively. Which method is more efficient depends on the sparsity of the E_i.

In summary, there are cheap methods for finding the next basic feasible solution (u^+, K^+) from (u, K). We simply wanted to give the reader a flavor of these techniques. We refer the reader to texts on linear programming for detailed presentations of methods for implementing efficiently the simplex method. In particular, the *revised simplex method* is presented in Chvatal [40], Papadimitriou and Steiglitz [134], Bertsimas and Tsitsiklis [21], and Vanderbei [181].

46.4 The Simplex Algorithm Using Tableaux

We now describe a formalism for presenting the simplex algorithm, namely *(full) tableaux*. This is the traditional formalism used in all books, modulo minor variations. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations *identical* to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

Since the quantities $c_j - c_K\gamma^j_K$ play a crucial role in determining which column A^j should come into the basis, the notation $\overline{c}_j$ is used to denote $c_j - c_K\gamma^j_K$, which is called the *reduced cost* of the variable x_j. The reduced costs actually depend on K so to be very precise we should denote them by $(\overline{c}_K)_j$, but to simplify notation we write $\overline{c}_j$ instead of $(\overline{c}_K)_j$. We will see shortly how $(\overline{c}_{K^+})_i$ is computed in terms of $(\overline{c}_K)_i$.

Observe that the data needed to execute the next step of the simplex algorithm are

(1) The current basic solution u_K and its basis $K = (k_1, \ldots, k_m)$.

(2) The reduced costs $\overline{c}_j = c_j - c_K A_K^{-1} A^j = c_j - c_K\gamma^j_K$, for all $j \notin K$.

(3) The vectors $\gamma^j_K = (\gamma^j_{k_i})_{i=1}^m$ for all $j \notin K$, that allow us to express each A^j as $A_K\gamma^j_K$.

All this information can be packed into a $(m+1)\times(n+1)$ matrix called a *(full) tableau* organized as follows:

$c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ^1_1	$\cdots$	γ^j_1	$\cdots$	γ^n_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ^1_m	$\cdots$	γ^j_m	$\cdots$	γ^n_m

It is convenient to think as the first row as Row 0, and of the first column as Column 0. Row 0 contains the current value of the objective function and the reduced costs. Column 0, except for its top entry, contains the components of the current basic solution u_K, and the remaining columns, except for their top entry, contain the vectors γ^j_K. Observe that the γ^j_K corresponding to indices j in K constitute a permutation of the identity matrix I_m. The entry $\gamma^{j^+}_{k^-}$ is called the *pivot* element. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma^j_{K^+}$, and the new reduced costs $(\overline{c}_{K^+})_j$.

If we define the $m \times n$ matrix Γ as the matrix $\Gamma = [\gamma^1_K \ \cdots \ \gamma^n_K]$ whose jth column is γ^j_K, and $\overline{c}$ as the row vector $\overline{c} = (\overline{c}_1 \ \cdots \ \overline{c}_n)$, then the above tableau is denoted concisely by

$c_K u_K$	$\overline{c}$
u_K	Γ

We now show that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref).

If $K = (k_1, \ldots, k_m)$, j^+ is the index of the incoming basis vector, $k^- = k_\ell$ is the index of the column leaving the basis, and if $K^+ = (k_1, \ldots, k_{\ell-1}, j^+, k_{\ell+1}, \ldots, k_m)$, since $A_{K^+} = A_K E(\gamma_K^{j^+})$, the new columns $\gamma_{K^+}^j$ are computed in terms of the old columns γ_K^j using $(*_\gamma)$ and the equations

$$\gamma_{K^+}^j = A_{K^+}^{-1} A^j = E(\gamma_K^{j^+})^{-1} A_K^{-1} A^j = E(\gamma_K^{j^+})^{-1} \gamma_K^j.$$

Consequently, the matrix Γ^+ is given in terms of Γ by

$$\Gamma^+ = E(\gamma_K^{j^+})^{-1} \Gamma.$$

But the matrix $E(\gamma_K^{j^+})^{-1}$ is of the form

$$E(\gamma_K^{j^+})^{-1} = \begin{pmatrix} 1 & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_1}^{j^+} & & & \\ & \ddots & & \vdots & & & \\ & & 1 & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell-1}}^{j^+} & & & \\ & & & (\gamma_{k^-}^{j^+})^{-1} & & & \\ & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell+1}}^{j^+} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_m}^{j^+} & & & 1 \end{pmatrix},$$

with the column involving the γs in the ℓth column, and Γ^+ is obtained by applying the following elementary row operations to Γ:

1. Multiply Row ℓ by $1/\gamma_{k^-}^{j^+}$ (the inverse of the pivot) to make the entry on Row ℓ and Column j^+ equal to 1.

2. Subtract $\gamma_{k_i}^{j^+} \times$ (the normalized) Row ℓ from Row i, for $i = 1, \ldots, \ell - 1, \ell + 1, \ldots, m$.

These are *exactly* the elementary row operations that reduce the ℓth column $\gamma_K^{j^+}$ of Γ to the ℓth column of the identity matrix I_m. Thus, this step is identical to the sequence of steps that the procedure to convert a matrix to row reduced echelon from executes on the ℓth column of the matrix. The only difference is the criterion for the choice of the pivot.

Since the new basic solution u_{K^+} is given by $u_{K^+} = A_{K^+}^{-1} b$, we have

$$u_{K^+} = E(\gamma_K^{j^+})^{-1} A_K^{-1} b = E(\gamma_K^{j^+})^{-1} u_K.$$

This means that u_+ is obtained from u_K by applying exactly the *same* elementary row operations that were applied to Γ. Consequently, just as in the procedure for reducing a matrix to rref, we can apply elementary row operations to the matrix $[u_k \; \Gamma]$, which consists of rows $1, \ldots, m$ of the tableau.

Once the new matrix Γ^+ is obtained, the new reduced costs are given by the following proposition.

Proposition 46.2. *Given any Linear Program $(P2)$ in standard form*

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax = b \;\textit{and}\; x \geq 0, \end{array}$$

where A is an $m \times n$ matrix of rank m, if (u, K) is a basic (not necessarily feasible) solution of $(P2)$ and if $K^+ = (K - \{k^-\}) \cup \{j^+\}$, with $K = (k_1, \ldots, k_m)$ and $k^- = k_\ell$, then for $i = 1, \ldots, n$ we have

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K).$$

Using the reduced cost notation, the above equation is

$$(\overline{c}_{K^+})_i = (\overline{c}_K)_i - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(\overline{c}_K)_{j^+}.$$

Proof. Without any loss of generality and to simplify notation assume that $K = (1, \ldots, m)$ and write j for j^+ and ℓ for k_m. Since $\gamma^i_K = A_K^{-1}A^i$, $\gamma^i_{K^+} = A^{-1}_{K^+}A^i$, and $A_{K^+} = A_K E(\gamma^j_K)$, we have

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_{K^+}A^{-1}_{K^+}A^i = c_i - c_{K^+}E(\gamma^j_K)^{-1}A_K^{-1}A^i = c_i - c_{K^+}E(\gamma^j_K)^{-1}\gamma^i_K,$$

where

$$E(\gamma^j_K)^{-1} = \begin{pmatrix} 1 & & & -(\gamma^j_\ell)^{-1}\gamma^j_1 & & & \\ & \ddots & & \vdots & & & \\ & & 1 & -(\gamma^j_\ell)^{-1}\gamma^j_{\ell-1} & & & \\ & & & (\gamma^j_\ell)^{-1} & & & \\ & & & -(\gamma^j_\ell)^{-1}\gamma^j_{\ell+1} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -(\gamma^j_\ell)^{-1}\gamma^j_m & & & 1 \end{pmatrix}$$

where the ℓth column contains the γs. Since $c_{K^+} = (c_1, \ldots, c_{\ell-1}, c_j, c_{\ell+1}, \ldots, c_m)$, we have

$$c_{K^+}E(\gamma^j_K)^{-1} = \left(c_1, \ldots, c_{\ell-1}, \frac{c_j}{\gamma^j_\ell} - \sum_{k=1, k\neq \ell}^{m} c_k\frac{\gamma^j_k}{\gamma^j_\ell}, c_{\ell+1}, \ldots, c_m\right),$$

and

$$
\begin{aligned}
c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i &= \begin{pmatrix} c_1 & \dots & c_{\ell-1} & \dfrac{c_j}{\gamma_\ell^j} - \displaystyle\sum_{k=1,k\neq\ell}^m c_k\frac{\gamma_k^j}{\gamma_\ell^j} & c_{\ell+1} & \dots & c_m \end{pmatrix} \begin{pmatrix} \gamma_1^i \\ \vdots \\ \gamma_{\ell-1}^i \\ \gamma_\ell^i \\ \gamma_{\ell+1}^i \\ \vdots \\ \gamma_m^i \end{pmatrix} \\
&= \sum_{k=1,k\neq\ell}^m c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1,k\neq\ell}^m c_k\gamma_k^j\right) \\
&= \sum_{k=1,k\neq\ell}^m c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j + c_\ell\gamma_\ell^j - \sum_{k=1}^m c_k\gamma_k^j\right) \\
&= \sum_{k=1}^m c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1}^m c_k\gamma_k^j\right) \\
&= c_K\gamma_K^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),
\end{aligned}
$$

and thus

$$
c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i = c_i - c_K\gamma_K^i - \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),
$$

as claimed. $\square$

Since $(\gamma_{k^-}^1, \dots, \gamma_{k^-}^n)$ is the ℓth row of Γ, we see that Proposition 46.2 shows that

$$
\overline{c}_{K^+} = \overline{c}_K - \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\,\Gamma_\ell, \tag{$\dagger$}
$$

where Γ_ℓ denotes the ℓ-th row of Γ and $\gamma_{k^-}^{j^+}$ is the pivot. This means that $\overline{c}_{K^+}$ is obtained by the elementary row operations which consist of first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then subtracting $(\overline{c}_K)_{j^+}\times$ the normalized Row ℓ from $\overline{c}_K$. *These are exactly the row operations that make the reduced cost $(\overline{c}_K)_{j^+}$ zero.*

Remark: It easy easy to show that we also have

$$
\overline{c}_{K^+} = c - c_{K^+}\Gamma^+.
$$

We saw in Section 46.2 that the change in the objective function after a pivoting step during which column j^+ comes in and column k^- leaves is given by

$$\theta^{j^+}\left(c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k\right) = \theta^{j^+}(\overline{c}_K)_{j^+},$$

where

$$\theta^{j^+} = \frac{u_{k^-}}{\gamma_{k^-}^{j^+}}.$$

If we denote the value of the objective function $c_K u_K$ by z_K, then we see that

$$z_{K^+} = z_K + \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\, u_{k^-}.$$

This means that the new value z_{K^+} of the objective function is obtained by first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then adding $(\overline{c}_K)_{j^+}\times$ the zeroth entry of the normalized ℓth line by $(\overline{c}_K)_{j^+}$ to the zeroth entry of line 0.

In updating the reduced costs, we subtract rather than add $(\overline{c}_K)_{j^+}\times$ the normalized row ℓ from row 0. This suggests storing $-z_K$ as the zeroth entry on line 0 rather than z_K, because then all the entries row 0 are updated by the *same* elementary row operations. Therefore, from now on, we use tableau of the form

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The simplex algorithm first chooses the incoming column j^+ by picking some column for which $\overline{c}_j > 0$, and then chooses the outgoing column k^- by considering the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+), and picking k^- to achieve the minimum of these ratios.

Here is an illustration of the simplex algorithm using elementary row operations on an example from Papadimitriou and Steiglitz [134] (Section 2.9).

Example 46.4. Consider the linear program

$$\begin{aligned}
&\text{maximize} \quad -2x_2 - x_4 - 5x_7\\
&\text{subject to}\\
&\qquad x_1 + x_2 + x_3 + x_4 = 4\\
&\qquad x_1 + x_5 = 2\\
&\qquad x_3 + x_6 = 3\\
&\qquad 3x_2 + x_3 + x_7 = 6\\
&\qquad x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geq 0.
\end{aligned}$$

We have the basic feasible solution $u = (0, 0, 0, 4, 2, 3, 6)$, with $K = (4, 5, 6, 7)$. Since $c_K = (-1, 0, 0, -5)$ and $c = (0, -2, 0, -1, 0, 0 - 5)$ the first tableau is

34	1	14	6	0	0	0	0
$u_4 = 4$	1	1	1	1	0	0	0
$u_5 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Since $\overline{c}_j = c_j - c_K \gamma_K^j$, Row 0 is obtained by subtracting $-1\times$ Row 1 and $-5\times$ Row 4 from $c = (0, -2, 0, -1, 0, \ 0, -5)$. Let us pick Column $j^+ = 1$ as the incoming column. We have the ratios (for positive entries on Column 1)

$$4/1, \ 2/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 5$. The pivot 1 is indicated in red. The new basis is $K = (4, 1, 6, 7)$. Next we apply row operations to reduce Column 1 to the second vector of the identity matrix I_4. For this, we subtract Row 2 from Row 1. We get the tableau

34	1	14	6	0	0	0	0
$u_4 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

To compute the new reduced costs, we want to set $\overline{c}_1$ to 0, so we apply the identical row operations and subtract Row 2 from Row 0 to obtain the tableau

32	0	14	6	0	-1	0	0
$u_4 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Next, pick Column $j^+ = 3$ as the incoming column. We have the ratios (for positive entries on Column 3)

$$2/1, \ 3/1, \ 6/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 4$. The pivot 1 is indicated in red and the new basis is $K = (3, 1, 6, 7)$. Next we apply row operations to reduce Column 3 to the first vector of the identity matrix I_4. For this, we subtract Row 1 from Row 3 and from Row 4 and obtain the tableau:

32	0	14	6	0	-1	0	0
$u_3 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

To compute the new reduced costs, we want to set $\bar{c}_3$ to 0, so we subtract $6\times$ Row 1 from Row 0 to get the tableau

20	0	8	0	-6	5	0	0
$u_3 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

Next we pick $j^+ = 2$ as the incoming column. We have the ratios (for positive entries on Column 2)

$$2/1,\ 4/2,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 3$. The pivot 1 is indicated in red and the new basis is $K = (2, 1, 6, 7)$. Next we apply row operations to reduce Column 2 to the first vector of the identity matrix I_4. For this, we add Row 1 to Row 3 and subtract $2\times$ Row 1 from Row 4 to obtain the tableau:

20	0	8	0	-6	5	0	0
$u_2 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	-2	-3	3	0	1

To compute the new reduced costs, we want to set $\bar{c}_2$ to 0, so we subtract $8\times$ Row 1 from Row 0 to get the tableau

4	0	0	-8	-14	13	0	0
$u_2 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	-2	-3	3	0	1

The only possible incoming column corresponds to $j^+ = 5$. We have the ratios (for positive entries on Column 5)

$$2/1,\ 0/3,$$

and since the minimum is 0, we pick the outgoing column to be Column $k^- = 7$. The pivot 3 is indicated in red and the new basis is $K = (2, 1, 6, 5)$. Since the minimum is 0, the basis $K = (2, 1, 6, 5)$ is degenerate (indeed, the component corresponding to the index 5 is 0). Next we apply row operations to reduce Column 5 to the fourth vector of the identity matrix I_4. For this, we multiply Row 4 by 1/3, and then add the normalized Row 4 to Row 1 and subtract the normalized Row 4 from Row 2 to obtain the tableau:

4	0	0	-8	-14	13	0	0
$u_2 = 2$	0	1	$1/3$	0	0	0	$1/3$
$u_1 = 2$	1	0	$2/3$	1	0	0	$-1/3$
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	$-2/3$	-1	1	0	$1/3$

To compute the new reduced costs, we want to set $\bar{c}_5$ to 0, so we subtract $13\times$ Row 4 from Row 0 to get the tableau

4	0	0	$2/3$	-1	0	0	$-13/3$
$u_2 = 2$	0	1	$1/3$	0	0	0	$1/3$
$u_1 = 2$	1	0	$2/3$	1	0	0	$-1/3$
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	$-2/3$	-1	1	0	$1/3$

The only possible incoming column corresponds to $j^+ = 3$. We have the ratios (for positive entries on Column 3)

$$2/(1/3) = 6, \; 2/(2/3) = 3, \; 3/1 = 3,$$

and since the minimum is 3, we pick the outgoing column to be Column $k^- = 1$. The pivot 2/3 is indicated in red and the new basis is $K = (2, 3, 6, 5)$. Next we apply row operations to reduce Column 3 to the second vector of the identity matrix I_4. For this, we multiply Row 2 by 3/2, subtract $(1/3)\times$ (normalized Row 2) from Row 1, and subtract normalized Row 2 from Row 3, and add $(2/3)\times$ (normalized Row 2) to Row 4 to obtain the tableau:

4	0	0	$2/3$	-1	0	0	$-13/3$
$u_2 = 1$	$-1/2$	1	0	$-1/2$	0	0	$1/2$
$u_3 = 3$	$3/2$	0	1	$3/2$	0	0	$-1/2$
$u_6 = 0$	$-3/2$	0	0	$-3/2$	0	1	$1/2$
$u_5 = 2$	1	0	0	0	1	0	0

To compute the new reduced costs, we want to set $\bar{c}_3$ to 0, so we subtract $(2/3)\times$ Row 2 from Row 0 to get the tableau

2	-1	0	0	-2	0	0	-4
$u_2 = 1$	$-1/2$	1	0	$-1/2$	0	0	$1/2$
$u_3 = 3$	$3/2$	0	1	$3/2$	0	0	$-1/2$
$u_6 = 0$	$-3/2$	0	0	$-3/2$	0	1	$1/2$
$u_5 = 2$	1	0	0	0	1	0	0

Since all the reduced cost are ≤ 0, we have reached an optimal solution, namely $(0, 1, 3, 0, 2, 0, 0, 0)$, with optimal value -2.

The progression of the simplex algorithm from one basic feasible solution to another corresponds to the visit of vertices of the polyhedron $\mathcal{P}$ associated with the constraints of the linear program illustrated in Figure 46.4.

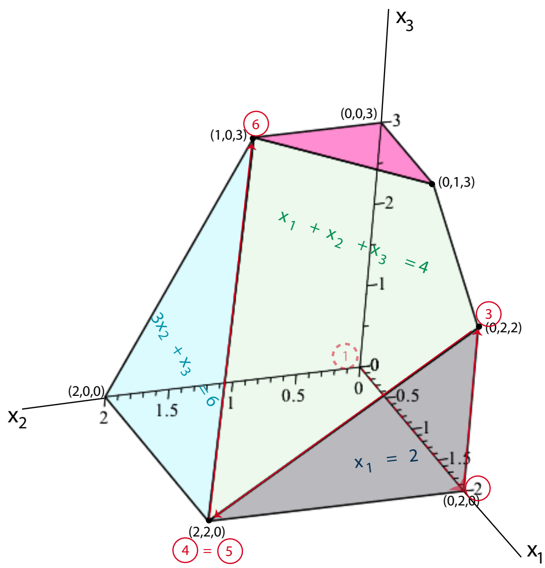


Figure 46.4: The polytope $\mathcal{P}$ associated with the linear program optimized by the tableau method. The red arrowed path traces the progression of the simplex method from the origin to the vertex $(0, 1, 3)$.

As a final comment, if it is necessary to run Phase I of the simplex algorithm, in the event that the simplex algorithm terminates with an optimal solution $(u^*, 0_m)$ and a basis K^* such that some $u_i = 0$, then the basis K^* contains indices of basic columns A^j corresponding to slack variables that need to be *driven out* of the basis. This is easy to achieve by performing a pivoting step involving some other column j^+ corresponding to one of the original variables (not a slack variable) for which $(\gamma_{K^*})_i^{j^+} \neq 0$. In such a step, it doesn't matter whether $(\gamma_{K^*})_i^{j^+} < 0$ or $(\overline{c}_{K^*})_{j^+} \leq 0$. If the original matrix A has no redundant equations, such a step is always possible. Otherwise, $(\gamma_{K^*})_i^j = 0$ for all non-slack variables, so we detected that the ith equation is redundant and we can delete it.

Other presentations of the tableau method can be found in Bertsimas and Tsitsiklis [21] and Papadimitriou and Steiglitz [134].

46.5 Computational Efficiency of the Simplex Method

Let us conclude with a few comments about the efficiency of the simplex algorithm. In *practice*, it was observed by Dantzig that for linear programs with $m < 50$ and $m+n < 200$, the simplex algorithms typically requires less than $3m/2$ iterations, but at most $3m$ iterations. This fact agrees with more recent empirical experiments with much larger programs that show that the number iterations is bounded by $3m$. Thus, it was somewhat of a shock in 1972 when Klee and Minty found a linear program with n variables and n equations for which the simplex algorithm with Dantzig's pivot rule requires requires $2^n - 1$ iterations. This program (taken from Chvatal [40], page 47) is reproduced below:

$$\begin{aligned}
&\text{maximize} \quad \sum_{j=1}^{n} 10^{n-j} x_j \\
&\text{subject to} \\
&\quad \left(2 \sum_{j=1}^{i-1} 10^{i-j} x_j\right) + x_i \leq 100^{i-1} \\
&\quad x_j \geq 0,
\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

If $p = \max(m, n)$, then, in terms of worse case behavior, for all currently known pivot rules, the simplex algorithm has exponential complexity in p. However, as we said earlier, in practice, nasty examples such as the Klee–Minty example seem to be rare, and the number of iterations appears to be linear in m.

Whether or not a pivot rule (a clairvoyant rule) for which the simplex algorithms runs in polynomial time in terms of m is still an *open problem*.

The *Hirsch conjecture* claims that there is some pivot rule such that the simplex algorithm finds an optimal solution in $O(p)$ steps. The best bound known so far due to Kalai and Kleitman is $m^{1+\ln n} = (2n)^{\ln m}$. For more on this topic, see Matousek and Gardner [123] (Section 5.9) and Bertsimas and Tsitsiklis [21] (Section 3.7).

Researchers have investigated the problem of finding upper bounds on the expected number of pivoting steps if a randomized pivot rule is used. Bounds better than 2^m (but of course, not polynomial) have been found.

Understanding the complexity of linear programing, in particular of the simplex algorithm, is still ongoing. The interested reader is referred to Matousek and Gardner [123] (Chapter 5, Section 5.9) for some pointers.

In the next section we consider important theoretical criteria for determining whether a set of constraints $Ax \leq b$ and $x \geq 0$ has a solution or not.

46.6 Summary

The main concepts and results of this chapter are listed below:

- Degenerate and nondegenerate basic feasible solution.
- Pivoting step.
- Pivot rule.
- Cycling.
- Bland's rule, Dantzig's rule, steepest edge rule, random edge rule, largest increase rule, lexicographic rule.
- Phase I and Phase II of the simplex algorithm.
- eta matrix, eta factorization.
- Revised simplex method.
- Reduced cost.
- Full tableaux.
- The Hirsch conjecture.

46.7 Problems

Problem 46.1. In Section 46.2 prove that if Case (A) arises, then the basic feasible solution u is an optimal solution. Prove that if Case (B1) arises, then the linear program is unbounded. Prove that if Case (B3) arises, then (u^+, K^+) is a basic feasible solution.

Problem 46.2. In Section 46.2 prove that the following equivalences hold:

$$\begin{aligned}
&\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset\\
&\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset\\
&\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset\\
&\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset.
\end{aligned}$$

Furthermore, prove that Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

Problem 46.3. Consider the linear program (due to E.M.L. Beale):

$$\begin{aligned} &\text{maximize} \quad (3/4)x_1 - 150x_2 + (1/50)x_3 - 6x_4 \\ &\text{subject to} \\ &\qquad (1/4)x_1 - 60x_2 - (1/25)x_3 + 9x_4 \leq 0 \\ &\qquad (1/4)x_1 - 90x_2 - (1/50)x_3 + 3x_4 \leq 0 \\ &\qquad x_3 \leq 1 \\ &\qquad x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

(1) Convert the above program to standard form.

(2) Show that if we apply the simplex algorithm with the pivot rule which selects the column entering the basis as the column of smallest index, then the method cycles.

Problem 46.4. Read carefully the proof given by Chvatal that the lexicographic pivot rule and Bland's pivot rule prevent cycling; see Chvatal [40] (Chapter 3, pages 34-38).

Problem 46.5. Solve the following linear program (from Chvatal [40], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} \quad 3x_1 + x_2 \\ &\text{subject to} \\ &\qquad x_1 - x_2 \leq -1 \\ &\qquad -x_1 - x_2 \leq -3 \\ &\qquad 2x_1 + x_2 \leq 4 \\ &\qquad x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 46.6. Solve the following linear program (from Chvatal [40], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} \quad 3x_1 + x_2 \\ &\text{subject to} \\ &\qquad x_1 - x_2 \leq -1 \\ &\qquad -x_1 - x_2 \leq -3 \\ &\qquad 2x_1 + x_2 \leq 2 \\ &\qquad x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 46.7. Solve the following linear program (from Chvatal [40], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned}
\text{maximize} \quad & 3x_1 + x_2 \\
\text{subject to} \quad & \\
& x_1 - x_2 \leq -1 \\
& -x_1 - x_2 \leq -3 \\
& 2x_1 - x_2 \leq 2 \\
& x_1 \geq 0, \; x_2 \geq 0.
\end{aligned}$$

Problem 46.8. Show that the following linear program (from Chvatal [40], Chapter 3, page 43) is unbounded.

$$\begin{aligned}
\text{maximize} \quad & x_1 + 3x_2 - x_3 \\
\text{subject to} \quad & \\
& 2x_1 + 2x_2 - x_3 \leq 10 \\
& 3x_1 - 2x_2 + x_3 \leq 10 \\
& x_1 - 3x_2 + x_3 \leq 10 \\
& x_1 \geq 0, \; x_2 \geq 0, \; x_3 \geq 0.
\end{aligned}$$

Hint. Try $x_1 = 0, x_3 = t$, and a suitable value for x_2.

Chapter 47

Linear Programming and Duality

47.1 Variants of the Farkas Lemma

If A is an $m \times n$ matrix and if $b \in \mathbb{R}^m$ is a vector, it is known from linear algebra that the linear system $Ax = b$ has no solution iff there is some linear form $y \in (\mathbb{R}^m)^*$ such that $yA = 0$ and $yb \neq 0$. This means that the linear from y vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form y defines the linear hyperplane H of equation $yz = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a linear hyperplane H containing $A^1, \ldots, A^n$ and not containing b. This is a kind of separation theorem that says that the vectors $A^1, \ldots, A^n$ and b can be separated by some linear hyperplane H.

What we would like to do is to generalize this kind of criterion, first to a system $Ax = b$ subject to the constraints $x \geq 0$, and next to sets of inequality constraints $Ax \leq b$ and $x \geq 0$. There are indeed such criteria going under the name of *Farkas lemma*.

The key is a separation result involving polyhedral cones known as the Farkas–Minkowski proposition. We have the following fundamental separation lemma.

Proposition 47.1. *Let $C \subseteq \mathbb{R}^n$ be a closed nonempty (convex) cone. For any point $a \in \mathbb{R}^n$, if $a \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*

2. *$a \notin H$*

3. *a lies in the other half-space determined by H.*

We say that H strictly separates C and a.

Proposition 47.1, which is illustrated in Figure 47.1, is an easy consequence of another separation theorem that asserts that given any two nonempty closed convex sets A and B of $\mathbb{R}^n$ with A compact, there is a hyperplane H strictly separating A and B (which means that $A \cap H = \emptyset$, $B \cap H = \emptyset$, that A lies in one of the two half-spaces determined by H, and B lies in the other half-space determined by H); see Gallier [72] (Chapter 7, Corollary 7.4 and Proposition 7.3). This proof is nontrivial and involves a geometric version of the Hahn–Banach theorem.

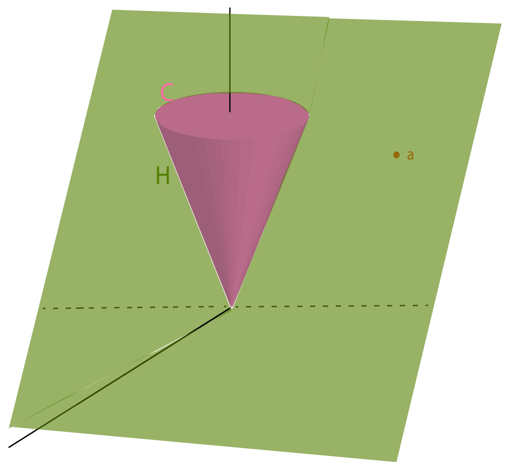


Figure 47.1: In $\mathbb{R}^3$, the olive green hyperplane H separates the cone C from the orange point a.

The Farkas–Minkowski proposition is Proposition 47.1 applied to a polyhedral cone

$$C = \{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \geq 0,\ i = 1, \ldots, n\}$$

where $\{a_1, \ldots, a_n\}$ is a *finite* number of vectors $a_i \in \mathbb{R}^n$. By Proposition 44.2, any polyhedral cone is closed, so Proposition 47.1 applies and we obtain the following separation lemma.

Proposition 47.2. *(Farkas–Minkowski) Let $C \subseteq \mathbb{R}^n$ be a nonempty polyhedral cone $C = \mathrm{cone}(\{a_1, \ldots, a_n\})$. For any point $b \in \mathbb{R}^n$, if $b \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*

2. *$b \notin H$*

3. *b lies in the other half-space determined by H.*

Equivalently, there is a nonzero linear form $y \in (\mathbb{R}^n)^*$ such that

1. *$ya_i \geq 0$ for $i = 1, \ldots, n$.*

2. *$yb < 0$.*

A direct proof of the Farkas–Minkowski proposition not involving Proposition 47.1 is given at the end of this section.

Remark: There is a generalization of the Farkas–Minkowski proposition applying to infinite dimensional real Hilbert spaces; see Theorem 48.12 (or Ciarlet [41], Chapter 9).

Proposition 47.2 implies our first version of Farkas' lemma.

Proposition 47.3. *(Farkas Lemma, Version I) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. First assume that there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0$ and $yb < 0$. If $x \geq 0$ is a solution of $Ax = b$, then we get

$$yAx = yb,$$

but if $yA \geq 0$ and $x \geq 0$, then $yAx \geq 0$, and yet by hypothesis $yb < 0$, a contradiction.

Next assume that $Ax = b$ has no solution $x \geq 0$. This means that b does not belong to the polyhedral cone $C = \mathrm{cone}(\{A^1, \ldots, A^n\})$ spanned by the columns of A. By Proposition 47.2, there is a nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

1. $yA^j \geq 0$ for $j = 1, \ldots, n$.
2. $yb < 0$,

which says that $yA \geq 0_n^\top$ and $yb < 0$. $\square$

Next consider the solvability of a system of inequalities of the form $Ax \leq b$ and $x \geq 0$.

Proposition 47.4. *(Farkas Lemma, Version II) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The system of inequalities $Ax \leq b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. We use the trick of linear programming which consists of adding "slack variables" z_i to convert inequalities $a_i x \leq b_i$ into equations $a_i x + z_i = b_i$ with $z_i \geq 0$ already discussed just before Definition 44.9. If we let $z = (z_1, \ldots, z_m)$, it is obvious that the system $Ax \leq b$ has a solution $x \geq 0$ iff the equation

$$\begin{pmatrix} A & I_m \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = b$$

has a solution $\begin{pmatrix} x \\ z \end{pmatrix}$ with $x \geq 0$ and $z \geq 0$. Now by Farkas I, the above system has no solution with with $x \geq 0$ and $z \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

$$y \begin{pmatrix} A & I_m \end{pmatrix} \geq 0_{n+m}^\top$$

and $yb < 0$, that is, $yA \geq 0_n^\top$, $y \geq 0_m^\top$ and $yb < 0$. $\square$

In the next section we use Farkas II to prove the duality theorem in linear programming. Observe that by taking the negation of the equivalence in Farkas II we obtain a criterion of solvability, namely:

The system of inequalities $Ax \leq b$ has a solution $x \geq 0$ iff for every nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, if $yA \geq 0_n^\top$, then $yb \geq 0$.

We now prove the Farkas–Minkowski proposition without using Proposition 47.1. This approach uses a basic property of the distance function from a point to a closed set.

Definition 47.1. Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. The *distance* $d(a, X)$ from a to X is defined as

$$d(a, X) = \inf_{x \in X} \|a - x\|.$$

Here, $\| \; \|$ denotes the Euclidean norm.

Proposition 47.5. *Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. If X is closed, then there is some $z \in X$ such that $\|a - z\| = d(a, X)$.*

Proof. Since X is nonempty, pick any $x_0 \in X$, and let $r = \|a - x_0\|$. If $B_r(a)$ is the closed ball $B_r(a) = \{x \in \mathbb{R}^n \mid \|x - a\| \leq r\}$, then clearly

$$d(a, X) = \inf_{x \in X} \|a - x\| = \inf_{x \in X \cap B_r(a)} \|a - x\|.$$

Since $B_r(a)$ is compact and X is closed, $K = X \cap B_r(a)$ is also compact. But the function $x \mapsto \|a - x\|$ defined on the compact set K is continuous, and the image of a compact set by a continuous function is compact, so by Heine–Borel it has a minimum that is achieved by some $z \in K \subseteq X$. $\square$

Remark: If U is a nonempty, closed and convex subset of a Hilbert space V, a standard result of Hilbert space theory (the projection lemma, see Proposition 48.5) asserts that for any $v \in V$ there is a *unique* $p \in U$ such that

$$\|v - p\| = \inf_{u \in U} \|v - u\| = d(v, U),$$

and

$$\langle p - v, u - p \rangle \geq 0 \quad \text{for all } u \in U.$$

Here $\|w\| = \sqrt{\langle w, w \rangle}$, where $\langle -, - \rangle$ is the inner product of the Hilbert space V.

We can now give a proof of the Farkas–Minkowski proposition (Proposition 47.2) that does not use Proposition 47.1. This proof is adapted from Matousek and Gardner [123] (Chapter 6, Sections 6.5).

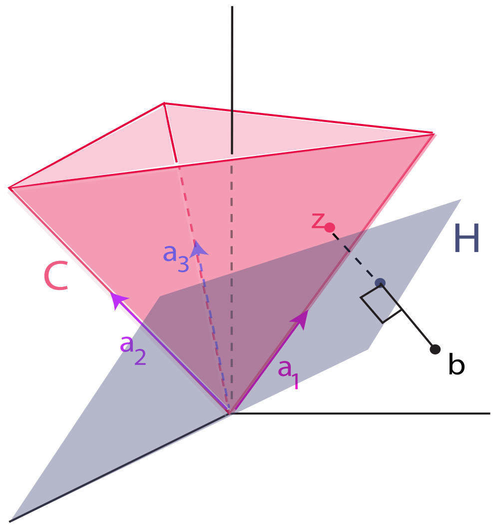


Figure 47.2: The hyperplane H, perpendicular to $z - b$, separates the point b from $C = \mathrm{cone}(\{a_1, a_2, a_3\})$.

Proof of the Farkas–Minkowski proposition. Let $C = \mathrm{cone}(\{a_1, \ldots, a_m\})$ be a polyhedral cone (nonempty) and assume that $b \notin C$. By Proposition 44.2, the polyhedral cone is closed, and by Proposition 47.5 there is some $z \in C$ such that $d(b, C) = \|b - z\|$; that is, z is a point of C closest to b. Since $b \notin C$ and $z \in C$ we have $u = z - b \neq 0$, and we claim that the linear hyperplane H orthogonal to u does the job, as illustrated in Figure 47.2.

First let us show that

$$\langle u, z \rangle = \langle z - b, z \rangle = 0. \tag{$*_1$}$$

This is trivial if $z = 0$, so assume $z \neq 0$. If $\langle u, z \rangle \neq 0$, then either $\langle u, z \rangle > 0$ or $\langle u, z \rangle < 0$. In either case we show that we can find some point $z' \in C$ closer to b than z is, a contradiction.

Case 1: $\langle u, z \rangle > 0$.

Let $z' = (1 - \alpha)z$ for any α such that $0 < \alpha < 1$. Then $z' \in C$ and since $u = z - b$,

$$z' - b = (1 - \alpha)z - (z - u) = u - \alpha z,$$

so

$$\|z' - b\|^2 = \|u - \alpha z\|^2 = \|u\|^2 - 2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 .$$

If we pick $\alpha > 0$ such that $\alpha < 2\langle u, z \rangle / \|z\|^2$, then $-2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, contradicting the fact that z is a point of C closest to b.

Case 2: $\langle u, z \rangle < 0$.

Let $z' = (1 + \alpha)z$ for any α such that $\alpha \geq -1$. Then $z' \in C$ and since $u = z - b$, we have $z' - b = (1 + \alpha)z - (z - u) = u + \alpha z$ so

$$\|z' - b\|^2 = \|u + \alpha z\|^2 = \|u\|^2 + 2\alpha\langle u, z \rangle + \alpha^2 \|z\|^2 ,$$

and if
$$0 < \alpha < -2\langle u, z\rangle / \|z\|^2,$$
then $2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, a contradiction as above.

Therefore $\langle u, z\rangle = 0$. We have

$$\langle u, u\rangle = \langle u, z - b\rangle = \langle u, z\rangle - \langle u, b\rangle = -\langle u, b\rangle,$$

and since $u \neq 0$, we have $\langle u, u\rangle > 0$, so $\langle u, u\rangle = -\langle u, b\rangle$ implies that

$$\langle u, b\rangle < 0. \tag{$*_2$}$$

It remains to prove that $\langle u, a_i\rangle \geq 0$ for $i = 1, \ldots, m$. Pick any $x \in C$ such that $x \neq z$. We claim that
$$\langle b - z, x - z\rangle \leq 0. \tag{$*_3$}$$
Otherwise $\langle b - z, x - z\rangle > 0$, that is, $\langle z - b, x - z\rangle < 0$, and we show that we can find some point $z' \in C$ on the line segment $[z, x]$ closer to b than z is.

For any α such that $0 \leq \alpha \leq 1$, we have $z' = (1 - \alpha)z + \alpha x = z + \alpha(x - z) \in C$, and since $z' - b = z - b + \alpha(x - z)$ we have

$$\|z' - b\|^2 = \|z - b + \alpha(x - z)\|^2 = \|z - b\|^2 + 2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2,$$

so for any $\alpha > 0$ such that

$$\alpha < -2\langle z - b, x - z\rangle / \|x - z\|^2,$$

we have $2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2 < 0$, which implies that $\|z' - b\|^2 < \|z - b\|^2$, contradicting that z is a point of C closest to b.

Since $\langle b - z, x - z\rangle \leq 0$, $u = z - b$, and by $(*_1)$, $\langle u, z\rangle = 0$, we have

$$0 \geq \langle b - z, x - z\rangle = \langle -u, x - z\rangle = -\langle u, x\rangle + \langle u, z\rangle = -\langle u, x\rangle,$$

which means that
$$\langle u, x\rangle \geq 0 \quad \text{for all } x \in C, \tag{$*_3$}$$
as claimed. In particular,
$$\langle u, a_i\rangle \geq 0 \quad \text{for } i = 1, \ldots, m. \tag{$*_4$}$$
Then by $(*_2)$ and $(*_4)$, the linear form defined by $y = u^\top$ satisfies the properties $yb < 0$ and $ya_i \geq 0$ for $i = 1, \ldots, m$, which proves the Farkas–Minkowski proposition. $\square$

There are other ways of proving the Farkas–Minkowski proposition, for instance using minimally infeasible systems or Fourier–Motzkin elimination; see Matousek and Gardner [123] (Chapter 6, Sections 6.6 and 6.7).

47.2 The Duality Theorem in Linear Programming

Let (P) be the linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \ \text{and} \ x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix, and assume that (P) has a feasible solution and is bounded above. Since by hypothesis the objective function $x \mapsto cx$ is bounded on $\mathcal{P}(A, b)$, it might be useful to deduce an *upper bound* for cx from the inequalities $Ax \leq b$, for any $x \in \mathcal{P}(A, b)$. We can do this as follows: for every inequality

$$a_i x \leq b_i \quad 1 \leq i \leq m,$$

pick a nonnegative scalar y_i, multiply both sides of the above inequality by y_i obtaining

$$y_i a_i x \leq y_i b_i \quad 1 \leq i \leq m,$$

(the direction of the inequality is preserved since $y_i \geq 0$), and then add up these m equations, which yields

$$(y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m.$$

If we can pick the $y_i \geq 0$ such that

$$c \leq y_1 a_1 + \cdots + y_m a_m,$$

then since $x_j \geq 0$, we have

$$cx \leq (y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m,$$

namely we found an upper bound of the value cx of the objective function of (P) for any feasible solution $x \in \mathcal{P}(A, b)$. If we let y be the linear form $y = (y_1, \ldots, y_m)$, then since

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

$y_1 a_1 + \cdots + y_m a_m = yA$, and $y_1 b_1 + \cdots + y_m b_m = yb$, what we did was to look for some $y \in (\mathbb{R}^m)^*$ such that

$$c \leq yA, \quad y \geq 0,$$

so that we have

$$cx \leq yb \quad \text{for all } x \in \mathcal{P}(A, b). \tag{$*$}$$

Then it is natural to look for a "best" value of yb, namely a minimum value, which leads to the definition of the *dual* of the linear program (P), a notion due to John von Neumann.

Definition 47.2. Given any Linear Program (P)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, the *dual* (D) of (P) is the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

The variables $y_1, \ldots, y_m$ are called the *dual variables*. The original Linear Program (P) is called the *primal* linear program and the original variables $x_1, \ldots, x_n$ are the *primal variables*.

Here is an explicit example of a linear program and its dual.

Example 47.1. Consider the linear program illustrated by Figure 47.3

$$\begin{aligned} &\text{maximize} \quad 2x_1 + 3x_2 \\ &\text{subject to} \\ &\qquad 4x_1 + 8x_2 \leq 12 \\ &\qquad 2x_1 + x_2 \leq 3 \\ &\qquad 3x_1 + 2x_2 \leq 4 \\ &\qquad x_1 \geq 0,\; x_2 \geq 0. \end{aligned}$$

Its dual linear program is illustrated in Figure 47.4

$$\begin{aligned} &\text{minimize} \quad 12y_1 + 3y_2 + 4y_3 \\ &\text{subject to} \\ &\qquad 4y_1 + 2y_2 + 3y_3 \geq 2 \\ &\qquad 8y_1 + y_2 + 2y_3 \geq 3 \\ &\qquad y_1 \geq 0,\; y_2 \geq 0,\; y_3 \geq 0. \end{aligned}$$

It can be checked that $(x_1, x_2) = (1/2, 5/4)$ is an optimal solution of the primal linear program, with the maximum value of the objective function $2x_1 + 3x_2$ equal to $19/4$, and that $(y_1, y_2, y_3) = (5/16, 0, 1/4)$ is an optimal solution of the dual linear program, with the minimum value of the objective function $12y_1 + 3y_2 + 4y_3$ also equal to $19/4$.

Observe that in the Primal Linear Program (P), we are looking for a *vector* $x \in \mathbb{R}^n$ maximizing the form cx, and that the constraints are determined by the action of the *rows* of the matrix A on x. On the other hand, in the Dual Linear Program (D), we are looking for a *linear form* $y \in (\mathbb{R}^*)^m$ minimizing the form yb, and the constraints are determined by

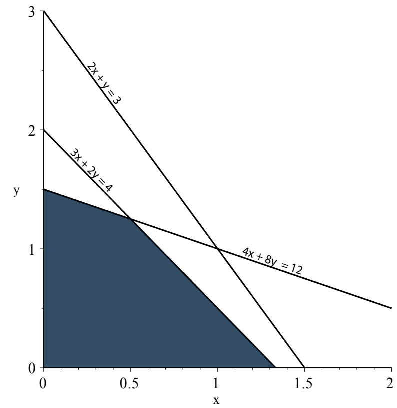


Figure 47.3: The $\mathcal{H}$-polytope for the linear program of Example 47.1. Note $x_1 \to x$ and $x_2 \to y$.

the action of y on the *columns* of A. This is the sense in which (D) is the *dual* (P). In most presentations, the fact that (P) and (D) perform a search for a solution in spaces that are dual to each other is obscured by excessive use of transposition.

To convert the Dual Program (D) to a standard maximization problem we change the objective function yb to $-b^\top y^\top$ and the inequality $yA \geq c$ to $-A^\top y^\top \leq -c^\top$. The Dual Linear Program (D) is now stated as (D')

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that the dual in maximization form (D'') of the Dual Program (D') gives back the Primal Program (P).

The above discussion established the following inequality known as *weak duality*.

Proposition 47.6. *(Weak Duality) Given any Linear Program (P)*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, for any feasible solution $x \in \mathbb{R}^n$ of the Primal Problem (P) and every feasible solution $y \in (\mathbb{R}^m)^*$ of the Dual Problem (D), we have

$$cx \leq yb.$$

Definition 47.3. We say that the Dual Linear Program (D) is *bounded below* if $\{yb \mid y^\top \in \mathcal{P}(-A^\top, -c^\top)\}$ is bounded below.

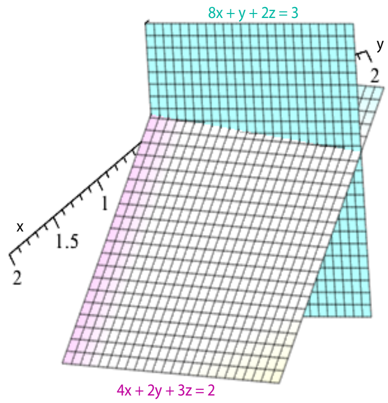


Figure 47.4: The $\mathcal{H}$-polyhedron for the dual linear program of Example 47.1 is the spacial region "above" the pink plane and in "front" of the blue plane. Note $y_1 \to x$, $y_2 \to y$, and $y_3 \to z$.

What happens if x^* is an optimal solution of (P) and if y^* is an optimal solution of (D)? We have $cx^* \leq y^*b$, but is there a "duality gap," that is, is it possible that $cx^* < y^*b$?

The answer is **no**, this is the *strong duality theorem*. Actually, the strong duality theorem asserts more than this.

Theorem 47.7. *(Strong Duality for Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \ \text{and} \ x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. The Primal Problem (P) has a feasible solution and is bounded above iff the Dual Problem (D) has a feasible solution and is bounded below. Furthermore, if (P) has a feasible solution and is bounded above, then for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

Proof. If (P) has a feasible solution and is bounded above, then we know from Proposition 45.1 that (P) has some optimal solution. Let x^* be any optimal solution of (P). First we will show that (D) has a feasible solution v.

Let $\mu = cx^*$ be the maximum of the objective function $x \mapsto cx$. Then for any $\epsilon > 0$, the system of inequalities

$$Ax \leq b, \quad x \geq 0, \quad cx \geq \mu + \epsilon$$

has no solution, since otherwise μ would not be the maximum value of the objective function cx. We would like to apply Farkas II, so first we transform the above system of inequalities into the system

$$\begin{pmatrix} A \\ -c \end{pmatrix} x \leq \begin{pmatrix} b \\ -(\mu+\epsilon) \end{pmatrix}.$$

By Proposition 47.4 (Farkas II), there is some linear form $(\lambda, z) \in (\mathbb{R}^{m+1})^*$ such that $\lambda \geq 0$, $z \geq 0$,

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} A \\ -c \end{pmatrix} \geq 0_m^\top,$$

and

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} b \\ -(\mu+\epsilon) \end{pmatrix} < 0,$$

which means that

$$\lambda A - zc \geq 0_m^\top, \quad \lambda b - z(\mu+\epsilon) < 0,$$

that is,

$$\begin{aligned} \lambda A &\geq zc \\ \lambda b &< z(\mu+\epsilon) \\ \lambda \geq 0&, \;\; z \geq 0. \end{aligned}$$

On the other hand, since $x^* \geq 0$ is an optimal solution of the system $Ax \leq b$, by Farkas II again (by taking the negation of the equivalence), since $\lambda A \geq 0$ (for the same λ as before), we must have

$$\lambda b \geq 0. \tag{$*_1$}$$

We claim that $z > 0$. Otherwise, since $z \geq 0$, we must have $z = 0$, but then

$$\lambda b < z(\mu+\epsilon)$$

implies

$$\lambda b < 0, \tag{$*_2$}$$

and since $\lambda b \geq 0$ by $(*_1)$, we have a contradiction. Consequently, we can divide by $z > 0$ without changing the direction of inequalities, and we obtain

$$\begin{aligned} \frac{\lambda}{z} A &\geq c \\ \frac{\lambda}{z} b &< \mu + \epsilon \\ \frac{\lambda}{z} &\geq 0, \end{aligned}$$

which shows that $v = \lambda/z$ is a feasible solution of the Dual Problem (D). However, weak duality (Proposition 47.6) implies that $cx^* = \mu \leq yb$ for any feasible solution $y \geq 0$ of the Dual Program (D), so (D) is bounded below and by Proposition 45.1 applied to the version of (D) written as a maximization problem, we conclude that (D) has some optimal solution.

For any optimal solution y^* of (D), since v is a feasible solution of (D) such that $vb < \mu + \epsilon$, we must have

$$\mu \leq y^* b < \mu + \epsilon,$$

and since our reasoning is valid for *any* $\epsilon > 0$, we conclude that $cx^* = \mu = y^* b$.

If we assume that the dual program (D) has a feasible solution and is bounded below, since the dual of (D) is (P), we conclude that (P) is also feasible and bounded above. $\square$

The strong duality theorem can also be proven by the simplex method, because when it terminates with an optimal solution of (P), the final tableau also produces an optimal solution y of (D) that can be read off the reduced costs of columns $n+1, \ldots, n+m$ by flipping their signs. We follow the proof in Ciarlet [41] (Chapter 10).

Theorem 47.8. *Consider the Linear Program (P),*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \ \textit{and} \ x \geq 0, \end{aligned}$$

its equivalent version $(P2)$ *in standard form,*

$$\begin{aligned} &\text{maximize} \quad \widehat{c}\,\widehat{x} \\ &\text{subject to} \quad \widehat{A}\widehat{x} = b \ \textit{and} \ \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ *is an* $m \times (n+m)$ *matrix,* $\widehat{c}$ *is a linear form in* $(\mathbb{R}^{n+m})^*$, *and* $\widehat{x} \in \mathbb{R}^{n+m}$, *given by*

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix},$$

and the Dual (D) *of* (P) *given by*

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \ \textit{and} \ y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. *If the simplex algorithm applied to the Linear Program* $(P2)$ *terminates with an optimal solution* $(\widehat{u}^*, K^*)$, *where* $\widehat{u}^*$ *is a basic feasible solution and* K^* *is a basis for* $\widehat{u}^*$, *then* $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$ *is an optimal solution for* (D) *such that* $\widehat{c}\,\widehat{u}^* = y^* b$. *Furthermore,* y^* *is given in terms of the reduced costs by* $y^* = -((\overline{c}_{K^*})_{n+1} \ \ldots (\overline{c}_{K^*})_{n+m})$.

Proof. We know that K^* is a subset of $\{1, \ldots, n+m\}$ consisting of m indices such that the corresponding columns of $\widehat{A}$ are linearly independent. Let $N^* = \{1, \ldots, n+m\} - K^*$. The simplex method terminates with an optimal solution in Case (A), namely when

$$\widehat{c}_j - \sum_{k \in k} \gamma_k^j \widehat{c}_k \leq 0 \quad \text{for all } j \in N^*,$$

where $\widehat{A}^j = \sum_{k\in K^*} \gamma_k^j \widehat{A}^k$, or using the notations of Section 46.3,

$$\widehat{c}_j - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^j \leq 0 \quad \text{for all } j \in N^*.$$

The above inequalities can be written as

$$\widehat{c}_{N^*} - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \leq 0_n^\top,$$

or equivalently as

$$\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_1$}$$

The value of the objective function for the optimal solution $\widehat{u}^*$ is $\widehat{c}\,\widehat{u}^* = \widehat{c}_{K^*}\widehat{u}^*_{K^*}$, and since $\widehat{u}^*_{K^*}$ satisfies the equation $\widehat{A}_{K^*}\widehat{u}^*_{K^*} = b$, the value of the objective function is

$$\widehat{c}_{K^*}\,\widehat{u}^*_{K^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}b. \tag{$*_2$}$$

Then if we let $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$, obviously we have $y^*b = \widehat{c}_{K^*}\widehat{u}_{K^*}$, so if we can prove that y^* is a feasible solution of the Dual Linear program (D), by weak duality, y^* is an optimal solution of (D). We have

$$y^*\widehat{A}_{K^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{K^*} = \widehat{c}_{K^*}, \tag{$*_3$}$$

and by $(*_1)$ we get

$$y^*\widehat{A}_{N^*} = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_4$}$$

Let P be the $(n+m)\times(n+m)$ permutation matrix defined so that

$$\widehat{A}\,P = \begin{pmatrix} A & I_m \end{pmatrix} P = \begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix}.$$

Then we also have

$$\widehat{c}\,P = \begin{pmatrix} c & 0_m^\top \end{pmatrix} P = \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix}.$$

Using Equations $(*_3)$ and $(*_4)$ we obtain

$$y^*\begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix} \geq \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix},$$

that is,

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} P \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix} P,$$

which is equivalent to

$$y^*\begin{pmatrix} A & I_m \end{pmatrix} \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix},$$

that is

$$y^*A \geq c, \quad y \geq 0,$$

and these are exactly the conditions that say that y^* is a feasible solution of the Dual Program (D).

The reduced costs are given by $(\widehat{c}_{K^*})_i = \widehat{c}_i - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^i$, for $i = 1, \ldots, n+m$. But for $i = n+j$ with $j = 1, \ldots, m$ each column $\widehat{A}^{n+j}$ is the jth vector of the identity matrix I_m and by definition $\widehat{c}_{n+j} = 0$, so

$$(\widehat{c}_{K^*})_{n+j} = -(\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1})_j = -y_j^* \quad j = 1, \ldots, m,$$

as claimed. $\square$

The fact that the above proof is fairly short is deceptive because this proof relies on the fact that there are versions of the simplex algorithm using pivot rules that prevent cycling, but the proof that such pivot rules work correctly is quite lengthy. Other proofs are given in Matousek and Gardner [123] (Chapter 6, Sections 6.3), Chvatal [40] (Chapter 5), and Papadimitriou and Steiglitz [134] (Section 2.7).

Observe that since the last m rows of the final tableau are actually obtained by multipling $[u\ \widehat{A}]$ by $\widehat{A}_{K^*}^{-1}$, the $m \times m$ matrix consisting of the last m columns and last m rows of the final tableau is $\widehat{A}_{K^*}^{-1}$ (basically, the simplex algorithm has performed the steps of a Gauss–Jordan reduction). This fact allows saving some steps in the primal dual method.

By combining weak duality and strong duality, we obtain the following theorem which shows that exactly four cases arise.

Theorem 47.9. *(Duality Theorem of Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \ \text{ and } \ x \geq 0, \end{aligned}$$

and let (D) be its dual program

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c \ \text{ and } \ y \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. Then exactly one of the following possibilities occur:

(1) Neither (P) nor (D) has a feasible solution.

(2) (P) is unbounded and (D) has no feasible solution.

(3) (P) has no feasible solution and (D) is unbounded.

(4) Both (P) and (D) have a feasible solution. Then both have an optimal solution, and for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

An interesting corollary of Theorem 47.9 is that there is a test to determine whether a Linear Program (P) has an optimal solution.

Corollary 47.10. *The Primal Program (P) has an optimal solution iff the following set of constraints is satisfiable:*

$$\begin{aligned} Ax &\le b \\ yA &\ge c \\ cx &\ge yb \\ x \ge 0,\ y &\ge 0_m^\top. \end{aligned}$$

In fact, for any feasible solution (x^*, y^*) of the above system, x^* is an optimal solution of (P) and y^* is an optimal solution of (D)

47.3 Complementary Slackness Conditions

Another useful corollary of the strong duality theorem is the following result known as the *equilibrium theorem*.

Theorem 47.11. *(Equilibrium Theorem) For any Linear Program (P) and its Dual Linear Program (D) (with set of inequalities $Ax \le b$ where A is an $m \times n$ matrix, and objective function $x \mapsto cx$), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$y_i = 0 \qquad \text{for all } i \text{ for which } \textstyle\sum_{j=1}^n a_{ij}x_j < b_i \tag{$*_D$}$$

and

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j. \tag{$*_P$}$$

Proof. First assume that $(*_D)$ and $(*_P)$ hold. The equations in $(*_D)$ say that $y_i = 0$ unless $\sum_{j=1}^n a_{ij}x_j = b_i$, hence

$$yb = \sum_{i=1}^m y_i b_i = \sum_{i=1}^m y_i \sum_{j=1}^n a_{ij}x_j = \sum_{i=1}^m \sum_{j=1}^n y_i a_{ij} x_j.$$

Similarly, the equations in $(*_P)$ say that $x_j = 0$ unless $\sum_{i=1}^m y_i a_{ij} = c_j$, hence

$$cx = \sum_{j=1}^n c_j x_j = \sum_{j=1}^n \sum_{i=1}^m y_i a_{ij} x_j.$$

Consequently, we obtain

$$cx = yb.$$

By weak duality (Proposition 47.6), we have

$$cx \le yb = cx$$

for all feasible solutions x of (P), so x is an optimal solution of (P). Similarly,

$$yb = cx \leq yb$$

for all feasible solutions y of (D), so y is an optimal solution of (D).

Let us now assume that x is an optimal solution of (P) and that y is an optimal solution of (D). Then as in the proof of Proposition 47.6,

$$\sum_{j=1}^{n} c_j x_j \leq \sum_{i=1}^{m} \sum_{j=1}^{n} y_i a_{ij} x_j \leq \sum_{i=1}^{m} y_i b_i.$$

By strong duality, since x and y are optimal solutions the above inequalities are actually equalities, so in particular we have

$$\sum_{j=1}^{n} \left(c_j - \sum_{i=1}^{m} y_i a_{ij} \right) x_j = 0.$$

Since x and y are feasible, $x_i \geq 0$ and $y_j \geq 0$, so if $\sum_{i=1}^{m} y_i a_{ij} > c_j$, we must have $x_j = 0$. Similarly, we have

$$\sum_{i=1}^{m} y_i \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right) = 0,$$

so if $\sum_{j=1}^{m} a_{ij} x_j < b_i$, then $y_i = 0$. $\square$

The equations in $(*_D)$ and $(*_P)$ are often called *complementary slackness conditions*. These conditions can be exploited to solve for an optimal solution of the primal problem with the help of the dual problem, and conversely. Indeed, if we guess a solution to one problem, then we may solve for a solution of the dual using the complementary slackness conditions, and then check that our guess was correct. This is the essence of the *primal-dual* method. To present this method, first we need to take a closer look at the dual of a linear program already in standard form.

47.4 Duality for Linear Programs in Standard Form

Let (P) be a linear program in standard form, where $Ax = b$ for some $m \times n$ matrix of rank m and some objective function $x \mapsto cx$ (of course, $x \geq 0$). To obtain the dual of (P) we convert the equations $Ax = b$ to the following system of inequalities involving a $(2m) \times n$ matrix:

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}.$$

Then if we denote the $2m$ dual variables by (y', y''), with $y', y'' \in (\mathbb{R}^m)^*$, the dual of the above program is

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, which is equivalent to

$$\begin{aligned} &\text{minimize} \quad (y' - y'')b \\ &\text{subject to} \quad (y' - y'')A \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$. If we write $y = y' - y''$, we find that the above linear program is equivalent to the following Linear Program (D):

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that y is *not required* to be nonnegative; it is arbitrary.

Next we would like to know what is the version of Theorem 47.8 for a linear program already in standard form. This is very simple.

Theorem 47.12. *Consider the Linear Program (P2) in standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

and its Dual (D) given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. If the simplex algorithm applied to the Linear Program $(P2)$ terminates with an optimal solution (u^*, K^*), where u^* is a basic feasible solution and K^* is a basis for u^*, then $y^* = c_{K^*}A_{K^*}^{-1}$ is an optimal solution for (D) such that $cu^* = y^*b$. Furthermore, if we assume that the simplex algorithm is started with a basic feasible solution (u_0, K_0) where $K_0 = (n-m+1, \ldots, n)$ (the indices of the last m columns of A) and $A_{(n-m+1,\ldots,n)} = I_m$ (the last m columns of A constitute the identity matrix I_m), then the optimal solution $y^* = c_{K^*}A_{K^*}^{-1}$ for (D) is given in terms of the reduced costs by

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

and the $m \times m$ matrix consisting of last m columns and the last m rows of the final tableau is $A_{K^*}^{-1}$.

Proof. The proof of Theorem 47.8 applies with A instead of $\widehat{A}$, and we can show that

$$c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*},$$

and that $y^* = c_{K^*}A_{K^*}^{-1}$ satisfies, $cu^* = y^*b$, and

$$\begin{aligned} y^*A_{K^*} &= c_{K^*}A_{K^*}^{-1}A_{K^*} = c_{K^*}, \\ y^*A_{N^*} &= c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*}. \end{aligned}$$

Let P be the $n \times n$ permutation matrix defined so that

$$AP = \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix}.$$

Then we also have

$$cP = \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

and using the above equations and inequalities we obtain

$$y^* \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix} \geq \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

that is, $y^*AP \geq cP$, which is equivalent to

$$y^*A \geq c,$$

which shows that y^* is a feasible solution of (D) (remember, y^* is arbitrary so there is no need for the constraint $y^* \geq 0$).

The reduced costs are given by

$$(\overline{c}_{K^*})_i = c_i - c_{K^*}A_{K^*}^{-1}A^i,$$

and since for $j = n-m+1, \ldots, n$ the column A^j is the $(j+m-n)$th column of the identity matrix I_m, we have

$$(\overline{c}_{K^*})_j = c_j - (c_{K^*}A_{K^*}^{-1})_{j+m-n} \quad j = n-m+1, \ldots, n,$$

that is,

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

as claimed. Since the last m rows of the final tableau is obtained by multiplying $[u_0 \;\; A]$ by $A_{K^*}^{-1}$, and the last m columns of A constitute I_m, the last m rows and the last m columns of the final tableau constitute $A_{K^*}^{-1}$. $\square$

Let us now take a look at the complementary slackness conditions of Theorem 47.11. If we go back to the version of (P) given by

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix} \;\text{ and }\; x \geq 0, \end{aligned}$$

and to the version of (D) given by

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, since the inequalities $Ax \leq b$ and $-Ax \leq -b$ together imply that $Ax = b$, we have equality for all these inequality constraints, and so the Conditions $(*_D)$ place no constraints at all on y' and y'', while the Conditions $(*_P)$ assert that

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m (y_i' - y_i'')a_{ij} > c_j.$$

If we write $y = y' - y''$, the above conditions are equivalent to

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j.$$

Thus we have the following version of Theorem 47.11.

Theorem 47.13. *(Equilibrium Theorem, Version 2) For any Linear Program $(P2)$ in standard form (with $Ax = b$ where A is an $m \times n$ matrix, $x \geq 0$, and objective function $x \mapsto cx$) and its Dual Linear Program (D), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$x_j = 0 \qquad \textit{for all } j \textit{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j. \tag{$*_P$}$$

Therefore, the slackness conditions applied to a Linear Program $(P2)$ in standard form and to its Dual (D) *only impose slackness conditions on the variables x_j of the primal problem*.

The above fact plays a crucial role in the primal-dual method.

47.5 The Dual Simplex Algorithm

Given a Linear Program $(P2)$ in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if no obvious feasible solution is available *but if* $c \leq 0$, rather than using the method for finding a feasible solution described in Section 46.2 we may use a method known as the dual simplex algorithm. This method uses basic solutions (u, K) where $Au = b$ and $u_j = 0$ for all $u_j \notin K$, but does not require $u \geq 0$, so u may not be feasible. However, $y = c_K A_K^{-1}$ is required to be feasible for the dual program

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^*)^m$. Since $c \leq 0$, observe that $y = 0_m^\top$ is a feasible solution of the dual.

If a basic solution u of $(P2)$ is found such that $u \geq 0$, then $cu = yb$ for $y = c_K A_K^{-1}$, and we have found an optimal solution u for $(P2)$ and y for (D). The dual simplex method makes progress by attempting to make negative components of u zero and by decreasing the objective function of the dual program.

The dual simplex method starts with a basic solution (u, K) of $Ax = b$ which is not feasible but for which $y = c_K A_K^{-1}$ is dual feasible. In many cases the original linear program is specified by a set of inequalities $Ax \leq b$ with some $b_i < 0$, so by adding slack variables it is easy to find such basic solution u, and if in addition $c \leq 0$, then because the cost associated with slack variables is 0, we see that $y = 0$ is a feasible solution of the dual.

Given a basic solution (u, K) of $Ax = b$ (feasible or not), $y = c_K A_K^{-1}$ is dual feasible iff $c_K A_K^{-1} A \geq c$, and since $c_K A_K^{-1} A_K = c_K$, the inequality $c_K A_K^{-1} A \geq c$ is equivalent to $c_K A_K^{-1} A_N \geq c_N$, that is,

$$c_N - c_K A_K^{-1} A_N \leq 0, \tag{$*_1$}$$

where $N = \{1, \ldots, n\} - K$. Equation $(*_1)$ is equivalent to

$$c_j - c_K \gamma_K^j \leq 0 \quad \text{for all } j \in N, \tag{$*_2$}$$

where $\gamma_K^j = A_K^{-1} A^j$. Recall that the notation $\overline{c}_j$ is used to denote $c_j - c_K \gamma_K^j$, which is called the *reduced cost* of the variable x_j.

As in the simplex algorithm we need to decide which column A^k leaves the basis K and which column A^j enters the new basis K^+, in such a way that $y^+ = c_{K^+} A_{K^+}^{-1}$ is a feasible solution of (D), that is, $c_{N^+} - c_{K^+} A_{K^+}^{-1} A_{N^+} \leq 0$, where $N^+ = \{1, \ldots, n\} - K^+$. We use Proposition 46.2 to decide wich column k^- should leave the basis.

Suppose (u, K) is a solution of $Ax = b$ for which $y = c_K A_K^{-1}$ is dual feasible.

Case (A). If $u \geq 0$, then u is an optimal solution of $(P2)$.

Case (B). There is some $k \in K$ such that $u_k < 0$. In this case pick some $k^- \in K$ such that $u_{k^-} < 0$ (according to some pivot rule).

Case (B1). Suppose that $\gamma_{k^-}^j \geq 0$ for all $j \notin K$ (in fact, for all j, since $\gamma_{k^-}^j \in \{0, 1\}$ for all $j \in K$). If so, we we claim that $(P2)$ is not feasible.

Indeed, let v be some basic feasible solution. We have $v \geq 0$ and $Av = b$, that is,

$$\sum_{j=1}^{n} v_j A^j = b,$$

so by multiplying both sides by A_K^{-1} and using the fact that by definition $\gamma_K^j = A_K^{-1} A^j$, we obtain

$$\sum_{j=1}^{n} v_j \gamma_K^j = A_K^{-1} b = u_K.$$

But recall that by hypothesis $u_{k^-} < 0$, yet $v_j \geq 0$ and $\gamma^j_{k^-} \geq 0$ for all j, so the component of index k^- is zero or positive on the left, and negative on the right, a contradiction. Therefore, $(P2)$ is indeed not feasible.

Case (B2). We have $\gamma^j_{k^-} < 0$ for some j.

We pick the column A^j entering the basis among those for which $\gamma^j_{k^-} < 0$. Since we assumed that $c_j - c_K\gamma^j_K \leq 0$ for all $j \in N$ by $(*_2)$, consider

$$\mu^+ = \max\left\{-\frac{c_j - c_K\gamma^j_K}{\gamma^j_{k^-}} \;\middle|\; \gamma^j_{k^-} < 0,\ j \in N\right\} = \max\left\{-\frac{\overline{c}_j}{\gamma^j_{k^-}} \;\middle|\; \gamma^j_{k^-} < 0,\ j \in N\right\} \leq 0,$$

and the set

$$N(\mu^+) = \left\{j \in N \;\middle|\; -\frac{\overline{c}_j}{\gamma^j_{k^-}} = \mu^+\right\}.$$

We pick some index $j^+ \in N(\mu^+)$ as the index of the column entering the basis (using some pivot rule).

Recall that by hypothesis $c_i - c_K\gamma^i_K \leq 0$ for all $j \notin K$ and $c_i - c_K\gamma^i_K = 0$ for all $i \in K$. Since $\gamma^{j^+}_{k^-} < 0$, for any index i such that $\gamma^i_{k^-} \geq 0$, we have $-\gamma^i_{k^-}/\gamma^{j^+}_{k^-} \geq 0$, and since by Proposition 46.2

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K),$$

we have $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. For any index i such that $\gamma^i_{k^-} < 0$, by the choice of $j^+ \in K^*$,

$$-\frac{c_i - c_K\gamma^i_K}{\gamma^i_{k^-}} \leq -\frac{c_{j^+} - c_K\gamma^{j^+}_K}{\gamma^{j^+}_{k^-}},$$

so

$$c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K) \leq 0,$$

and again, $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. Therefore, if we let $K^+ = (K - \{k^-\}) \cup \{j^+\}$, then $y^+ = c_{K^+}A^{-1}_{K^+}$ is dual feasible. As in the simplex algorithm, θ^+ is given by

$$\theta^+ = u_{k^-}/\gamma^{j^+}_{k^-} \geq 0,$$

and u^+ is also computed as in the simplex algorithm by

$$u^+_i = \begin{cases} u_i - \theta^{j^+}\gamma^{j^+}_i & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}.$$

The change in the objective function of the primal and dual program (which is the same, since $u_K = A_K^{-1}b$ and $y = c_K A_K^{-1}$ is chosen such that $cu = c_K u_K = yb$) is the same as in the simplex algorithm, namely

$$\theta^+ \left(c^{j^+} - c_K \gamma_K^{j^+}\right).$$

We have $\theta^+ > 0$ and $c^{j^+} - c_K\gamma_K^{j^+} \leq 0$, so if $c^{j^+} - c_K\gamma_K^{j^+} < 0$, then the objective function of the dual program decreases strictly.

Case (B3). $\mu^+ = 0$.

The possibity that $\mu^+ = 0$, that is, $c^{j^+} - c_K\gamma_K^{j^+} = 0$, may arise. In this case, the objective function doesn't change. This is a case of degeneracy similar to the degeneracy that arises in the simplex algorithm. We still pick $j^+ \in N(\mu^+)$, but we need a pivot rule that prevents cycling. Such rules exist; see Bertsimas and Tsitsiklis [21] (Section 4.5) and Papadimitriou and Steiglitz [134] (Section 3.6).

The reader surely noticed that the dual simplex algorithm is very similar to the simplex algorithm, except that the simplex algorithm preserves the property that (u, K) is (primal) feasible, whereas the dual simplex algorithm preserves the property that $y = c_K A_K^{-1}$ is dual feasible. One might then wonder whether the dual simplex algorithm is equivalent to the simplex algorithm applied to the dual problem. This is indeed the case, there is a one-to-one correspondence between the dual simplex algorithm and the simplex algorithm applied to the dual problem in maximization form. This correspondence is described in Papadimitriou and Steiglitz [134] (Section 3.7).

The comparison between the simplex algorithm and the dual simplex algorithm is best illustrated if we use a description of these methods in terms of *(full) tableaux*.

Recall that a *(full) tableau* is an $(m + 1) \times (n + 1)$ matrix organized as follows:

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The top row contains the current value of the objective function and the reduced costs, the first column except for its top entry contain the components of the current basic solution u_K, and the remaining columns except for their top entry contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $\overline{c}_i - (\gamma_{k^-}^i / \gamma_{k^-}^{j^+})\overline{c}_{j^+}$.

When executing the simplex algorithm, we have $u_k \geq 0$ for all $k \in K$ (and $u_j = 0$ for all $j \notin K$), and the incoming column j^+ is determined by picking one of the column indices such that $\overline{c}_j > 0$. Then the index k^- of the leaving column is determined by looking at the minimum of the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+).

On the other hand, when executing the dual simplex algorithm, we have $\overline{c}_j \leq 0$ for all $j \notin K$ (and $\overline{c}_k = 0$ for all $k \in K$), and the outgoing column k^- is determined by picking one of the row indices such that $u_k < 0$. The index j^+ of the incoming column is determined by looking at the maximum of the ratios $-\overline{c}_j/\gamma^j_{k^-}$ for which $\gamma^j_{k^-} < 0$ (along row k^-).

More details about the comparison between the simplex algorithm and the dual simplex algorithm can be found in Bertsimas and Tsitsiklis [21] and Papadimitriou and Steiglitz [134].

Here is an example of the the dual simplex method.

Example 47.2. Consider the following linear program in standard form:

$$\text{Maximize} \quad -4x_1 - 2x_2 - x_3$$

$$\text{subject to} \quad \begin{pmatrix} -1 & -1 & 2 & 1 & 0 & 0 \\ -4 & -2 & 1 & 0 & 1 & 0 \\ 1 & 1 & -4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 2 \end{pmatrix} \quad \text{and } x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

We initialize the dual simplex procedure with (u, K) where $u = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -3 \\ -4 \\ 2 \end{pmatrix}$ and $K = (4, 5, 6)$.

The initial tableau, before explicitly calculating the reduced cost, is

0	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$	$\overline{c}_5$	$\overline{c}_6$
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since u has negative coordinates, Case (B) applies, and we will set $k^- = 4$. We must now determine whether Case (B1) or Case (B2) applies. This determination is accomplished by scanning the first three columns in the tableau and observing each column has a negative entry. Thus Case (B2) is applicable, and we need to determine the reduced costs. Observe that $c = (-4, -2, -1, 0, 0, 0)$, which in turn implies $c_{(4,5,6)} = (0, 0, 0)$. Equation $(*_2)$ implies that the nonzero reduced costs are

$$\overline{c}_1 = c_1 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -4 \\ 1 \end{pmatrix} = -4$$
$$\overline{c}_2 = c_2 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = -2$$
$$\overline{c}_3 = c_3 - c_{(4,5,6)} \begin{pmatrix} -2 \\ 1 \\ 4 \end{pmatrix} = -1,$$

and our tableau becomes

0	-4	-2	-1	0	0	0
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since $k^- = 4$, our pivot row is the first row of the tableau. To determine candidates for j^+, we scan this row, locate negative entries and compute

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_4^j} \;\middle|\; \gamma_4^j < 0,\ j \in \{1, 2, 3\} \right\} = \max\left\{ \frac{-2}{1}, \frac{-4}{1} \right\} = -2.$$

Since μ^+ occurs when $j = 2$, we set $j^+ = 2$. Our new basis is $K^+ = (2, 5, 6)$. We must normalize the first row of the tableau, namely multiply by -1, then add twice this normalized row to the second row, and subtract the normalized row from the third row to obtain the updated tableau.

0	-4	-2	-1	0	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

It remains to update the reduced costs and the value of the objective function by adding twice the normalized row to the top row.

6	-2	0	-5	-2	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

We now repeat the procedure of Case (B2) and set $k^- = 6$ (since this is the only negative entry of u^+). Our pivot row is now the third row of the updated tableau, and the new μ^+

becomes

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_6^j} \;\middle|\; \gamma_6^j < 0,\ j \in \{1,3,4\} \right\} = \max\left\{ \frac{-5}{2} \right\} = -\frac{5}{2},$$

which implies that $j^+ = 3$. Hence the new basis is $K^+ = (2, 5, 3)$, and we update the tableau by taking $-\frac{1}{2}$ of Row 3, adding twice the normalized Row 3 to Row 1, and adding three times the normalized Row 3 to Row 2.

6	-2	0	-5	-2	0	0
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-7/2$	1	$-3/2$
$u_3 = 1/2$	0	0	1	$-1/2$	0	$-1/2$

It remains to update the objective function and the reduced costs by adding five times the normalized row to the top row.

$17/2$	-2	0	0	$-9/2$	0	$-5/2$
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-\frac{7}{2}$	1	$-3/2$
$u_3 = 1/2$	0	0	1	$-1/2$	0	$-1/2$

Since u^+ has no negative entries, the dual simplex method terminates and objective function $-4x_1 - 2x_2 - x_3$ is maximized with $-\frac{17}{2}$ at $(0, 4, \frac{1}{2})$. See Figure 47.5.

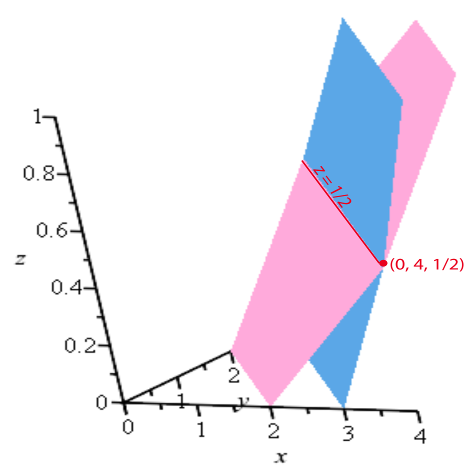


Figure 47.5: The objective function $-4x_1 - 2x_2 - x_3$ is maximized at the intersection between the blue plane $-x_1 - x_2 + 2x_3 = -3$ and the pink plane $x_1 + x_2 - 4x_3 = 2$.

47.6 The Primal-Dual Algorithm

Let $(P2)$ be a linear program in standard form

$$\begin{aligned}
&\text{maximize} && cx\\
&\text{subject to} && Ax = b \text{ and } x \geq 0,
\end{aligned}$$

where A is an $m \times n$ matrix of rank m, and (D) be its dual given by

$$\begin{aligned}
&\text{minimize} && yb\\
&\text{subject to} && yA \geq c,
\end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

First we may assume that $b \geq 0$ by changing every equation $\sum_{j=1}^n a_{ij}x_j = b_i$ with $b_i < 0$ to $\sum_{j=1}^n -a_{ij}x_j = -b_i$. If we happen to have some feasible solution y of the dual program (D), we know from Theorem 47.13 that a feasible solution x of $(P2)$ is an optimal solution iff the equations in $(*_P)$ hold. If we denote by J the subset of $\{1, \ldots, n\}$ for which the equalities

$$yA^j = c_j$$

hold, then by Theorem 47.13 a feasible solution x of $(P2)$ is an optimal solution iff

$$x_j = 0 \quad \text{for all } j \notin J.$$

Let $|J| = p$ and $N = \{1, \ldots, n\} - J$. The above suggests looking for $x \in \mathbb{R}^n$ such that

$$\begin{aligned}
\sum_{j \in J} x_j A^j &= b\\
x_j &\geq 0 \quad \text{for all } j \in J\\
x_j &= 0 \quad \text{for all } j \notin J,
\end{aligned}$$

or equivalently

$$A_J x_J = b, \quad x_J \geq 0, \tag{$*_1$}$$

and

$$x_N = 0_{n-p}.$$

To search for such an x, we just need to look for a feasible x_J, and for this we can use the *Restricted Primal* linear program (RP) defined as follows:

$$\begin{aligned}
&\text{maximize} && -(\xi_1 + \cdots + \xi_m)\\
&\text{subject to} && \begin{pmatrix} A_J & I_m \end{pmatrix} \begin{pmatrix} x_J \\ \xi \end{pmatrix} = b \text{ and } x, \xi \geq 0.
\end{aligned}$$

Since by hypothesis $b \geq 0$ and the objective function is bounded above by 0, this linear program has an optimal solution (x^*_J, ξ^*).

If $\xi^* = 0$, then the vector $u^* \in \mathbb{R}^n$ given by $u^*_J = x^*_J$ and $u^*_N = 0_{n-p}$ is an optimal solution of (P).

Otherwise, $\xi^* > 0$ and we have failed to solve $(*_1)$. However we may try to use ξ^* to improve y. For this consider the Dual (DRP) of (RP):

$$\begin{aligned} &\text{minimize} && zb \\ &\text{subject to} && zA_J \geq 0 \\ & && z \geq -\mathbf{1}^\top_m. \end{aligned}$$

Observe that the Program (DRP) has the same objective function as the original Dual Program (D). We know by Theorem 47.12 that the optimal solution (x^*_J, ξ^*) of (RP) yields an optimal solution z^* of (DRP) such that

$$z^* b = -(\xi^*_1 + \cdots + \xi^*_m) < 0.$$

In fact, if K^* is the basis associated with (x^*_J, ξ^*) and if we write

$$\widehat{A} = \begin{pmatrix} A_J & I_m \end{pmatrix}$$

and $\widehat{c} = [0^\top_p \quad -\mathbf{1}^\top]$, then by Theorem 47.12 we have

$$z^* = \widehat{c}_{K^*} \widehat{A}^{-1}_{K^*} = -\mathbf{1}^\top_m - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

where $(\overline{c}_{K^*})_{(p+1,\ldots,p+m)}$ denotes the row vector of reduced costs in the final tableau corresponding to the last m columns.

If we write

$$y(\theta) = y + \theta z^*,$$

then the new value of the objective function of (D) is

$$y(\theta)b = yb + \theta z^* b, \tag{$*_2$}$$

and since $z^* b < 0$, we have a chance of improving the objective function of (D), that is, decreasing its value for $\theta > 0$ small enough if $y(\theta)$ is feasible for (D). This will be the case iff $y(\theta)A \geq c$ iff

$$yA + \theta z^* A \geq c. \tag{$*_3$}$$

Now since y is a feasible solution of (D) we have $yA \geq c$, so if $z^* A \geq 0$, then $(*_3)$ is satisfied and $y(\theta)$ is a solution of (D) for all $\theta > 0$, which means that (D) is unbounded. But this implies that (P) is not feasible.

Let us take a closer look at the inequalities $z^*A \geq 0$. For $j \in J$, since z^* is an optimal solution of (DRP), we know that $z^*A_J \geq 0$, so if $z^*A^j \geq 0$ for all $j \in N$, then $(P2)$ is not feasible.

Otherwise, there is some $j \in N = \{1, \ldots, n\} - J$ such that

$$z^*A^j < 0,$$

and then since by the definition of N we have $yA^j > c_j$ for all $j \in N$, if we pick θ such that

$$0 < \theta \leq \frac{yA^j - c_j}{-z^*A^j} \quad j \in N,\ z^*A^j < 0,$$

then we decrease the objective function $y(\theta)b = yb + \theta z^*b$ of (D) (since $z^*b < 0$). Therefore we pick the best θ, namely

$$\theta^+ = \min\left\{ \frac{yA^j - c_j}{-z^*A^j} \;\middle|\; j \notin J,\ z^*A^j < 0 \right\} > 0. \tag{$*_4$}$$

Next we update y to $y^+ = y(\theta^+) = y + \theta^+ z^*$, we create the new restricted primal with the new subset

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+A^j = c_j\},$$

and repeat the process.

Here are the steps of the primal-dual algorithm.

Step 1. Find some feasible solution y of the Dual Program (D). We will show later that this is always possible.

Step 2. Compute

$$J^+ = \{j \in \{1, \ldots, n\} \mid yA^j = c_j\}.$$

Step 3. Set $J = J^+$ and solve the Problem (RP) using the simplex algorithm, starting from the optimal solution determined during the previous round, obtaining the optimal solution (x_J^*, ξ^*) with the basis K^*.

Step 4.

If $\xi^* = 0$, then stop with an optimal solution u^* for (P) such that $u_J^* = x_J^*$ and the other components of u^* are zero.

Else let

$$z^* = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

be the optimal solution of (DRP) corresponding to (x_J^*, ξ^*) and the basis K^*.

If $z^*A^j \geq 0$ for all $j \notin J$, then stop; the Program (P) has no feasible solution.

Else compute

$$\theta^+ = \min\left\{-\frac{yA^j - c_j}{z^*A^j} \;\middle|\; j \notin J,\, z^*A^j < 0\right\}, \qquad y^+ = y + \theta^+ z^*,$$

and

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+ A^j = c_j\}.$$

Go back to Step 3.

The following proposition shows that at each iteration we can start the Program (RP) with the optimal solution obtained at the previous iteration.

Proposition 47.14. *Every $j \in J$ such that A^j is in the basis of the optimal solution ξ^* belongs to the next index set J^+.*

Proof. Such an index $j \in J$ correspond to a variable ξ_j such that $\xi_j > 0$, so by complementary slackness, the constraint $z^*A^j \geq 0$ of the Dual Program (DRP) must be an equality, that is, $z^*A^j = 0$. But then we have

$$y^+A^j = yA^j + \theta^+ z^* A^j = c_j,$$

which shows that $j \in J^+$. $\square$

If (u^*, ξ^*) with the basis K^* is the optimal solution of the Program (RP), Proposition 47.14 together with the last property of Theorem 47.12 allows us to restart the (RP) in Step 3 with $(u^*, \xi^*)_{K^*}$ as initial solution (with basis K^*). For every $j \in J - J^+$, column j is deleted, and for every $j \in J^+ - J$, the new column A^j is computed by multiplying $\widehat{A}_{K^*}^{-1}$ and A^j, but $\widehat{A}_{K^*}^{-1}$ is the matrix $\Gamma^*[1{:}m; p+1{:}p+m]$ consisting of the last m columns of Γ^* in the final tableau, and the new reduced $\overline{c}_j$ is given by $c_j - z^*A^j$. Reusing the optimal solution of the previous (RP) may improve efficiency significantly.

Another crucial observation is that for any index $j_0 \in N$ such that $\theta^+ = (yA^{j_0} - c_{j_0})/(-z^*A^{j_0})$, we have

$$y^+A_{j_0} = yA_{j_0} + \theta^+ z^* A^{j_0} = c_{j_0},$$

and so $j_0 \in J^+$. This fact that be used to ensure that the primal-dual algorithm terminates in a finite number of steps (using a pivot rule that prevents cycling); see Papadimitriou and Steiglitz [134] (Theorem 5.4).

It remains to discuss how to pick some initial feasible solution y of the Dual Program (D). If $c_j \leq 0$ for $j = 1, \ldots, n$, then we can pick $y = 0$. If we are dealing with a minimization problem, the weight c_j are often nonnegative, so from the point of view of maximization we will have $-c_j \leq 0$ for all j, and we will be able to use $y = 0$ as a starting point.

Going back to our primal problem in maximization form and its dual in minimization form, we still need to deal with the situation where $c_j > 0$ for some j, in which case there may not be any obvious y feasible for (D). Preferably we would like to find such a y very cheaply.

There is a trick to deal with this situation. We pick some very large positive number M and add to the set of equations $Ax = b$ the new equation

$$x_1 + \cdots + x_n + x_{n+1} = M,$$

with the new variable x_{n+1} constrained to be nonnegative. If the Program (P) has a feasible solution, such an M exists. In fact it can shown that for any basic feasible solution $u = (u_1, \ldots, u_n)$, each $|u_i|$ is bounded by some expression depending only on A and b; see Papadimitriou and Steiglitz [134] (Lemma 2.1). The proof is not difficult and relies on the fact that the inverse of a matrix can be expressed in terms of certain determinants (the adjugates). Unfortunately, this bound contains $m!$ as a factor, which makes it quite impractical.

Having added the new equation above, we obtain the new set of equations

$$\begin{pmatrix} A & 0_n \\ \mathbf{1}_n^\top & 1 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} b \\ M \end{pmatrix},$$

with $x \geq 0, x_{n+1} \geq 0$, and the new objective function given by

$$\begin{pmatrix} c & 0 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = cx.$$

The dual of the above linear program is

$$\begin{aligned} &\text{minimize} \quad yb + y_{m+1}M \\ &\text{subject to} \quad yA^j + y_{m+1} \geq c_j \quad j = 1, \ldots, n \\ &\qquad\qquad\quad y_{m+1} \geq 0. \end{aligned}$$

If $c_j > 0$ for some j, observe that the linear form $\widetilde{y}$ given by

$$\widetilde{y}_i = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ \max_{1 \leq j \leq n}\{c_j\} > 0 \end{cases}$$

is a feasible solution of the new dual program. In practice, we can choose M to be a number close to the largest integer representable on the computer being used.

Here is an example of the primal-dual algorithm given in the Math 588 class notes of T. Molla [128].

Example 47.3. Consider the following linear program in standard form:

$$\begin{aligned} &\text{Maximize} \quad -x_1 - 3x_2 - 3x_3 - x_4 \\ &\text{subject to} \quad \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, x_2, x_3, x_4 \geq 0. \end{aligned}$$

The associated Dual Program (D) is

$$\begin{aligned} &\text{Minimize} && 2y_1 + y_2 + 4y_3 \\ &\text{subject to} && \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \geq \begin{pmatrix} -1 & -3 & -3 & -1 \end{pmatrix}. \end{aligned}$$

We initialize the primal-dual algorithm with the dual feasible point $y = (-1/3 \;\; 0 \;\; 0)$. Observe that only the first inequality of (D) is actually an equality, and hence $J = \{1\}$. We form the Restricted Primal Program $(RP1)$

$$\begin{aligned} &\text{Maximize} && -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} && \begin{pmatrix} 3 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 6 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and } x_1, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

We now solve $(RP1)$ via the simplex algorithm. The initial tableau with $K = (2, 3, 4)$ and $J = \{1\}$ is

	x_1	ξ_1	ξ_2	ξ_3
7	12	0	0	0
$\xi_1 = 2$	3	1	0	0
$\xi_2 = 1$	3	0	1	0
$\xi_3 = 4$	6	0	0	1

.

For $(RP1)$, $\hat{c} = (0, -1, -1, -1)$, $(x_1, \xi_1, \xi_2, \xi_3) = (0, 2, 1, 4)$, and the nonzero reduced cost is given by

$$0 - (-1 \;\; -1 \;\; -1) \begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix} = 12.$$

Since there is only one nonzero reduced cost, we must set $j^+ = 1$. Since $\min\{\xi_1/3, \xi_2/3, \xi_3/6\} = 1/3$, we see that $k^- = 3$ and $K = (2, 1, 4)$. Hence we pivot through the red circled 3 (namely we divide row 2 by 3, and then subtract $3\times$ (row 2) from row 1, $6\times$ (row 2) from row 3, and $12\times$ (row 2) from row 0), to obtain the tableau

	x_1	ξ_1	ξ_2	ξ_3
3	0	0	-4	0
$\xi_1 = 1$	0	1	-1	0
$x_1 = 1/3$	1	0	$1/3$	0
$\xi_3 = 2$	0	0	-2	1

.

At this stage the simplex algorithm for $(RP1)$ terminates since there are no positive reduced costs. Since the upper left corner of the final tableau is not zero, we proceed with Step 4 of the primal dual algorithm and compute

$$z^* = (-1 \;\; -1 \;\; -1) - (0 \;\; -4 \;\; 0) = (-1 \;\; 3 \;\; -1),$$

$$yA^2 - c_2 = (-1/3\ \ 0\ \ 0)\begin{pmatrix}4\\-2\\4\end{pmatrix} + 3 = \frac{5}{3}, \qquad z^*A^2 = -(-1\ \ 3\ \ -1)\begin{pmatrix}4\\-2\\4\end{pmatrix} = 14,$$

$$yA^4 - c_4 = (-1/3\ \ 0\ \ 0)\begin{pmatrix}1\\-1\\1\end{pmatrix} + 1 = \frac{2}{3}, \qquad z^*A^4 = -(-1\ \ 3\ \ -1)\begin{pmatrix}1\\-1\\1\end{pmatrix} = 5,$$

so

$$\theta^+ = \min\left\{\frac{5}{42}, \frac{2}{15}\right\} = \frac{5}{42},$$

and we conclude that the new feasible solution for (D) is

$$y^+ = (-1/3\ \ 0\ \ 0) + \frac{5}{42}(-1\ \ 3\ \ -1) = (-19/42\ \ 5/14\ \ -5/42).$$

When we substitute y^+ into (D), we discover that the first two constraints are equalities, and that the new J is $J = \{1, 2\}$. The new Reduced Primal $(RP2)$ is

$$\begin{aligned} &\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} \quad \begin{pmatrix}3 & 4 & 1 & 0 & 0\\ 3 & -2 & 0 & 1 & 0\\ 6 & 4 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\\xi_1\\\xi_2\\\xi_3\end{pmatrix} = \begin{pmatrix}2\\1\\4\end{pmatrix} \quad \text{and } x_1, x_2, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

Once again, we solve $(RP2)$ via the simplex algorithm, where $\hat{c} = (0, 0, -1, -1, -1)$, $(x_1, x_2, \xi_1, \xi_2, \xi_3) = (1/3, 0, 1, 0, 2)$ and $K = (3, 1, 5)$. The initial tableau is obtained from the final tableau of the previous $(RP1)$ by adding a column corresponding the the variable x_2, namely

$$\widehat{A}_K^{-1}A^2 = \begin{pmatrix}1 & -1 & 0\\ 0 & 1/3 & 0\\ 0 & -2 & 1\end{pmatrix}\begin{pmatrix}4\\-2\\4\end{pmatrix} = \begin{pmatrix}6\\-2/3\\8\end{pmatrix},$$

with

$$\overline{c}_2 = c_2 - z^*A^2 = 0 - \begin{pmatrix}-1 & 3 & -1\end{pmatrix}\begin{pmatrix}4\\-2\\4\end{pmatrix} = 14,$$

and we get

	x_1	x_2	ξ_1	ξ_2	ξ_3
3	0	14	0	-4	0
$\xi_1 = 1$	0	6	1	-1	0
$x_1 = 1/3$	1	$-2/3$	0	$1/3$	0
$\xi_3 = 2$	0	8	0	-2	1

.

Note that $j^+ = 2$ since the only positive reduced cost occurs in column 2. Also observe that since $\min\{\xi_1/6, \xi_3/8\} = \xi_1/6 = 1/6$, we set $k^- = 3$, $K = (2, 1, 5)$ and pivot along the red 6 to obtain the tableau

	x_1	x_2	ξ_1	ξ_2	ξ_3
$2/3$	0	0	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$-4/3$	$-2/3$	1

.

Since the reduced costs are either zero or negative the simplex algorithm terminates, and we compute

$$
\begin{aligned}
z^* &= (-1\ \ -1\ \ -1) - (-7/3\ \ -5/3\ \ 0) = (4/3\ \ 2/3\ \ -1),\\
y^+A^4 - c_4 &= (-19/42\ \ 5/14\ \ -5/42)\begin{pmatrix}1\\-1\\1\end{pmatrix} + 1 = 1/14,\\
z^*A^4 &= -(4/3\ \ 2/3\ \ -1)\begin{pmatrix}1\\-1\\1\end{pmatrix} = 1/3,
\end{aligned}
$$

so

$$
\begin{aligned}
\theta^+ &= \frac{3}{14},\\
y^+ &= (-19/42\ \ 5/14\ \ -5/42) + \frac{5}{14}(4/3\ \ 2/3\ \ -1) = (-1/6\ \ 1/2\ \ -1/3).
\end{aligned}
$$

When we plug y^+ into (D), we discover that the first, second, and fourth constraints are equalities, which implies $J = \{1, 2, 4\}$. Hence the new Restricted Primal $(RP3)$ is

$$
\begin{aligned}
&\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)\\
&\text{subject to} \quad \begin{pmatrix}3 & 4 & 1 & 1 & 0 & 0\\ 3 & -2 & -1 & 0 & 1 & 0\\ 6 & 4 & 1 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_4\\\xi_1\\\xi_2\\\xi_3\end{pmatrix} = \begin{pmatrix}2\\1\\4\end{pmatrix} \quad \text{and } x_1, x_2, x_4, \xi_1, \xi_2, \xi_3 \geq 0.
\end{aligned}
$$

The initial tableau for $(RP3)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_2, x_4, \xi_1, \xi_2, \xi_3) = (4/9, 1/6, 0, 0, 0, 2/3)$ and $K = (2, 1, 6)$, is obtained from the final tableau of the previous $(RP2)$ by adding a column corresponding the the variable x_4, namely

$$
\widehat{A}_K^{-1}A^4 = \begin{pmatrix}1/6 & -1/6 & 0\\ 1/9 & 2/9 & 0\\ -4/3 & -2/3 & 1\end{pmatrix}\begin{pmatrix}1\\-1\\1\end{pmatrix} = \begin{pmatrix}1/3\\-1/9\\1/3\end{pmatrix},
$$

with

$$\overline{c}_4 = c_4 - z^* A^4 = 0 - \begin{pmatrix} 4/3 & 2/3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

and we get

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$2/3$	0	0	$1/3$	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/3$	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$-1/9$	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$1/3$	$-4/3$	$-2/3$	1

.

Since the only positive reduced cost occurs in column 3, we set $j^+ = 3$. Furthermore since $\min\{x_2/(1/3), \xi_3/(1/3)\} = x_2/(1/3) = 1/2$, we let $k^- = 2$, $K = (3, 1, 6)$, and pivot around the red circled $1/3$ to obtain

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	-1	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	3	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/3$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	-1	0	$-3/2$	$-1/2$	1

.

At this stage there are no positive reduced costs, and we must compute

$$z^* = (-1 \;\; -1 \;\; -1) - (-5/2 \;\; -3/2 \;\; 0) = (3/2 \;\; 1/2 \;\; -1),$$

$$y^+ A^3 - c_3 = (-1/6 \;\; 1/2 \;\; -1/3) \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} + 3 = 13/2,$$

$$z^* A^3 = -(3/2 \;\; 1/2 \;\; -1) \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

so

$$\theta^+ = \frac{13}{3},$$
$$y^+ = (-1/6 \;\; 1/2 \;\; -1/3) + \frac{13}{3}(3/2 \;\; 1/2 \;\; -1) = (19/3 \;\; 8/3 \;\; -14/3).$$

We plug y^+ into (D) and discover that the first, third, and fourth constraints are equalities. Thus, $J = \{1, 3, 4\}$ and the Restricted Primal $(RP4)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$

$$\text{subject to} \quad \begin{pmatrix} 3 & -3 & 1 & 1 & 0 & 0 \\ 3 & 6 & -1 & 0 & 1 & 0 \\ 6 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and } x_1, x_3, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP4)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_3, x_4, \xi_1, \xi_2, \xi_3) = (1/2, 0, 1/2, 0, 0, 1/2)$ and $K = (3, 1, 6)$ is obtained from the final tableau of the previous $(RP3)$ by replacing the column corresponding to the variable x_2 by a column corresponding to the variable x_3, namely

$$\widehat{A}_K^{-1} A^3 = \begin{pmatrix} 1/2 & -1/2 & 0 \\ 1/6 & 1/6 & 0 \\ -3/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = \begin{pmatrix} -9/2 \\ 1/2 \\ 3/2 \end{pmatrix},$$

with

$$\overline{c}_3 = c_3 - z^* A^3 = 0 - \begin{pmatrix} 3/2 & 1/2 & -1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

and we get

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	$3/2$	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	$-9/2$	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/2$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	$3/2$	0	$-3/2$	$-1/2$	1

By analyzing the top row of reduced cost, we see that $j^+ = 2$. Furthermore, since $\min\{x_1/(1/2), \xi_3/(3/2)\} = \xi_3/(3/2) = 1/3$, we let $k^- = 6$, $K = (3, 1, 2)$, and pivot along the red circled $3/2$ to obtain

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
0	0	0	0	-1	-1	-1
$x_4 = 2$	0	0	1	-4	-2	3
$x_1 = 1/3$	1	0	0	$2/3$	$1/3$	$-1/3$
$x_3 = 1/3$	0	1	0	-1	$-1/3$	$2/3$

Since the upper left corner of the final tableau is zero and the reduced costs are all ≤ 0, we are finally finished. Then $y = (19/3 \; 8/3 \; -14/3)$ is an optimal solution of (D), but more importantly $(x_1, x_2, x_3, x_4) = (1/3, 0, 1/3, 2)$ is an optimal solution for our original linear program and provides an optimal value of $-10/3$.

The primal-dual algorithm for linear programming doesn't seem to be the favorite method to solve linear programs nowadays. But it is important because its basic principle, to use a restricted (simpler) primal problem involving an objective function with fixed weights, namely 1, and the dual problem to provide feedback to the primal by improving the objective function of the dual, has led to a whole class of combinatorial algorithms (often approximation algorithms) based on the primal-dual paradigm. The reader will get a taste of this kind of algorithm by consulting Papadimitriou and Steiglitz [134], where it is explained how classical algorithms such as Dijkstra's algorithm for the shortest path problem, and Ford and Fulkerson's algorithm for max flow can be derived from the primal-dual paradigm.

47.7 Summary

The main concepts and results of this chapter are listed below:

- Strictly separating hyperplane.
- Farkas–Minkowski proposition.
- Farkas lemma, version I, Farkas lemma, version II.
- Distance of a point to a subset.
- Dual linear program, primal linear program.
- Dual variables, primal variables.
- Complementary slackness conditions.
- Dual simplex algorithm.
- Primal-dual algorithm.
- Restricted primal linear program.

47.8 Problems

Problem 47.1. Let $(v_1, \ldots, v_n)$ be a sequence of n vectors in $\mathbb{R}^d$ and let V be the $d \times n$ matrix whose j-th column is v_j. Prove the equivalence of the following two statements:

(a) There is no nontrivial positive linear dependence among the v_j, which means that there is no nonzero vector, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, with $y_j \geq 0$ for $j = 1, \ldots, n$, so that

$$y_1 v_1 + \cdots + y_n v_n = 0$$

or equivalently, $Vy = 0$.

(b) There is some vector, $c \in \mathbb{R}^d$, so that $c^\top V > 0$, which means that $c^\top v_j > 0$, for $j = 1, \ldots, n$.

Problem 47.2. Check that the dual in maximization form (D'') of the Dual Program (D') (which is the dual of (P) in maximization form),

$$\begin{aligned} \text{maximize} \quad & -b^\top y^\top \\ \text{subject to} \quad & -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$, gives back the Primal Program (P).

Problem 47.3. In a General Linear Program (P) with n primal variables $x_1, \ldots, x_n$ and objective function $\sum_{j=1}^n c_j x_j$ (to be maximized), the m constraints are of the form

$$\begin{aligned} \sum_{j=1}^n a_{ij} x_j &\leq b_i, \\ \sum_{j=1}^n a_{ij} x_j &\geq b_i, \\ \sum_{j=1}^n a_{ij} x_j &= b_i, \end{aligned}$$

for $i = 1, \ldots, m$, and the variables x_j satisfy an inequality of the form

$$\begin{aligned} x_j &\geq 0, \\ x_j &\leq 0, \\ x_j &\in \mathbb{R}, \end{aligned}$$

for $j = 1, \ldots, n$. If $y_1, \ldots, y_m$ are the dual variables, show that the dual program of the linear program in standard form equivalent to (P) is equivalent to the linear program whose objective function is $\sum_{i=1}^m y_i b_i$ (to be minimized) and whose constraints are determined as follows:

$$\text{if} \quad \begin{Bmatrix} x_j \geq 0 \\ x_j \leq 0 \\ x_j \in \mathbb{R} \end{Bmatrix}, \quad \text{then} \quad \begin{Bmatrix} \sum_{i=1}^m a_{ij} y_i \geq c_j \\ \sum_{i=1}^m a_{ij} y_i \leq c_j \\ \sum_{i=1}^m a_{ij} y_i = c_j \end{Bmatrix},$$

and

$$\text{if} \quad \begin{Bmatrix} \sum_{j=1}^{n} a_{ij}x_j \leq b_i \\ \sum_{j=1}^{n} a_{ij}x_j \geq b_i \\ \sum_{j=1}^{n} a_{ij}x_j = b_i \end{Bmatrix}, \quad \text{then} \quad \begin{Bmatrix} y_i \geq 0 \\ y_i \leq 0 \\ y_i \in \mathbb{R} \end{Bmatrix}.$$

Problem 47.4. Apply the procedure of Problem 47.3 to show that the dual of the (general) linear program

$$\begin{aligned} &\text{maximize} \quad 3x_1 + 2x_2 + 5x_3 \\ &\text{subject to} \\ &\qquad\qquad 5x_1 + 3x_2 + x_3 = -8 \\ &\qquad\qquad 4x_1 + 2x_2 + 8x_3 \leq 23 \\ &\qquad\qquad 6x_1 + 7x_2 + 3x_3 \geq 1 \\ &\qquad\qquad x_1 \leq 4,\ x_3 \geq 0 \end{aligned}$$

is the (general) linear program:

$$\begin{aligned} &\text{minimize} \quad -8y_1 + 23y_2 - y_3 + 4y_4 \\ &\text{subject to} \\ &\qquad\qquad 5y_1 + 4y_2 - 6y_3 + y_4 = 3 \\ &\qquad\qquad 3y_1 + 2y_2 - 7y_3 = 2 \\ &\qquad\qquad y_1 + 8y_2 - 3y_3 \geq 5 \\ &\qquad\qquad y_2,\ y_3,\ y_4 \geq 0. \end{aligned}$$

Problem 47.5. (1) Prove that the dual of the (general) linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \ \text{and} \ x \in \mathbb{R}^n \end{aligned}$$

is

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA = c \ \text{and} \ y \in \mathbb{R}^m. \end{aligned}$$

(2) Prove that the dual of the (general) linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \geq b \ \text{and} \ x \geq 0 \end{aligned}$$

is

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \ \text{ and } \ y \leq 0. \end{aligned}$$

Problem 47.6. Use the complementary slackness conditions to confirm that

$$x_1 = 2,\ x_2 = 4,\ x_3 = 0,\ x_4 = 0,\ x_5 = 7,\ x_6 = 0$$

is an optimal solution of the following linear program (from Chavatal [40], Chapter 5):

$$\begin{aligned} &\text{maximize} \quad 18x_1 - 7x_2 + 12x_3 + 5x_4 + 8x_6 \\ &\text{subject to} \end{aligned}$$

$$\begin{aligned} 2x_1 - 6x_2 + 2x_3 + 7x_4 + 3x_5 + 8x_6 &\leq 1 \\ -3x_1 - x_2 + 4x_3 - 3x_4 + x_5 + 2x_6 &\leq -2 \\ 8x_1 - 3x_2 + 5x_3 - 2x_4 + 2x_6 &\leq 4 \\ 4x_1 + 8x_3 + 7x_4 - x_5 + 3x_6 &\leq 1 \\ 5x_1 + 2x_2 - 3x_3 + 6x_4 - 2x_5 - x_6 &\leq 5 \\ x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0. \end{aligned}$$

Problem 47.7. Check carefully that the dual simplex method is equivalent to the simplex method applied to the dual program in maximization form.

Part VIII

NonLinear Optimization

Chapter 48

Basics of Hilbert Spaces

Most of the "deep" results about the existence of minima of real-valued functions proven in Chapter 49 rely on two fundamental results of Hilbert space theory:

(1) The projection lemma, which is a result about nonempty, closed, convex subsets of a Hilbert space V.

(2) The Riesz representation theorem, which allows us to express a continuous linear form on a Hilbert space V in terms of a vector in V and the inner product on V.

The correctness of the Karush–Kuhn–Tucker conditions appearing in Lagrangian duality follows from a version of the Farkas–Minkowski proposition, which also follows from the projection lemma.

Thus, we feel that it is indispensable to review some basic results of Hilbert space theory, although in most applications considered here the Hilbert space in question will be finite-dimensional. However, in optimization theory, there are many problems where we seek to find a *function* minimizing some type of energy functional (often given by a bilinear form), in which case we are dealing with an infinite dimensional Hilbert space, so it necessary to develop tools to deal with the more general situation of infinite-dimensional Hilbert spaces.

48.1 The Projection Lemma

Given a Hermitian space $\langle E, \varphi \rangle$, we showed in Section 14.1 that the function $\| \, \| \colon E \to \mathbb{R}$ defined such that $\|u\| = \sqrt{\varphi(u,u)}$, is a norm on E. Thus, E is a normed vector space. If E is also complete, then it is a very interesting space.

Recall that completeness has to do with the convergence of Cauchy sequences. A normed vector space $\langle E, \| \, \| \rangle$ is automatically a metric space under the metric d defined such that $d(u,v) = \|v - u\|$ (see Chapter 37 for the definition of a normed vector space and of a metric space, or Lang [111, 112], or Dixmier [51]). Given a metric space E with metric d, a sequence

$(a_n)_{n\geq 1}$ of elements $a_n \in E$ is a *Cauchy sequence* iff for every $\epsilon > 0$, there is some $N \geq 1$ such that

$$d(a_m, a_n) < \epsilon \quad \text{for all} \quad m, n \geq N.$$

We say that E is *complete* iff every Cauchy sequence converges to a limit (which is unique, since a metric space is Hausdorff).

Every finite dimensional vector space over $\mathbb{R}$ *or* $\mathbb{C}$ *is complete.* For example, one can show by induction that given any basis $(e_1, \ldots, e_n)$ of E, the linear map $h\colon \mathbb{C}^n \to E$ defined such that

$$h((z_1, \ldots, z_n)) = z_1 e_1 + \cdots + z_n e_n$$

is a homeomorphism (using the *sup*-norm on $\mathbb{C}^n$). One can also use the fact that any two norms on a finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$ are equivalent (see Chapter 9, or Lang [112], Dixmier [51], Schwartz [150]).

However, if E has infinite dimension, it may not be complete. When a Hermitian space is complete, a number of the properties that hold for finite dimensional Hermitian spaces also hold for infinite dimensional spaces. For example, any closed subspace has an orthogonal complement, and in particular, a finite dimensional subspace has an orthogonal complement. Hermitian spaces that are also complete play an important role in analysis. Since they were first studied by Hilbert, they are called Hilbert spaces.

Definition 48.1. A (complex) Hermitian space $\langle E, \varphi\rangle$ which is a complete normed vector space under the norm $\| \ \|$ induced by φ is called a *Hilbert space*. A real Euclidean space $\langle E, \varphi\rangle$ which is complete under the norm $\| \ \|$ induced by φ is called a *real Hilbert space*.

All the results in this section hold for complex Hilbert spaces as well as for real Hilbert spaces. We state all results for the complex case only, since they also apply to the real case, and since the proofs in the complex case need a little more care.

Example 48.1. The space ℓ^2 of all countably infinite sequences $x = (x_i)_{i\in\mathbb{N}}$ of complex numbers such that $\sum_{i=0}^{\infty} |x_i|^2 < \infty$ is a Hilbert space. It will be shown later that the map $\varphi\colon \ell^2 \times \ell^2 \to \mathbb{C}$ defined such that

$$\varphi\left((x_i)_{i\in\mathbb{N}}, (y_i)_{i\in\mathbb{N}}\right) = \sum_{i=0}^{\infty} x_i \overline{y_i}$$

is well defined, and that ℓ^2 is a Hilbert space under φ. In fact, we will prove a more general result (Proposition A.3).

Example 48.2. The set $\mathcal{C}^\infty[a, b]$ of smooth functions $f\colon [a, b] \to \mathbb{C}$ is a Hermitian space under the Hermitian form

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx,$$

but it is not a Hilbert space because it is not complete. It is possible to construct its completion $\mathrm{L}^2([a, b])$, which turns out to be the space of Lebesgue integrable functions on $[a, b]$.

Theorem 37.63 yields a quick proof of the fact that any Hermitian space E (with Hermitian product $\langle -, - \rangle$) can be embedded in a Hilbert space E_h.

Theorem 48.1. *Given a Hermitian space $(E, \langle -, - \rangle)$ (resp. Euclidean space), there is a Hilbert space $(E_h, \langle -, - \rangle_h)$ and a linear map $\varphi \colon E \to E_h$, such that*

$$\langle u, v \rangle = \langle \varphi(u), \varphi(v) \rangle_h$$

for all $u, v \in E$, and $\varphi(E)$ is dense in E_h. Furthermore, E_h is unique up to isomorphism.

Proof. Let $(\widehat{E}, \| \, \|_{\widehat{E}})$ be the Banach space, and let $\varphi \colon E \to \widehat{E}$ be the linear isometry, given by Theorem 37.63. Let $\|u\| = \sqrt{\langle u, u \rangle}$ (with $u \in E$) and $E_h = \widehat{E}$. If E is a real vector space, we know from Section refsec5bis that the inner product $\langle -, - \rangle$ can be expressed in terms of the norm $\|u\|$ by the polarity equation

$$\langle u, v \rangle = \frac{1}{2}(\|u + v\|^2 - \|u\|^2 - \|v\|^2),$$

and if E is a complex vector space, we know from Section 14.1 that we have the polarity equation

$$\langle u, v \rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2).$$

By the Cauchy-Schwarz inequality, $|\langle u, v \rangle| \leq \|u\| \|v\|$, the map $\langle -, - \rangle \colon E \times E \to \mathbb{C}$ (resp. $\langle -, - \rangle \colon E \times E \to \mathbb{R}$) is continuous. However, it is not uniformly continuous, but we can get around this problem by using the polarity equations to extend it to a continuous map. By continuity, the polarity equations also hold in E_h, which shows that $\langle -, - \rangle$ extends to a positive definite Hermitian inner product (resp. Euclidean inner product) $\langle -, - \rangle_h$ on E_h induced by $\| \, \|_{\widehat{E}}$ extending $\langle -, - \rangle$. ☐

Remark: We followed the approach in Schwartz [149] (Chapter XXIII, Section 42. Theorem 2). For other approaches, see Munkres [131] (Chapter 7, Section 43), and Bourbaki [27].

One of the most important facts about finite-dimensional Hermitian (and Euclidean) spaces is that they have orthonormal bases. This implies that, up to isomorphism, *every finite-dimensional Hermitian space is isomorphic to* $\mathbb{C}^n$ (for some $n \in \mathbb{N}$) and that the inner product is given by

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i \overline{y_i}.$$

Furthermore, every subspace W has an orthogonal complement $W^{\perp}$, and the inner product induces a natural duality between E and E^* (actually, between $\overline{E}$ and E^*) where E^* is the space of linear forms on E.

When E is a Hilbert space, E may be infinite dimensional, often of uncountable dimension. Thus, we can't expect that E always have an orthonormal basis. However, if we modify the notion of basis so that a "Hilbert basis" is an orthogonal family that is also dense in E, i.e., every $v \in E$ is the limit of a sequence of finite combinations of vectors from the Hilbert basis, then we can recover most of the "nice" properties of finite-dimensional Hermitian spaces. For instance, if $(u_k)_{k\in K}$ is a Hilbert basis, for every $v \in E$, we can define the Fourier coefficients $c_k = \langle v, u_k\rangle/\|u_k\|$, and then, v is the "sum" of its Fourier series $\sum_{k\in K} c_k u_k$. However, the cardinality of the index set K can be very large, and it is necessary to define what it means for a family of vectors indexed by K to be summable. We will do this in Section A.1. It turns out that every Hilbert space is isomorphic to a space of the form $\ell^2(K)$, where $\ell^2(K)$ is a generalization of the space of Example 48.1 (see Theorem A.8, usually called the Riesz-Fischer theorem).

Our first goal is to prove that a closed subspace of a Hilbert space has an orthogonal complement. We also show that duality holds if we redefine the dual E' of E to be the space of *continuous* linear maps on E. Our presentation closely follows Bourbaki [27]. We also were inspired by Rudin [140], Lang [111, 112], Schwartz [150, 149], and Dixmier [51]. In fact, we highly recommend Dixmier [51] as a clear and simple text on the basics of topology and analysis. To achieve this goal, we must first prove the so-called projection lemma.

Recall that in a metric space E, a subset X of E is *closed* iff for every convergent sequence (x_n) of points $x_n \in X$, the limit $x = \lim_{n\to\infty} x_n$ also belongs to X. The *closure* $\overline{X}$ *of* X is the set of all limits of convergent sequences (x_n) of points $x_n \in X$. Obviously, $X \subseteq \overline{X}$. We say that the subset X of E is *dense in* E iff $E = \overline{X}$, the closure of X, which means that every $a \in E$ is the limit of some sequence (x_n) of points $x_n \in X$. Convex sets will again play a crucial role. In a complex vector space E, a subset $C \subseteq E$ is convex if $(1-\lambda)x + \lambda y \in C$ for all $x, y \in C$ and all *real* $\lambda \in [0, 1]$. Observe that a subspace is convex.

First we state the following easy "parallelogram law," whose proof is left as an exercise.

Proposition 48.2. *If E is a Hermitian space, for any two vectors $u, v \in E$, we have*

$$\|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

From the above, we get the following proposition:

Proposition 48.3. *If E is a Hermitian space, given any $d, \delta \in \mathbb{R}$ such that $0 \leq \delta < d$, let*

$$B = \{u \in E \mid \|u\| < d\} \quad \text{and} \quad C = \{u \in E \mid \|u\| \leq d + \delta\}.$$

For any convex set such A that $A \subseteq C - B$, we have

$$\|v-u\| \leq \sqrt{12d\delta},$$

for all $u, v \in A$ (see Figure 48.1).

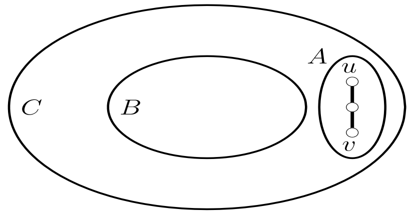


Figure 48.1: Inequality of Proposition 48.3.

Proof. Since A is convex, $\frac{1}{2}(u+v) \in A$ if $u, v \in A$, and thus, $\|\frac{1}{2}(u+v)\| \geq d$. From the parallelogram equality written in the form

$$\left\|\frac{1}{2}(u+v)\right\|^2 + \left\|\frac{1}{2}(u-v)\right\|^2 = \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right),$$

since $\delta < d$, we get

$$\left\|\frac{1}{2}(u-v)\right\|^2 = \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right) - \left\|\frac{1}{2}(u+v)\right\|^2 \leq (d+\delta)^2 - d^2 = 2d\delta + \delta^2 \leq 3d\delta,$$

from which

$$\|v - u\| \leq \sqrt{12 d\delta}.$$

$\square$

Definition 48.2. If X is a nonempty subset of a metric space (E, d), for any $a \in E$, recall that we define the *distance* $d(a, X)$ *of* a *to* X as

$$d(a, X) = \inf_{b \in X} d(a, b).$$

Also, the *diameter* $\delta(X)$ *of* X is defined by

$$\delta(X) = \sup\{d(a, b) \mid a, b \in X\}.$$

It is possible that $\delta(X) = \infty$.

We leave the following standard two facts as an exercise (see Dixmier [51]):

Proposition 48.4. *Let E be a metric space.*

(1) For every subset $X \subseteq E$, $\delta(X) = \delta(\overline{X})$.

(2) If E is a complete metric space, for every sequence (F_n) of closed nonempty subsets of E such that $F_{n+1} \subseteq F_n$, if $\lim_{n\to\infty} \delta(F_n) = 0$, then $\bigcap_{n=1}^{\infty} F_n$ consists of a single point.

We are now ready to prove the crucial projection lemma.

Proposition 48.5. *(Projection lemma) Let E be a Hilbert space and let $X \subseteq E$ be any nonempty convex and closed subset.*

(1) For any $u \in E$, there is a unique vector $p_X(u) \in X$ such that

$$\|u - p_X(u)\| = \inf_{v \in X} \|u - v\| = d(u, X).$$

See Figure 48.2.

(2) The vector $p_X(u)$ is the unique vector $w \in E$ satisfying the following property (see Figure 48.3):

$$w \in X \quad \text{and} \quad \Re \langle u - w, z - w \rangle \leq 0 \quad \text{for all } z \in X. \tag{$*$}$$

(3) If X is a nonempty closed subspace of E, then the vector $p_X(u)$ is the unique vector $w \in E$ satisfying the following property:

$$w \in X \quad \text{and} \quad \langle u - w, z \rangle = 0 \quad \text{for all } z \in X. \tag{$**$}$$

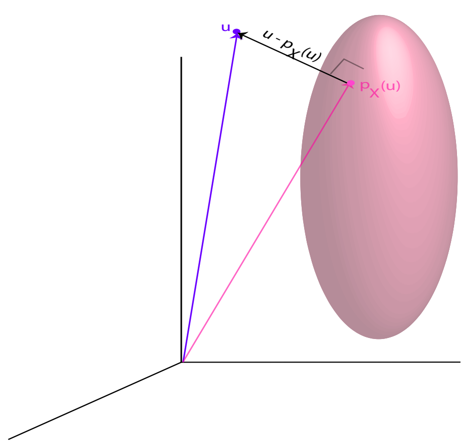


Figure 48.2: Let X be the solid pink ellipsoid. The projection of the purple point u onto X is the magenta point $p_X(u)$.

Proof. (1) Let $d = \inf_{v \in X} \|u - v\| = d(u, X)$. We define a sequence X_n of subsets of X as follows: for every $n \geq 1$,

$$X_n = \left\{ v \in X \mid \|u - v\| \leq d + \frac{1}{n} \right\}.$$

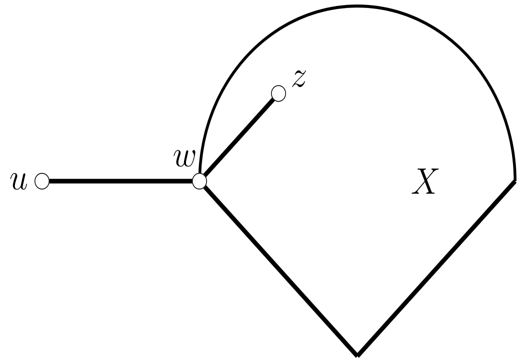


Figure 48.3: Inequality of Proposition 48.5.

It is immediately verified that each X_n is nonempty (by definition of d), convex, and that $X_{n+1} \subseteq X_n$. Also, by Proposition 48.3, (where $B = \{v \in E \mid \|u - v\| \leq d\}$, $C = \left\{v \in E \mid \|u - v\| \leq d + \frac{1}{n}\right\}$, and $A = X_n$), we have

$$\sup\{\|z - v\| \mid v, z \in X_n\} \leq \sqrt{12d/n},$$

and thus, $\bigcap_{n \geq 1} X_n$ contains at most one point; see Proposition 48.4(2). We will prove that $\bigcap_{n \geq 1} X_n$ contains exactly one point, namely, $p_X(u)$. For this, define a sequence $(w_n)_{n \geq 1}$ by picking some $w_n \in X_n$ for every $n \geq 1$. We claim that $(w_n)_{n \geq 1}$ is a Cauchy sequence. Given any $\epsilon > 0$, if we pick N such that

$$N > \frac{12d}{\epsilon^2},$$

since $(X_n)_{n \geq 1}$ is a monotonic decreasing sequence, which means that $X_{n+1} \subseteq X_n$ for all $n \geq 1$, for all $m, n \geq N$, we have

$$\|w_m - w_n\| \leq \sqrt{12d/N} < \epsilon,$$

as desired. Since E is complete, the sequence $(w_n)_{n \geq 1}$ has a limit w, and since $w_n \in X$ and X is closed, we must have $w \in X$. Also observe that

$$\|u - w\| \leq \|u - w_n\| + \|w_n - w\|,$$

and since w is the limit of $(w_n)_{n \geq 1}$ and

$$\|u - w_n\| \leq d + \frac{1}{n},$$

given any $\epsilon > 0$, there is some n large enough so that

$$\frac{1}{n} < \frac{\epsilon}{2} \quad \text{and} \quad \|w_n - w\| \leq \frac{\epsilon}{2},$$

and thus
$$\|u - w\| \leq d + \epsilon.$$
Since the above holds for every $\epsilon > 0$, we have $\|u - w\| = d$. Thus, $w \in X_n$ for all $n \geq 1$, which proves that $\bigcap_{n\geq 1} X_n = \{w\}$. Now any $z \in X$ such that $\|u - z\| = d(u, X) = d$ also belongs to every X_n, and thus $z = w$, proving the uniqueness of w, which we denote as $p_X(u)$. See Figure 48.4.

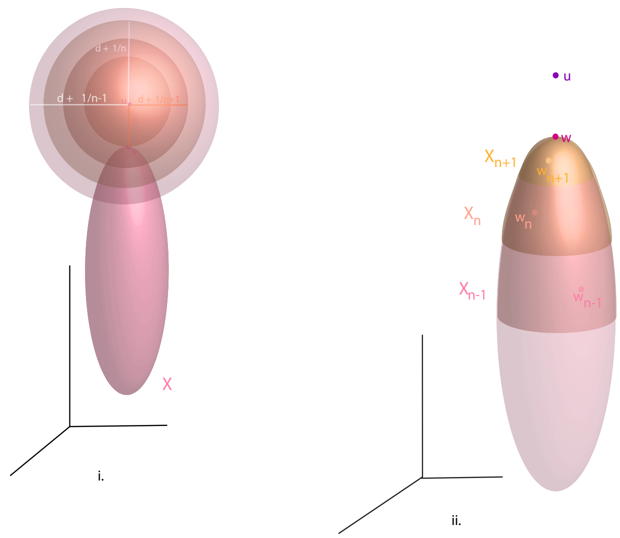


Figure 48.4: Let X be the solid pink ellipsoid with $p_X(u) = w$ at its apex. Each X_n is the intersection of X and a solid sphere centered at u with radius $d + 1/n$. These intersections are the colored "caps" of Figure ii. The Cauchy sequence $(w_n)_{n\geq 1}$ is obtained by selecting a point in each colored X_n.

(2) Let $z \in X$. Since X is convex, $v = (1 - \lambda)p_X(u) + \lambda z \in X$ for every λ, $0 \leq \lambda \leq 1$. Then by the definition of u, we have
$$\|u - v\| \geq \|u - p_X(u)\|$$
for all λ, $0 \leq \lambda \leq 1$, and since
$$\begin{aligned}\|u - v\|^2 &= \|u - p_X(u) - \lambda(z - p_X(u))\|^2 \\ &= \|u - p_X(u)\|^2 + \lambda^2\|z - p_X(u)\|^2 - 2\lambda \Re \langle u - p_X(u), z - p_X(u)\rangle,\end{aligned}$$
for all λ, $0 < \lambda \leq 1$, we get
$$\Re \langle u - p_X(u), z - p_X(u)\rangle = \frac{1}{2\lambda}\left(\|u - p_X(u)\|^2 - \|u - v\|^2\right) + \frac{\lambda}{2}\|z - p_X(u)\|^2. \qquad (\dagger)$$
Since
$$\|u - v\| \geq \|u - p_X(u)\|,$$

we have

$$\|u - p_X(u)\|^2 - \|u - v\|^2 = (\|u - p_X(u)\| - \|u - v\|)(\|u - p_X(u)\| + \|u - v\|) \leq 0,$$

and since Equation (†) holds for all λ such that $0 < \lambda \leq 1$, if $\|u - p_X(u)\|^2 - \|u - v\|^2 < 0$, then for $\lambda > 0$ small enough we have

$$\frac{1}{2\lambda}\left(\|u - p_X(u)\|^2 - \|u - v\|^2\right) + \frac{\lambda}{2}\|z - p_X(u)\|^2 < 0,$$

and if $\|u - p_X(u)\|^2 - \|u - v\|^2 = 0$, then the limit of $\frac{\lambda}{2}\|z - p_X(u)\|^2$ as $\lambda > 0$ goes to zero is zero, so in all cases, by (†), we have

$$\Re\langle u - p_X(u), z - p_X(u)\rangle \leq 0.$$

Conversely, assume that $w \in X$ satisfies the condition

$$\Re\langle u - w, z - w\rangle \leq 0$$

for all $z \in X$. For all $z \in X$, we have

$$\|u - z\|^2 = \|u - w\|^2 + \|z - w\|^2 - 2\Re\langle u - w, z - w\rangle \geq \|u - w\|^2,$$

which implies that $\|u - w\| = d(u, X) = d$, and from (1), that $w = p_X(u)$.

(3) If X is a subspace of E and $w \in X$, when z ranges over X the vector $z - w$ also ranges over the whole of X so Condition $(*)$ is equivalent to

$$w \in X \quad \text{and} \quad \Re\langle u - w, z\rangle \leq 0 \quad \text{for all } z \in X. \tag{$*_1$}$$

Since X is a subspace, if $z \in X$, then $-z \in X$, which implies that $(*_1)$ is equivalent to

$$w \in X \quad \text{and} \quad \Re\langle u - w, z\rangle = 0 \quad \text{for all } z \in X. \tag{$*_2$}$$

Finally, since X is a subspace, if $z \in X$,then $iz \in X$, and this implies that

$$0 = \Re\langle u - w, iz\rangle = -i\Im\langle u - w, z\rangle,$$

so $\Im\langle u - w, z\rangle = 0$, but since we also have $\Re\langle u - w, z\rangle = 0$, we see that $(*_2)$ is equivalent to

$$w \in X \quad \text{and} \quad \langle u - w, z\rangle = 0 \quad \text{for all } z \in X, \tag{$**$}$$

as claimed. □

Definition 48.3. The vector $p_X(u)$ is called the *projection of u onto X*, and the map $p_X \colon E \to X$ is called the *projection of E onto X*.

In the case of a real Hilbert space, there is an intuitive geometric interpretation of the condition

$$\langle u - p_X(u), z - p_X(u)\rangle \leq 0$$

for all $z \in X$. If we restate the condition as

$$\langle u - p_X(u), p_X(u) - z\rangle \geq 0$$

for all $z \in X$, this says that the absolute value of the measure of the angle between the vectors $u - p_X(u)$ and $p_X(u) - z$ is at most $\pi/2$. See Figure 48.5. This makes sense, since X is convex, and points in X must be on the side opposite to the "tangent space" to X at $p_X(u)$, which is orthogonal to $u - p_X(u)$. Of course, this is only an intuitive description, since the notion of tangent space has not been defined!

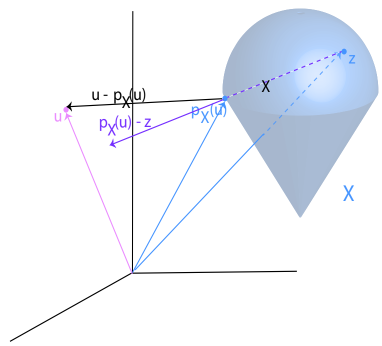


Figure 48.5: Let X be the solid blue ice cream cone. The acute angle between the black vector $u - p_X(u)$ and the purple vector $p_X(u) - z$ is less than $\pi/2$.

If X is a closed subspace of E, then Condition $(**)$ says that the vector $u - p_X(u)$ is *orthogonal* to X, in the sense that $u - p_X(u)$ is orthogonal to every vector $z \in X$.

The map $p_X \colon E \to X$ is continuous as shown below.

Proposition 48.6. *Let E be a Hilbert space. For any nonempty convex and closed subset $X \subseteq E$, the map $p_X \colon E \to X$ is continuous. In fact, p_X satisfies the Lipschitz condition*

$$\|p_X(v) - p_X(u)\| \leq \|v - u\| \quad \textit{for all } u, v \in E.$$

Proof. For any two vectors $u, v \in E$, let $x = p_X(u) - u$, $y = p_X(v) - p_X(u)$, and $z = v - p_X(v)$. Clearly, (as illustrated in Figure 48.6),

$$v - u = x + y + z,$$

and from Proposition 48.5(2), we also have

$$\Re\langle x, y\rangle \geq 0 \quad \text{and} \quad \Re\langle z, y\rangle \geq 0,$$

from which we get

$$\begin{aligned}\|v-u\|^2 &= \|x+y+z\|^2 = \|x+z+y\|^2 \\ &= \|x+z\|^2 + \|y\|^2 + 2\Re\langle x, y\rangle + 2\Re\langle z, y\rangle \\ &\geq \|y\|^2 = \|p_X(v) - p_X(u)\|^2.\end{aligned}$$

However, $\|p_X(v) - p_X(u)\| \leq \|v-u\|$ obviously implies that p_X is continuous. $\square$

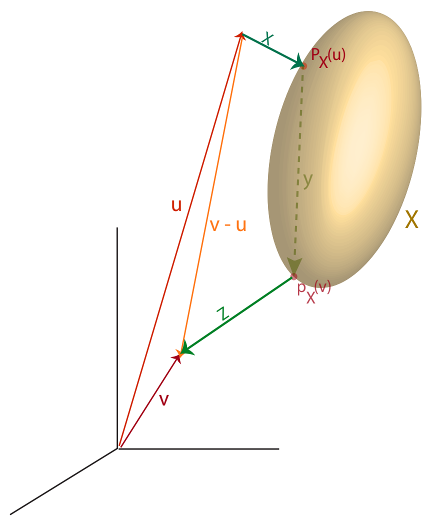


Figure 48.6: Let X be the solid gold ellipsoid. The vector $v-u$ is the sum of the three green vectors, each of which is determined by the appropriate projections.

We can now prove the following important proposition.

Proposition 48.7. *Let E be a Hilbert space.*

(1) For any closed subspace $V \subseteq E$, we have $E = V \oplus V^{\perp}$, and the map $p_V\colon E \to V$ is linear and continuous.

(2) For any $u \in E$, the projection $p_V(u)$ is the unique vector $w \in E$ such that

$$w \in V \quad \textit{and} \quad \langle u - w, z\rangle = 0 \quad \textit{for all } z \in V.$$

Proof. (1) First, we prove that $u - p_V(u) \in V^{\perp}$ for all $u \in E$. For any $v \in V$, since V is a subspace, $z = p_V(u) + \lambda v \in V$ for all $\lambda \in \mathbb{C}$, and since V is convex and nonempty (since it is a subspace), and closed by hypothesis, by Proposition 48.5(2), we have

$$\Re(\overline{\lambda}\langle u - p_V(u), v\rangle) = \Re(\langle u - p_V(u), \lambda v\rangle = \Re\langle u - p_V(u), z - p_V(u)\rangle \leq 0$$

for all $\lambda \in \mathbb{C}$. In particular, the above holds for $\lambda = \langle u - p_V(u), v\rangle$, which yields

$$|\langle u - p_V(u), v\rangle| \leq 0,$$

and thus, $\langle u - p_V(u), v\rangle = 0$. See Figure 48.7. As a consequence, $u - p_V(u) \in V^\perp$ for all $u \in E$. Since $u = p_V(u) + u - p_V(u)$ for every $u \in E$, we have $E = V + V^\perp$. On the other hand, since $\langle -, -\rangle$ is positive definite, $V \cap V^\perp = \{0\}$, and thus $E = V \oplus V^\perp$.

We already proved in Proposition 48.6 that $p_V \colon E \to V$ is continuous. Also, since

$$p_V(\lambda u + \mu v) - (\lambda p_V(u) + \mu p_V(v)) = p_V(\lambda u + \mu v) - (\lambda u + \mu v) + \lambda(u - p_V(u)) + \mu(v - p_V(v)),$$

for all $u, v \in E$, and since the left-hand side term belongs to V, and from what we just showed, the right-hand side term belongs to $V^\perp$, we have

$$p_V(\lambda u + \mu v) - (\lambda p_V(u) + \mu p_V(v)) = 0,$$

showing that p_V is linear.

(2) This is basically obvious from (1). We proved in (1) that $u - p_V(u) \in V^\perp$, which is exactly the condition

$$\langle u - p_V(u), z\rangle = 0$$

for all $z \in V$. Conversely, if $w \in V$ satisfies the condition

$$\langle u - w, z\rangle = 0$$

for all $z \in V$, since $w \in V$, every vector $z \in V$ is of the form $y - w$, with $y = z + w \in V$, and thus, we have

$$\langle u - w, y - w\rangle = 0$$

for all $y \in V$, which implies the condition of Proposition 48.5(2):

$$\Re\langle u - w, y - w\rangle \leq 0$$

for all $y \in V$. By Proposition 48.5, $w = p_V(u)$ is the projection of u onto V. $\square$

Remark: If $p_V \colon E \to V$ is linear, then V is a subspace of E. It follows that if V is a closed convex subset of E, then $p_V \colon E \to V$ is linear iff V is a subspace of E.

Example 48.3. Let us illustrate the power of Proposition 48.7 on the following "least squares" problem. Given a real $m \times n$-matrix A and some vector $b \in \mathbb{R}^m$, we would like to solve the linear system

$$Ax = b$$

in the least-squares sense, which means that we would like to find some solution $x \in \mathbb{R}^n$ that minimizes the Euclidean norm $\|Ax - b\|$ of the error $Ax - b$. It is actually not clear that the

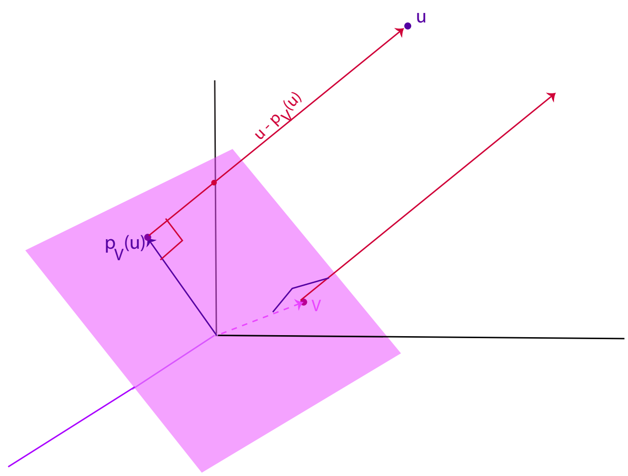


Figure 48.7: Let V be the pink plane. The vector $u - p_V(u)$ is perpendicular to any $v \in V$.

problem has a solution, but it does! The problem can be restated as follows: Is there some $x \in \mathbb{R}^n$ such that

$$\|Ax - b\| = \inf_{y \in \mathbb{R}^n} \|Ay - b\|,$$

or equivalently, is there some $z \in \mathrm{Im}\,(A)$ such that

$$\|z - b\| = d(b, \mathrm{Im}\,(A)),$$

where $\mathrm{Im}\,(A) = \{Ay \in \mathbb{R}^m \mid y \in \mathbb{R}^n\}$, the image of the linear map induced by A. Since $\mathrm{Im}\,(A)$ is a closed subspace of $\mathbb{R}^m$, because we are in finite dimension, Proposition 48.7 tells us that there is a unique $z \in \mathrm{Im}\,(A)$ such that

$$\|z - b\| = \inf_{y \in \mathbb{R}^n} \|Ay - b\|,$$

and thus the problem always has a solution since $z \in \mathrm{Im}\,(A)$, and since there is at least some $x \in \mathbb{R}^n$ such that $Ax = z$ (by definition of $\mathrm{Im}\,(A)$). Note that such an x is not necessarily unique. Furthermore, Proposition 48.7 also tells us that $z \in \mathrm{Im}\,(A)$ is the solution of the equation

$$\langle z - b, w \rangle = 0 \quad \text{for all } w \in \mathrm{Im}\,(A),$$

or equivalently, that $x \in \mathbb{R}^n$ is the solution of

$$\langle Ax - b, Ay \rangle = 0 \quad \text{for all } y \in \mathbb{R}^n,$$

which is equivalent to

$$\langle A^\top (Ax - b), y \rangle = 0 \quad \text{for all } y \in \mathbb{R}^n,$$

and thus, since the inner product is positive definite, to $A^\top(Ax - b) = 0$, i.e.,

$$A^\top A x = A^\top b.$$

Therefore, the solutions of the original least-squares problem are precisely the solutions of the the so-called *normal equations*

$$A^\top Ax = A^\top b,$$

discovered by Gauss and Legendre around 1800. We also proved that the normal equations always have a solution.

Computationally, it is best not to solve the normal equations directly, and instead, to use methods such as the QR-decomposition (applied to A) or the SVD-decomposition (in the form of the pseudo-inverse). We will come back to this point later on.

Here is another important corollary of Proposition 48.7.

Corollary 48.8. *For any continuous nonnull linear map $h\colon E \to \mathbb{C}$, the null space*

$$H = \mathrm{Ker}\, h = \{u \in E \mid h(u) = 0\} = h^{-1}(0)$$

is a closed hyperplane H, and thus, $H^\perp$ is a subspace of dimension one such that $E = H \oplus H^\perp$.

The above suggests defining the dual space of E as the set of all continuous maps $h\colon E \to \mathbb{C}$.

Remark: If $h\colon E \to \mathbb{C}$ is a linear map which is **not** continuous, then it can be shown that the hyperplane $H = \mathrm{Ker}\, h$ is dense in E! Thus, $H^\perp$ is reduced to the trivial subspace $\{0\}$. This goes against our intuition of what a hyperplane in $\mathbb{R}^n$ (or $\mathbb{C}^n$) is, and warns us not to trust our "physical" intuition too much when dealing with infinite dimensions. As a consequence, the map $\flat\colon E \to E^*$ introduced in Section 14.2 (see just after Definition 48.4 below) is not surjective, since the linear forms of the form $u \mapsto \langle u, v\rangle$ (for some fixed vector $v \in E$) are continuous (the inner product is continuous).

48.2 Duality and the Riesz Representation Theorem

We now show that by redefining the dual space of a Hilbert space as the set of continuous linear forms on E we recover Theorem 14.6.

Definition 48.4. Given a Hilbert space E, we define the *dual space E' of E* as the vector space of all continuous linear forms $h\colon E \to \mathbb{C}$. Maps in E' are also called *bounded linear operators, bounded linear functionals, or simply operators or functionals*.

As in Section 14.2, for all $u, v \in E$, we define the maps $\varphi^l_u\colon E \to \mathbb{C}$ and $\varphi^r_v\colon E \to \mathbb{C}$ such that

$$\varphi^l_u(v) = \overline{\langle u, v\rangle},$$

and

$$\varphi^r_v(u) = \langle u, v\rangle\,.$$

In fact, $\varphi_u^l = \varphi_u^r$, and because the inner product $\langle -, - \rangle$ is continuous, it is obvious that φ_v^r is continuous and linear, so that $\varphi_v^r \in E'$. To simplify notation, we write φ_v instead of φ_v^r.

Theorem 14.6 is generalized to Hilbert spaces as follows.

Proposition 48.9. *(Riesz representation theorem) Let E be a Hilbert space. Then the map $\flat\colon E \to E'$ defined such that*

$$\flat(v) = \varphi_v,$$

is semilinear, continuous, and bijective. Furthermore, for any continuous linear map $\psi \in E'$, if $u \in E$ is the unique vector such that

$$\psi(v) = \langle v, u \rangle \quad \text{for all } v \in E,$$

then we have $\|\psi\| = \|u\|$, where

$$\|\psi\| = \sup \left\{ \frac{|\psi(v)|}{\|v\|} \;\middle|\; v \in E,\ v \neq 0 \right\}.$$

Proof. The proof is basically identical to the proof of Theorem 14.6, except that a different argument is required for the surjectivity of $\flat\colon E \to E'$, since E may not be finite dimensional. For any nonnull linear operator $h \in E'$, the hyperplane $H = \operatorname{Ker} h = h^{-1}(0)$ is a closed subspace of E, and by Proposition 48.7, $H^{\perp}$ is a subspace of dimension one such that $E = H \oplus H^{\perp}$. Then picking any nonnull vector $w \in H^{\perp}$, observe that H is also the kernel of the linear operator φ_w, with

$$\varphi_w(u) = \langle u, w \rangle,$$

and thus, since any two nonzero linear forms defining the same hyperplane must be proportional, there is some nonzero scalar $\lambda \in \mathbb{C}$ such that $h = \lambda\varphi_w$. But then, $h = \varphi_{\overline{\lambda}w}$, proving that $\flat\colon E \to E'$ is surjective.

By the Cauchy–Schwarz inequality we have

$$|\psi(v)| = |\langle v, u \rangle| \leq \|v\| \, \|u\|,$$

so by definition of $\|\psi\|$ we get

$$\|\psi\| \leq \|u\|.$$

Obviously $\psi = 0$ iff $u = 0$ so assume $u \neq 0$. We have

$$\|u\|^2 = \langle u, u \rangle = \psi(u) \leq \|\psi\| \, \|u\|,$$

which yields $\|u\| \leq \|\psi\|$, and therefore $\|\psi\| = \|u\|$, as claimed. ∎

Proposition 48.9 is known as the *Riesz representation theorem or "Little Riesz Theorem."* It shows that the inner product on a Hilbert space induces a natural semilinear isomorphism between E and its dual E' (equivalently, a linear isomorphism between $\overline{E}$ and E'). This isomorphism is an isometry (it is preserves the norm).

Remark: Many books on quantum mechanics use the so-called Dirac notation to denote objects in the Hilbert space E and operators in its dual space E'. In the Dirac notation, an element of E is denoted as $|x\rangle$, and an element of E' is denoted as $\langle t|$. The scalar product is denoted as $\langle t| \cdot |x\rangle$. This uses the isomorphism between E and E', except that the inner product is assumed to be semi-linear on the left rather than on the right.

Proposition 48.9 allows us to define the adjoint of a linear map, as in the Hermitian case (see Proposition 14.8). Actually, we can prove a slightly more general result which is used in optimization theory.

If $\varphi\colon E \times E \to \mathbb{C}$ is a sesquilinear map on a normed vector space $(E, \| \|)$, then Proposition 37.59 is immediately adapted to prove that φ is continuous iff there is some constant $k \geq 0$ such that

$$|\varphi(u, v)| \leq k \, \|u\| \, \|v\| \quad \text{for all } u, v \in E.$$

Thus we define $\|\varphi\|$ as in Definition 37.42 by

$$\|\varphi\| = \sup \{ |\varphi(x, y)| \mid \|x\| \leq 1, \ \|y\| \leq 1, \ x, y \in E \}.$$

Proposition 48.10. *Given a Hilbert space E, for every continuous sesquilinear map $\varphi\colon E \times E \to \mathbb{C}$, there is a unique continuous linear map $f_\varphi\colon E \to E$, such that*

$$\varphi(u, v) = \langle u, f_\varphi(v) \rangle \quad \textit{for all } u, v \in E.$$

We also have $\|f_\varphi\| = \|\varphi\|$. If φ is Hermitian, then f_φ is self-adjoint, that is

$$\langle u, f_\varphi(v) \rangle = \langle f_\varphi(u), v \rangle \quad \textit{for all } u, v \in E.$$

Proof. The proof is adapted from Rudin [141] (Theorem 12.8). To define the function f_φ, we proceed as follows. For any fixed $v \in E$, define the linear map φ_v by

$$\varphi_v(u) = \varphi(u, v) \quad \text{for all } u \in E.$$

Since φ is continuous, φ_v is continuous. So by Proposition 48.9, there is a unique vector in E that we denote $f_\varphi(v)$ such that

$$\varphi_v(u) = \langle u, f_\varphi(v) \rangle \quad \text{for all } u \in E,$$

and $\|f_\varphi(v)\| = \|\varphi_v\|$. Let us check that the map $v \mapsto f_\varphi(v)$ is linear.

We have

$$\begin{aligned}
\varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2) && \varphi \text{ is additive} \\
&= \langle u, f_\varphi(v_1) \rangle + \langle u, f_\varphi(v_2) \rangle && \text{by definition of } f_\varphi \\
&= \langle u, f_\varphi(v_1) + f_\varphi(v_2) \rangle && \langle -, - \rangle \text{ is additive}
\end{aligned}$$

for all $u \in E$, and since $f_\varphi(v_1+v_2)$ is the unique vector such that $\varphi(u, v_1+v_2) = \langle u, f_\varphi(v_1+v_2)\rangle$ for all $u \in E$, we must have

$$f_\varphi(v_1 + v_2) = f_\varphi(v_1) + f_\varphi(v_2).$$

For any $\lambda \in \mathbb{C}$ we have

$$\begin{aligned}
\varphi(u, \lambda v) &= \overline{\lambda}\varphi(u, v) && \varphi \text{ is sesquilinear} \\
&= \overline{\lambda}\langle u, f_\varphi(v)\rangle && \text{by definition of } f_\varphi \\
&= \langle u, \lambda f_\varphi(v)\rangle && \langle -, -\rangle \text{ is sesquilinear}
\end{aligned}$$

for all $u \in E$, and since $f_\varphi(\lambda v)$ is the unique vector such that $\varphi(u, \lambda v) = \langle u, f_\varphi(\lambda v)\rangle$ for all $u \in E$, we must have

$$f_\varphi(\lambda v) = \lambda f_\varphi(v).$$

Therefore f_φ is linear.

Then by definition of $\|\varphi\|$, we have

$$|\varphi_v(u)| = |\varphi(u, v)| \leq \|\varphi\| \|u\| \|v\|,$$

which shows that $\|\varphi_v\| \leq \|\varphi\| \|v\|$. Since $\|f_\varphi(v)\| = \|\varphi_v\|$, we have

$$\|f_\varphi(v)\| \leq \|\varphi\| \|v\|,$$

which shows that f_φ is continuous and that $\|f_\varphi\| \leq \|\varphi\|$. But by the Cauchy–Schwarz inequality we also have

$$|\varphi(u, v)| = |\langle u, f_\varphi(v)\rangle| \leq \|u\| \|f_\varphi(v)\| \leq \|u\| \|f_\varphi\| \|v\|,$$

so $\|\varphi\| \leq \|f_\varphi\|$, and thus

$$\|f_\varphi\| = \|\varphi\|.$$

If φ is Hermitian, $\varphi(v, u) = \overline{\varphi(u, v)}$, so

$$\langle f_\varphi(u), v\rangle = \overline{\langle v, f_\varphi(u)\rangle} = \overline{\varphi(v, u)} = \varphi(u, v) = \langle u, f_\varphi(v)\rangle,$$

which shows that f_φ is self-adjoint. ☐

Proposition 48.11. *Given a Hilbert space E, for every continuous linear map $f\colon E \to E$, there is a unique continuous linear map $f^*\colon E \to E$, such that*

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle \quad \text{for all } u, v \in E,$$

and we have $\|f^*\| = \|f\|$. The map f^* is called the adjoint of f.

Proof. The proof is adapted from Rudin [141] (Section 12.9). By the Cauchy–Schwarz inequality, since

$$|\langle x, y\rangle| \leq \|x\| \, \|y\| \,,$$

we see that the sesquilinear map $(x, y) \mapsto \langle x, y\rangle$ on $E \times E$ is continuous. Let $\varphi\colon E \times E \to \mathbb{C}$ be the sesquilinear map given by

$$\varphi(u, v) = \langle f(u), v\rangle \quad \text{for all } u, v \in E.$$

Since f is continuous and the inner product $\langle -, -\rangle$ is continuous, this is a continuous map. By Proposition 48.10, there is a unique linear map $f^*\colon E \to E$ such that

$$\langle f(u), v\rangle = \varphi(u, v) = \langle u, f^*(v)\rangle \quad \text{for all } u, v \in E,$$

with $\|f^*\| = \|\varphi\|$.

We can also prove that $\|\varphi\| = \|f\|$. First, by definition of $\|\varphi\|$ we have

$$\begin{aligned}
\|\varphi\| &= \sup \{|\varphi(x, y)| \mid \|x\| \leq 1, \ \|y\| \leq 1\} \\
&= \sup \{|\langle f(x), y\rangle| \mid \|x\| \leq 1, \ \|y\| \leq 1\} \\
&\leq \sup \{\|f(x)\| \, \|y\| \mid \|x\| \leq 1, \|y\| \leq 1\} \\
&\leq \sup \{\|f(x)\| \mid \|x\| \leq 1\} \\
&= \|f\| \,.
\end{aligned}$$

In the other direction we have

$$\|f(x)\|^2 = \langle f(x), f(x)\rangle = \varphi(x, f(x)) \leq \|\varphi\| \, \|x\| \, \|f(x)\| \,,$$

and if $f(x) \neq 0$ we get $\|f(x)\| \leq \|\varphi\| \, \|x\|$. This inequality holds trivially if $f(x) = 0$, so we conclude that $\|f\| \leq \|\varphi\|$. Therefore we have

$$\|\varphi\| = \|f\| \,,$$

as claimed, and consequently $\|f^*\| = \|\varphi\| = \|f\|$. $\square$

It is easy to show that the adjoint satisfies the following properties:

$$\begin{aligned}
(f + g)^* &= f^* + g^* \\
(\lambda f)^* &= \overline{\lambda} f^* \\
(f \circ g)^* &= g^* \circ f^* \\
f^{**} &= f.
\end{aligned}$$

One can also show that $\|f^* \circ f\| = \|f\|^2$ (see Rudin [141], Section 12.9).

As in the Hermitian case, given two Hilbert spaces E and F, the above results can be adapted to show that for any linear map $f\colon E \to F$, there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v\rangle_2 = \langle u, f^*(v)\rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the adjoint of f.

48.3 Farkas–Minkowski Lemma in Hilbert Spaces

In this section $(V, \langle -, - \rangle)$ is assumed to be a *real* Hilbert space. The projection lemma can be used to show an interesting version of the Farkas–Minkowski lemma in a Hilbert space.

Given a finite sequence of vectors $(a_1, \ldots, a_m)$ with $a_i \in V$, let C be the polyhedral cone

$$C = \mathrm{cone}(a_1, \ldots, a_m) = \left\{ \sum_{i=1}^{m} \lambda_i a_i \mid \lambda_i \geq 0,\ i = 1, \ldots, m \right\}.$$

For any vector $b \in V$, the Farkas–Minkowski lemma gives a criterion for checking whether $b \in C$.

In Proposition 44.2 we proved that every polyhedral cone $\mathrm{cone}(a_1, \ldots, a_m)$ with $a_i \in \mathbb{R}^n$ is closed. Close examination of the proof shows that it goes through if $a_i \in V$ where V is any vector space possibly of infinite dimension, because the important fact is that the number m of these vectors is finite, not their dimension.

Theorem 48.12. *(Farkas–Minkowski Lemma in Hilbert Spaces) Let $(V, \langle -, - \rangle)$ be a real Hilbert space. For any finite sequence of vectors $(a_1, \ldots, a_m)$ with $a_i \in V$, if C is the polyhedral cone $C = \mathrm{cone}(a_1, \ldots, a_m)$, for any vector $b \in V$, we have $b \notin C$ iff there is a vector $u \in V$ such that*

$$\langle a_i, u \rangle \geq 0 \quad i = 1, \ldots, m, \quad \text{and} \quad \langle b, u \rangle < 0.$$

Equivalently, $b \in C$ iff for all $u \in V$,

$$\text{if } \langle a_i, u \rangle \geq 0 \quad i = 1, \ldots, m, \quad \text{then} \quad \langle b, u \rangle \geq 0.$$

Proof. We follow Ciarlet [41] (Chapter 9, Theorem 9.1.1). We already established in Proposition 44.2 that the polyhedral cone $C = \mathrm{cone}(a_1, \ldots, a_m)$ is closed. Next we claim the following:

Claim: If C is a nonempty, closed, convex subset of a Hilbert space V, and $b \in V$ is any vector such that $b \notin C$, then there exist some $u \in V$ and infinitely many scalars $\alpha \in \mathbb{R}$ such that

$$\begin{aligned} \langle v, u \rangle &> \alpha \quad \text{for every } v \in C \\ \langle b, u \rangle &< \alpha. \end{aligned}$$

We use the projection lemma (Proposition 48.5) which says that since $b \notin C$ there is some unique $c = p_C(b) \in C$ such that

$$\begin{aligned} \|b - c\| &= \inf_{v \in C} \|b - v\| > 0 \\ \langle b - c, v - c \rangle &\leq 0 \quad \text{for all } v \in C, \end{aligned}$$

or equivalently

$$\|b - c\| = \inf_{v \in C} \|b - v\| > 0$$
$$\langle v - c, c - b\rangle \geq 0 \quad \text{for all } v \in C.$$

As a consequence, since $b \notin C$ and $c \in C$, we have $c - b \neq 0$, so

$$\langle v, c - b\rangle \geq \langle c, c - b\rangle > \langle b, c - b\rangle$$

because $\langle c, c - b\rangle - \langle b, c - b\rangle = \langle c - b, c - b\rangle > 0$, and if we pick $u = c - b$ and any α such that

$$\langle c, c - b\rangle > \alpha > \langle b, c - b\rangle,$$

the claim is satisfied.

We now prove the Farkas–Minkowski lemma. Assume that $b \notin C$. Since C is nonempty, convex, and closed, by the claim there is some $u \in V$ and some $\alpha \in \mathbb{R}$ such that

$$\langle v, u\rangle > \alpha \quad \text{for every } v \in C$$
$$\langle b, u\rangle < \alpha.$$

But C is a polyhedral cone containing 0, so we must have $\alpha < 0$. Then for every $v \in C$, since C a polyhedral cone if $v \in C$ then $\lambda v \in C$ for all $\lambda > 0$, so by the above

$$\langle v, u\rangle > \frac{\alpha}{\lambda} \quad \text{for every } \lambda > 0,$$

which implies that

$$\langle v, u\rangle \geq 0.$$

Since $a_i \in C$ for $i = 1, \ldots, m$, we proved that

$$\langle a_i, u\rangle \geq 0 \quad i = 1, \ldots, m \quad \text{and} \quad \langle b, u\rangle < \alpha < 0,$$

which proves Farkas lemma. $\square$

Remark: Observe that the claim established during the proof of Theorem 48.12 shows that the affine hyperplane $H_{u,\alpha}$ of equation $\langle v, u\rangle = \alpha$ for all $v \in V$ separates strictly C and $\{b\}$.

48.4 Summary

The main concepts and results of this chapter are listed below:

- Hilbert space.
- Projection lemma.

- Distance of a point to a subset, diameter.
- Projection onto a closed and convex subset.
- Orthogonal complement of a closed subspace.
- Dual of a Hilbert space.
- Bounded linear operator (or functional).
- Riesz representation theorem.
- Adjoint of a continuous linear map.
- Farkas–Minkowski lemma.

48.5 Problems

Problem 48.1. Let V be a Hilbert space. Prove that a subspace W of V is dense in V if and only if there is no nonzero vector orthogonal to W.

Problem 48.2. Prove that the adjoint satisfies the following properties:

$$\begin{aligned}(f+g)^* &= f^* + g^* \\ (\lambda f)^* &= \overline{\lambda} f^* \\ (f \circ g)^* &= g^* \circ f^* \\ f^{**} &= f.\end{aligned}$$

Problem 48.3. Prove that $\|f^* \circ f\| = \|f\|^2$.

Problem 48.4. Let V be a (real) Hilbert space and let C be a nonempty closed convex subset of V. Define the map $h\colon V \to \mathbb{R} \cup \{+\infty\}$ by

$$h(u) = \sup_{v \in C} \langle u, v \rangle.$$

Prove that

$$C = \bigcap_{u \in V} \{v \in V \mid \langle u, v\rangle \leq h(u)\} = \bigcap_{u \in \Lambda_C} \{v \in V \mid \langle u, v\rangle \leq h(u)\},$$

where $\Lambda_C = \{u \in V \mid h(u) \neq +\infty\}$.

Describe Λ_C when C is also a subspace of V.

Problem 48.5. Let A be a real $m \times n$ matrix, and let (u_k) be a sequence of vectors $u_k \in \mathbb{R}^n$ such that $u_k \geq 0$. Prove that if the sequence (Au_k) converges, then there is some $u \in \mathbb{R}^n$ such that

$$Au = \lim_{k \mapsto \infty} Au_k \quad \text{and} \quad u \geq 0.$$

Problem 48.6. Let V be a real Hilbert space, $(a_1, \ldots, a_m)$ a sequence of m vectors in V, b some vector in V, $(\alpha_1, \ldots, \alpha_m)$ a sequence of m real numbers, and β some real number. Prove that the inclusion

$$\{w \in V \mid \langle a_i, w\rangle \geq \alpha_i,\ 1 \leq i \leq m\} \subseteq \{w \in V \mid \langle b, w\rangle \geq \beta\}$$

holds if and only if there exist $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ such that $\lambda_i \geq 0$ for $i = 1, \ldots, m$ and

$$\begin{aligned} b &= \sum_{i=1}^{m} \lambda_i a_i \\ \beta &\leq \sum_{i=1}^{m} \lambda_i \alpha_i. \end{aligned}$$

Chapter 49

General Results of Optimization Theory

This chapter is devoted to some general results of optimization theory. A main theme is to find sufficient conditions that ensure that an objective function has a minimum which is achieved. We define the notion of a coercive function. The most general result is Theorem 49.2, which applies to a coercive convex function on a convex subset of a separable Hilbert space. In the special case of a coercive quadratic functional, we obtain the Lions–Stampacchia theorem (Theorem 49.6), and the Lax–Milgram theorem (Theorem 49.7). We define elliptic functionals, which generalize quadratic functions defined by symmetric positive definite matrices. We define gradient descent methods, and discuss their convergence. A gradient descent method looks for a descent direction and a stepsize parameter, which is obtained either using an exact line search or a backtracking line search. A popular technique to find the search direction is steepest descent. In addition to steepest descent for the Euclidean norm, we discuss steepest descent for an arbitrary norm. We also consider a special case of steepest descent, Newton's method. This method converges faster than the other gradient descent methods, but it is quite expensive since it requires computing and storing Hessians. We also present the method of conjugate gradients and prove its correctness. We briefly discuss the method of gradient projection and the penalty method in the case of constrained optima.

49.1 Optimization Problems; Basic Terminology

The main goal of *optimization theory* is to construct *algorithms* to find solutions (often approximate) of problems of the form

$$\begin{aligned}&\text{find}\quad u\\ &\text{such that}\ \ u \in U\ \text{ and }\ J(u) = \inf_{v\in U} J(v),\end{aligned}$$

where U is a given subset of a (real) vector space V (possibly infinite dimensional) and $J\colon \Omega \to \mathbb{R}$ is a function defined on some open subset Ω of V such that $U \subseteq \Omega$.

To be very clear, $\inf_{v\in U} J(v)$ denotes the *greatest lower bound* of the set of real numbers $\{J(u) \mid u \in U\}$. To make sure that we are on firm grounds, let us review the notions of greatest lower bound and least upper bound of a set of real numbers.

Let X be any nonempty subset of $\mathbb{R}$. The set $LB(X)$ of *lower bounds* of X is defined as

$$LB(X) = \{b \in \mathbb{R} \mid b \leq x \text{ for all } x \in X\}.$$

If the set X is not bounded below, which means that for every $r \in \mathbb{R}$ there is some $x \in X$ such that $x < r$, then $LB(X)$ is empty. Otherwise, if $LB(X)$ is nonempty, since it is bounded above by every element of X, by a fundamental property of the real numbers, the set $LB(X)$ has a greatest element denoted $\inf X$. The real number $\inf X$ is thus the *greatest lower bound* of X. In general, $\inf X$ does not belong to X, but if it does, then it is the least element of X.

If $LB(X) = \emptyset$, then X is *unbounded below* and $\inf X$ is undefined. In this case (with an abuse of notation), we write

$$\inf X = -\infty.$$

By convention, when $X = \emptyset$ we set

$$\inf \emptyset = +\infty.$$

For example, if $X = \{x \in \mathbb{R} \mid x \leq 0\}$, then $LB(X) = \emptyset$. On the other hand, if $X = \{1/n \mid n \in \mathbb{N} - \{0\}\}$, then $LB(X) = \{x \in \mathbb{R} \mid x \leq 0\}$ and $\inf X = 0$, which is not in X.

Similarly, the set $UB(X)$ of *upper bounds* of X is given by

$$UB(X) = \{u \in \mathbb{R} \mid x \leq u \text{ for all } x \in X\}.$$

If X is not bounded above, then $UB(X) = \emptyset$. Otherwise, if $UB(X) \neq \emptyset$, then it has least element denoted $\sup X$. Thus $\sup X$ is the *least upper bound* of X. If $\sup X \in X$, then it is the greatest element of X. If $UB(X) = \emptyset$, then

$$\sup X = +\infty.$$

By convention, when $X = \emptyset$ we set

$$\sup \emptyset = -\infty.$$

For example, if $X = \{x \in \mathbb{R} \mid x \geq 0\}$, then $UB(X) = \emptyset$. On the other hand, if $X = \{1 - 1/n \mid n \in \mathbb{N} - \{0\}\}$, then $UB(X) = \{x \in \mathbb{R} \mid x \geq 1\}$ and $\sup X = 1$, which is not in X.

The element $\inf_{v\in U} J(v)$ is just $\inf\{J(v) \mid v \in U\}$. The notation J^* is often used to denote $\inf_{v\in U} J(v)$. If the function J is not bounded below, which means that for every $r \in \mathbb{R}$, there is some $u \in U$ such that $J(u) < r$, then

$$\inf_{v\in U} J(v) = -\infty,$$

and we say that our minimization problem has no solution, or that it is unbounded (below). For example, if $V = \Omega = \mathbb{R}$, $U = \{x \in \mathbb{R} \mid x \leq 0\}$, and $J(x) = x$, then the function $J(x)$ is not bounded below and $\inf_{v \in U} J(v) = -\infty$.

The issue is that J^* may not belong to $\{J(u) \mid u \in U\}$, that is, it may not be achieved by some element $u \in U$, and solving the above problem consists in finding some $u \in U$ that achieves the value J^* in the sense that $J(u) = J^*$. If no such $u \in U$ exists, again we say that our minimization problem has no solution.

The minimization problem

$$\begin{aligned} &\text{find} \quad u \\ &\text{such that} \quad u \in U \text{ and } J(u) = \inf_{v \in U} J(v) \end{aligned}$$

is often presented in the following more informal way:

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad v \in U. \end{aligned} \qquad \textbf{(Problem M)}$$

A vector $u \in U$ such that $J(u) = \inf_{v \in U} J(v)$ is often called a *minimizer* of J over U. Some authors denote the set of minimizers of J over U by $\arg\min_{v \in U} J(v)$ and write

$$u \in \arg\min_{v \in U} J(v)$$

to express that u is such a minimizer. When such a minimizer is unique, by abuse of notation, this unique minimizer u is denoted by

$$u = \arg\min_{v \in U} J(v).$$

We prefer not to use this notation, although it seems to have invaded the literature.

If we need to maximize rather than minimize a function, then we try to find some $u \in U$ such that

$$J(u) = \sup_{v \in U} J(v).$$

Here $\sup_{v \in U} J(v)$ is the least upper bound of the set $\{J(u) \mid u \in U\}$. Some authors denote the set of *maximizers* of J over U by $\arg\max_{v \in U} J(v)$.

Remark: Some authors define an *extended real-valued function* as a function $f \colon \Omega \to \mathbb{R}$ which is allowed to take the value $-\infty$ or even $+\infty$ for some of its arguments. Although this may be convenient to deal with situations where we need to consider $\inf_{v \in U} J(v)$ or $\sup_{v \in U} J(v)$, such "functions" are really partial functions and we prefer not to use the notion of extended real-valued function.

In most cases, U is defined as the set of solutions of a finite sets of *constraints*, either equality constraints $\varphi_i(v) = 0$, or inequality constraints $\varphi_i(v) \leq 0$, where the $\varphi_i\colon \Omega \to \mathbb{R}$ are some given functions. The function J is often called the *functional* of the optimization problem. This is a slightly odd terminology, but it is justified if V is a function space.

The following questions arise naturally:

(1) Results concerning the *existence and uniqueness* of a solution for Problem (M). In the next section we state sufficient conditions either on the domain U or on the function J that ensure the existence of a solution.

(2) The *characterization* of the possible solutions of Problem M. These are conditions for any element $u \in U$ to be a solution of the problem. Such conditions usually involve the derivative dJ_u of J, and possibly the derivatives of the functions φ_i defining U. Some of these conditions become sufficient when the functions φ_i are convex,

(3) The effective construction of *algorithms*, typically iterative algorithms that construct a sequence $(u_k)_{k\geq 1}$ of elements of U whose limit is a solution $u \in U$ of our problem. It is then necessary to understand when and how quickly such sequences converge. Gradient descent methods fall under this category. As a general rule, unconstrained problems (for which $U = \Omega = V$) are (much) easier to deal with than constrained problems (where $U \neq V$).

The material of this chapter is heavily inspired by Ciarlet [41]. *In this chapter it is assumed that V is a real vector space with an inner product $\langle -, - \rangle$.* If V is infinite dimensional, then we assume that it is a real Hilbert space (it is complete). As usual, we write $\|u\| = \langle u, u \rangle^{1/2}$ for the norm associated with the inner product $\langle -, - \rangle$. The reader may want to review Section 48.1, especially the projection lemma and the Riesz representation theorem.

As a matter of terminology, if U is defined by inequality and equality constraints as

$$U = \{v \in \Omega \mid \varphi_i(v) \leq 0,\ i = 1, \ldots, m,\ \psi_j(v) = 0,\ j = 1, \ldots, p\},$$

if J and all the functions φ_i and ψ_j are affine, the problem is said to be *linear* (or a *linear program*), and otherwise *nonlinear*. If J is of the form

$$J(v) = \langle Av, v \rangle - \langle b, v \rangle$$

where A is a nonzero symmetric positive semidefinite matrix and the constraints are affine, the problem is called a *quadratic programming problem*. If the inner product $\langle -, - \rangle$ is the standard Euclidean inner product, J is also expressed as

$$J(v) = v^\top Av - b^\top v.$$

49.2 Existence of Solutions of an Optimization Problem

We begin with the case where U is a closed but possibly unbounded subset of $\mathbb{R}^n$. In this case the following type of functions arise.

Definition 49.1. A real-valued function $J: V \to \mathbb{R}$ defined on a normed vector space V is *coercive* iff for any sequence $(v_k)_{k\geq 1}$ of vectors $v_k \in V$, if $\lim_{k\mapsto\infty} \|v_k\| = \infty$, then

$$\lim_{k\mapsto\infty} J(v_k) = +\infty.$$

For example, the function $f(x) = x^2 + 2x$ is coercive, but an affine function $f(x) = ax + b$ is not.

Proposition 49.1. *Let U be a nonempty, closed subset of $\mathbb{R}^n$, and let $J: \mathbb{R}^n \to \mathbb{R}$ be a continuous function which is coercive if U is unbounded. Then there is a least one element $u \in \mathbb{R}^n$ such that*

$$u \in U \quad \textit{and} \quad J(u) = \inf_{v\in U} J(v).$$

Proof. Since $U \neq \emptyset$, pick any $u_0 \in U$. Since J is coercive, there is some $r > 0$ such that for all $v \in \mathbb{R}^n$, if $\|v\| > r$ then $J(u_0) < J(v)$. It follows that J is minimized over the set

$$U_0 = U \cap \{v \in \mathbb{R}^n \mid \|v\| \leq r\}.$$

Since U is closed and since the closed ball $\{v \in \mathbb{R}^n \mid \|v\| \leq r\}$ is compact, U_0 is compact, but we know that any continuous function on a compact set has a minimum which is achieved. ☐

The key point in the above proof is the fact that U_0 is compact. In order to generalize Proposition 49.1 to the case of an infinite dimensional vector space, we need some additional assumptions, and it turns out that the convexity of U and of the function J is sufficient. The key is that convex, closed and bounded subsets of a Hilbert space are "weakly compact."

Definition 49.2. Let V be a Hilbert space. A sequence $(u_k)_{k\geq 1}$ of vectors $u_k \in V$ *converges weakly* if there is some $u \in V$ such that

$$\lim_{k\mapsto\infty} \langle v, u_k\rangle = \langle v, u\rangle \quad \text{for every } v \in V.$$

Recall that a Hibert space is separable if it has a countable Hilbert basis (see Definition A.4). Also, in a Euclidean space (of finite dimension) V, the inner product induces an isomorphism between V and its dual V^*. In our case, we need the isomorphism $\sharp$ from V^* to V defined such that for every linear form $\omega \in V^*$, the vector $\omega^\sharp \in V$ is uniquely defined by the equation

$$\omega(v) = \langle v, \omega^\sharp\rangle \quad \text{for all } v \in V.$$

In a Hilbert space, the dual space V' is the set of all continuous linear forms $\omega\colon V \to \mathbb{R}$, and the existence of the isomorphism $\sharp$ between V' and V is given by the Riesz representation theorem; see Proposition 48.9. This theorem allows a generalization of the notion of gradient. Indeed, if $f\colon V \to \mathbb{R}$ is a function defined on the Hilbert space V and if f is differentiable at some point $u \in V$, then by definition, the derivative $df_u\colon V \to \mathbb{R}$ is a continuous linear form, so by the Riesz representation theorem (Proposition 48.9) there is a unique vector, denoted $\nabla f_u \in V$, such that

$$df_u(v) = \langle v, \nabla f_u\rangle \quad \text{for all } v \in V.$$

Definition 49.3. The unique vector ∇f_u such that

$$df_u(v) = \langle v, \nabla f_u\rangle \quad \text{for all } v \in V$$

is called the *gradient* of f at u.

Similarly, since the second derivative $\mathrm{D}^2 f_u\colon V \to V'$ of f induces a continuous symmetric billinear form from $V \times V$ to $\mathbb{R}$, by Proposition 48.10, there is a unique continuous self-adjoint linear map $\nabla^2 f_u\colon V \to V$ such that

$$\mathrm{D}^2 f_u(v, w) = \langle \nabla^2 f_u(v), w\rangle \quad \text{for all } v, w \in V.$$

The map $\nabla^2 f_u$ is a generalization of the *Hessian*.

The next theorem is a rather general result about the existence of minima of convex functions defined on convex domains. The proof is quite involved and can be omitted upon first reading.

Theorem 49.2. *Let U be a nonempty, convex, closed subset of a separable Hilbert space V, and let $J\colon V \to \mathbb{R}$ be a convex, differentiable function which is coercive if U is unbounded. Then there is a least one element $u \in V$ such that*

$$u \in U \quad \textit{and} \quad J(u) = \inf_{v \in U} J(v).$$

Proof. As in the proof of Proposition 49.1, since the function J is coercive, we may assume that U is bounded and convex (however, if V infinite dimensional, then U is not compact in general). The proof proceeds in four steps.

Step 1. Consider a *minimizing sequence* $(u_k)_{k\geq 0}$, namely a sequence of elements $u_k \in V$ such that

$$u_k \in U \quad \text{for all } k \geq 0, \quad \lim_{k \mapsto \infty} J(u_k) = \inf_{v \in U} J(v).$$

At this stage, it is possible that $\inf_{v\in U} J(v) = -\infty$, but we will see that this is actually impossible. However, since U is bounded, the sequence $(u_k)_{k\geq 0}$ is bounded. Our goal is to prove that there is some subsequence of $(w_\ell)_{\ell\geq 0}$ of $(u_k)_{k\geq 0}$ that converges weakly.

Since the sequence $(u_k)_{k\geq 0}$ is bounded there is some constant $C > 0$ such that $\|u_k\| \leq C$ for all $k \geq 0$. Then by the Cauchy–Schwarz inequality, for every $v \in V$ we have

$$|\langle v, u_k\rangle| \leq \|v\| \, \|u_k\| \leq C \, \|v\| \, ,$$

which shows that the sequence $(\langle v, u_k\rangle)_{k\geq 0}$ is bounded. Since V is a separable Hilbert space, there is a countable family $(v_k)_{k\geq 0}$ of vectors $v_k \in V$ which is dense in V. Since the sequence $(\langle v_1, u_k\rangle)_{k\geq 0}$ is bounded (in $\mathbb{R}$), we can find a convergent subsequence $(\langle v_1, u_{i_1(j)}\rangle)_{j\geq 0}$. Similarly, since the sequence $(\langle v_2, u_{i_1(j)}\rangle)_{j\geq 0}$ is bounded, we can find a convergent subsequence $(\langle v_2, u_{i_2(j)}\rangle)_{j\geq 0}$, and in general, since the sequence $(\langle v_k, u_{i_{k-1}(j)}\rangle)_{j\geq 0}$ is bounded, we can find a convergent subsequence $(\langle v_k, u_{i_k(j)}\rangle)_{j\geq 0}$.

We obtain the following infinite array:

$$\begin{pmatrix}
\langle v_1, u_{i_1(1)}\rangle & \langle v_2, u_{i_2(1)}\rangle & \cdots & \langle v_k, u_{i_k(1)}\rangle & \cdots \\
\langle v_1, u_{i_1(2)}\rangle & \langle v_2, u_{i_2(2)}\rangle & \cdots & \langle v_k, u_{i_k(2)}\rangle & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\langle v_1, u_{i_1(k)}\rangle & \langle v_2, u_{i_2(k)}\rangle & \cdots & \langle v_k, u_{i_k(k)}\rangle & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}$$

Consider the "diagonal" sequence $(w_\ell)_{\ell\geq 0}$ defined by

$$w_\ell = u_{i_\ell(\ell)}, \quad \ell \geq 0.$$

We are going to prove that for every $v \in V$, the sequence $(\langle v, w_\ell\rangle)_{\ell\geq 0}$ has a limit.

By construction, for every $k \geq 0$, the sequence $(\langle v_k, w_\ell\rangle)_{\ell\geq 0}$ has a limit, which is the limit of the sequence $(\langle v_k, u_{i_k(j)}\rangle)_{j\geq 0}$, since the sequence $(i_\ell(\ell))_{\ell\geq 0}$ is a subsequence of every sequence $(i_\ell(j))_{j\geq 0}$ for every $\ell \geq 0$.

Pick any $v \in V$ and any $\epsilon > 0$. Since $(v_k)_{k\geq 0}$ is dense in V, there is some v_k such that

$$\|v - v_k\| \leq \epsilon/(4C).$$

Then we have

$$\begin{aligned}
|\langle v, w_\ell\rangle - \langle v, w_m\rangle| &= |\langle v, w_\ell - w_m\rangle| \\
&= |\langle v_k + v - v_k, w_\ell - w_m\rangle| \\
&= |\langle v_k, w_\ell - w_m\rangle + \langle v - v_k, w_\ell - w_m\rangle| \\
&\leq |\langle v_k, w_\ell\rangle - \langle v_k, w_m\rangle| + |\langle v - v_k, w_\ell - w_m\rangle|.
\end{aligned}$$

By Cauchy–Schwarz and since $\|w_\ell - w_m\| \leq \|w_\ell\| + \|w_m\| \leq C + C = 2C$,

$$|\langle v - v_k, w_\ell - w_m\rangle| \leq \|v - v_k\| \, \|w_\ell - w_m\| \leq (\epsilon/(4C))2C = \epsilon/2,$$

so

$$|\langle v, w_\ell\rangle - \langle v, w_m\rangle| \leq |\langle v_k, w_\ell - w_m\rangle| + \epsilon/2.$$

With the element v_k held fixed, by a previous argument the sequence $(\langle v_k, w_\ell\rangle)_{\ell\geq 0}$ converges, so it is a Cauchy sequence. Consequently there is some ℓ_0 (depending on ϵ and v_k) such that

$$|\langle v_k, w_\ell\rangle - \langle v_k, w_m\rangle| \leq \epsilon/2 \quad \text{for all } \ell, m \geq \ell_0,$$

so we get

$$|\langle v, w_\ell\rangle - \langle v, w_m\rangle| \leq \epsilon/2 + \epsilon/2 = \epsilon \quad \text{for all } \ell, m \geq \ell_0.$$

This proves that the sequence $(\langle v, w_\ell\rangle)_{\ell\geq 0}$ is a Cauchy sequence, and thus it converges.

Define the function $g\colon V \to \mathbb{R}$ by

$$g(v) = \lim_{\ell\mapsto\infty} \langle v, w_\ell\rangle, \quad \text{for all } v \in V.$$

Since

$$|\langle v, w_\ell\rangle| \leq \|v\| \, \|w_\ell\| \leq C \, \|v\| \quad \text{for all } \ell \geq 0,$$

we have

$$|g(v)| \leq C \, \|v\| \, ,$$

so g is a continuous linear map. By the Riesz representation theorem (Proposition 48.9), there is a unique $u \in V$ such that

$$g(v) = \langle v, u\rangle \quad \text{for all } v \in V,$$

which shows that

$$\lim_{\ell\mapsto\infty} \langle v, w_\ell\rangle = \langle v, u\rangle \quad \text{for all } v \in V,$$

namely the subsequence $(w_\ell)_{\ell\geq 0}$ of the sequence $(u_k)_{k\geq 0}$ converges weakly to $u \in V$.

Step 2. We prove that the "weak limit" u of the sequence $(w_\ell)_{\ell\geq 0}$ belongs to U.

Consider the projection $p_U(u)$ of $u \in V$ onto the closed convex set U. Since $w_\ell \in U$, by Proposition 48.5(2) and the fact that U is convex and closed, we have

$$\langle p_U(u) - u, w_\ell - p_U(u)\rangle \geq 0 \quad \text{for all } \ell \geq 0.$$

The weak convergence of the sequence $(w_\ell)_{\ell\geq 0}$ to u implies that

$$\begin{aligned} 0 \leq \lim_{\ell\mapsto\infty} \langle p_U(u) - u, w_\ell - p_U(u)\rangle &= \langle p_U(u) - u, u - p_U(u)\rangle \\ &= -\|p_U(u) - u\| \leq 0, \end{aligned}$$

so $\|p_U(u) - u\| = 0$, which means that $p_U(u) = u$, and so $u \in U$.

Step 3. We prove that

$$J(v) \leq \liminf_{\ell\mapsto\infty} J(z_\ell)$$

for every sequence $(z_\ell)_{\ell\geq 0}$ converging weakly to some element $v \in V$.

Since J is assumed to be differentiable and convex, by Proposition 40.11(1) we have

$$J(v) + \langle \nabla J_v, z_\ell - v\rangle \leq J(z_\ell) \quad \text{for all } \ell \geq 0,$$

and by definition of weak convergence

$$\lim_{\ell \mapsto \infty} \langle \nabla J_v, z_\ell \rangle = \langle \nabla J_v, v \rangle,$$

so $\lim_{\ell \mapsto \infty} \langle \nabla J_v, z_\ell - v \rangle = 0$, and by definition of lim inf we get

$$J(v) \leq \liminf_{\ell \mapsto \infty} J(z_\ell)$$

for every sequence $(z_\ell)_{\ell \geq 0}$ converging weakly to some element $v \in V$.

Step 4. The weak limit $u \in U$ of the subsequence $(w_\ell)_{\ell \geq 0}$ extracted from the minimizing sequence $(u_k)_{k \geq 0}$ satisfies the equation

$$J(u) = \inf_{v \in U} J(v).$$

By Step (1) and Step (2) the subsequence $(w_\ell)_{\ell \geq 0}$ of the sequence $(u_k)_{k \geq 0}$ converges weakly to some element $u \in U$, so by Step (3) we have

$$J(u) \leq \liminf_{\ell \mapsto \infty} J(w_\ell).$$

On the other hand, by definition of $(w_\ell)_{\ell \geq 0}$ as a subsequence of $(u_k)_{k \geq 0}$, since the sequence $(J(u_k))_{k \geq 0}$ converges to $J(v)$, we have

$$J(u) \leq \liminf_{\ell \mapsto \infty} J(w_\ell) = \lim_{k \mapsto \infty} J(u_k) = \inf_{v \in U} J(v),$$

which proves that $u \in U$ achieves the minimum of J on U. $\square$

Remark: Theorem 49.2 still holds if we only assume that J is convex and continuous. It also holds in a reflexive Banach space, of which Hilbert spaces are a special case; see Brezis [31], Corollary 3.23.

Theorem 49.2 is a rather general theorem whose proof is quite involved. For functions J of a certain type, we can obtain existence and uniqueness results that are easier to prove. This is true in particular for quadratic functionals.

49.3 Minima of Quadratic Functionals

Definition 49.4. Let V be a real Hilbert space. A function $J: V \to \mathbb{R}$ is called a *quadratic functional* if it is of the form

$$J(v) = \frac{1}{2} a(v, v) - h(v),$$

where $a: V \times V \to \mathbb{R}$ is a bilinear form which is symmetric and continuous, and $h: V \to \mathbb{R}$ is a continuous linear form.

Definition 49.4 is a natural extension of the notion of a quadratic functional on $\mathbb{R}^n$. Indeed, by Proposition 48.10, there is a unique continuous self-adjoint linear map $A\colon V \to V$ such that

$$a(u, v) = \langle Au, v\rangle \quad \text{for all } u, v \in V,$$

and by the Riesz representation theorem (Proposition 48.9), there is a unique $b \in V$ such that

$$h(v) = \langle b, v\rangle \quad \text{for all } v \in V.$$

Consequently, J can be written as

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \quad \text{for all } v \in V. \tag{1}$$

Since a is bilinear and h is linear, by Propositions 39.3 and 39.5, observe that the derivative of J is given by

$$dJ_u(v) = a(u, v) - h(v) \quad \text{for all } v \in V,$$

or equivalently by

$$dJ_u(v) = \langle Au, v\rangle - \langle b, v\rangle = \langle Au - b, v\rangle, \quad \text{for all } v \in V.$$

Thus the gradient of J is given by

$$\nabla J_u = Au - b, \tag{2}$$

just as in the case of a quadratic function of the form $J(v) = (1/2)v^\top Av - b^\top v$, where A is a symmetric $n \times n$ matrix and $b \in \mathbb{R}^n$. To find the second derivative $\mathrm{D}^2 J_u$ of J at u we compute

$$dJ_{u+v}(w) - dJ_u(w) = a(u + v, w) - h(w) - (a(u, w) - h(w)) = a(v, w),$$

so

$$\mathrm{D}^2 J_u(v, w) = a(v, w) = \langle Av, w\rangle,$$

which yields

$$\nabla^2 J_u = A. \tag{3}$$

We will also make use of the following formula.

Proposition 49.3. *If J is a quadratic functional, then*

$$J(u + \rho v) = \frac{\rho^2}{2} a(v, v) + \rho(a(u, v) - h(v)) + J(u).$$

Proof. Since a is symmetric bilinear and h is linear, we have

$$\begin{aligned} J(u+\rho v) &= \frac{1}{2}a(u+\rho v, u+\rho v) - h(u+\rho v) \\ &= \frac{\rho^2}{2}a(v,v) + \rho a(u,v) + \frac{1}{2}a(u,u) - h(u) - \rho h(v) \\ &= \frac{\rho^2}{2}a(v,v) + \rho(a(u,v) - h(v)) + J(u). \end{aligned}$$

Since $dJ_u(v) = a(u,v) - h(v) = \langle Au - b, v\rangle$ and $\nabla J_u = Au - b$, we can also write

$$J(u+\rho v) = \frac{\rho^2}{2}a(v,v) + \rho\langle \nabla J_u, v\rangle + J(u),$$

as claimed. □

We have the following theorem about the existence and uniqueness of minima of quadratic functionals.

Theorem 49.4. *Given any real Hilbert space V, let $J\colon V \to \mathbb{R}$ be a quadratic functional of the form*

$$J(v) = \frac{1}{2}a(v,v) - h(v).$$

Assume that there is some real number $\alpha > 0$ such that

$$a(v,v) \geq \alpha \left\|v\right\|^2 \quad \text{for all } v \in V. \tag{$*_\alpha$}$$

If U is any nonempty, closed, convex subset of V, then there is a unique $u \in U$ such that

$$J(u) = \inf_{v\in U} J(v).$$

The element $u \in U$ satisfies the condition

$$a(u, v-u) \geq h(v-u) \quad \text{for all } v \in U. \tag{$*$}$$

Conversely (with the same assumptions on U as above), if an element $u \in U$ satisfies $(*)$, then

$$J(u) = \inf_{v\in U} J(v).$$

If U is a subspace of V, then the above inequalities are replaced by the equations

$$a(u,v) = h(v) \quad \text{for all } v \in U. \tag{$**$}$$

Proof. The key point is that the bilinear form a is actually an inner product in V. This is because it is positive definite, since $(*_\alpha)$ implies that

$$\sqrt{\alpha}\left\|v\right\| \leq (a(v,v))^{1/2},$$

and on the other hand the continuity of a implies that

$$a(v, v) \leq \|a\| \, \|v\|^2 ,$$

so we get

$$\sqrt{\alpha} \, \|v\| \leq (a(v, v))^{1/2} \leq \sqrt{\|a\|} \, \|v\| .$$

The above also shows that the norm $v \mapsto (a(v, v))^{1/2}$ induced by the inner product a is equivalent to the norm induced by the inner product $\langle -, - \rangle$ on V. Thus h is still continuous with respect to the norm $v \mapsto (a(v, v))^{1/2}$. Then by the Riesz representation theorem (Proposition 48.9), there is some unique $c \in V$ such that

$$h(v) = a(c, v) \quad \text{for all } v \in V.$$

Consequently, we can express $J(v)$ as

$$J(v) = \frac{1}{2}a(v, v) - a(c, v) = \frac{1}{2}a(v - c, v - c) - \frac{1}{2}a(c, c).$$

But then minimizing $J(v)$ over U is equivalent to minimizing $(a(v - c, v - c))^{1/2}$ over $v \in U$, and by the projection lemma (Proposition 48.5(1)) this is equivalent to finding the projection $p_U(c)$ of c on the closed convex set U with respect to the inner product a. Therefore, there is a unique $u = p_U(c) \in U$ such that

$$J(u) = \inf_{v \in U} J(v).$$

Also by Proposition 48.5(2), this unique element $u \in U$ is characterized by the condition

$$a(u - c, v - u) \geq 0 \quad \text{for all } v \in U.$$

Since

$$a(u - c, v - u) = a(u, v - u) - a(c, v - u) = a(u, v - u) - h(v - u),$$

the above inequality is equivalent to

$$a(u, v - u) \geq h(v - u) \quad \text{for all } v \in U. \tag{$*$}$$

If U is a subspace of V, then by Proposition 48.5(3) we have the condition

$$a(u - c, v) = 0 \quad \text{for all } v \in U,$$

which is equivalent to

$$a(u, v) = a(c, v) = h(v) \quad \text{for all } v \in U, \tag{$**$}$$

a claimed. $\square$

Note that the symmetry of the bilinear form a played a crucial role. Also, the inequalities

$$a(u, v-u) \geq h(v-u) \quad \text{for all } v \in U$$

are sometimes called *variational inequalities*.

Definition 49.5. A bilinear form $a\colon V \times V \to \mathbb{R}$ such that there is some real $\alpha > 0$ such that

$$a(v,v) \geq \alpha \left\| v \right\|^2 \quad \text{for all } v \in V$$

is said to be *coercive*.

Theorem 49.4 is the special case of Stampacchia's theorem and the Lax–Milgram theorem when $U = V$, and where a is a symmetric bilinear form. To prove Stampacchia's theorem in general, we need to recall the *contraction mapping theorem*.

Definition 49.6. Let (E, d) be a metric space. A map $f\colon E \to E$ is a *contraction* (or a *contraction mapping*) if there is some real number k such that $0 \leq k < 1$ and

$$d(f(u), f(v)) \leq kd(u, v) \quad \text{for all } u, v \in E.$$

The number k is often called a *Lipschitz constant*.

The following theorem is proven in Section 37.10; see Theorem 37.54. A proof can be also found in Apostol [4], Dixmier [51], or Schwartz [150], among many sources. For the reader's convenience we restate this theorem.

Theorem 49.5. *(Contraction Mapping Theorem) Let (E, d) be a complete metric space. Every contraction $f\colon E \to E$ has a unique fixed point (that is, an element $u \in E$ such that $f(u) = u$).*

The contraction mapping theorem is also known as the *Banach fixed point theorem*.

Theorem 49.6. *(Lions–Stampacchia) Given a Hilbert space V, let $a\colon V \times V \to \mathbb{R}$ be a continuous bilinear form (not necessarily symmetric), let $h \in V'$ be a continuous linear form, and let J be given by*

$$J(v) = \frac{1}{2}\, a(v,v) - h(v), \quad v \in V.$$

If a is coercive, then for every nonempty, closed, convex subset U of V, there is a unique $u \in U$ such that

$$a(u, v-u) \geq h(v-u) \quad \textit{for all } v \in U. \tag{$*$}$$

If a is symmetric, then $u \in U$ is the unique element of U such that

$$J(u) = \inf_{v \in U} J(v).$$

Proof. As discussed just after Definition 49.4, by Proposition 48.10, there is a unique continuous linear map $A\colon V \to V$ such that

$$a(u,v) = \langle Au, v\rangle \quad \text{for all } u, v \in V,$$

with $\|A\| = \|a\| = C$, and by the Riesz representation theorem (Proposition 48.9), there is a unique $b \in V$ such that

$$h(v) = \langle b, v\rangle \quad \text{for all } v \in V.$$

Consequently, J can be written as

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \quad \text{for all } v \in V. \tag{$*_1$}$$

Since $\|A\| = \|a\| = C$, we have $\|Av\| \leq \|A\| \|v\| = C \|v\|$ for all $v \in V$. Using $(*_1)$, the inequality $(*)$ is equivalent to finding u such that

$$\langle Au, v-u\rangle \geq \langle b, v-u\rangle \quad \text{for all } v \in U. \tag{$*_2$}$$

Let $\rho > 0$ be a constant to be determined later. Then $(*_2)$ is equivalent to

$$\langle \rho b - \rho Au + u - u, v - u\rangle \leq 0 \quad \text{for all } v \in U. \tag{$*_3$}$$

By the projection lemma (Proposition 48.5 (1) and (2)), $(*_3)$ is equivalent to finding $u \in U$ such that

$$u = p_U(\rho b - \rho Au + u). \tag{$*_4$}$$

We are led to finding a fixed point of the function $F\colon U \to U$ given by

$$F(v) = p_U(\rho b - \rho Av + v).$$

By Proposition 48.6, the projection map p_U does not increase distance, so

$$\|F(v_1) - F(v_2)\| \leq \|v_1 - v_2 - \rho(Av_1 - Av_2)\|.$$

Since a is coercive we have

$$a(v,v) \geq \alpha \|v\|^2,$$

since $a(v,v) = \langle Av, v\rangle$ we have

$$\langle Av, v\rangle \geq \alpha \|v\|^2 \quad \text{for all } v \in V, \tag{$*_5$}$$

and since

$$\|Av\| \leq C \|v\| \quad \text{for all } v \in V, \tag{$*_6$}$$

we get

$$\begin{aligned}\|F(v_1) - F(v_2)\|^2 &\leq \|v_1 - v_2\|^2 - 2\rho\langle Av_1 - Av_2, v_1 - v_2\rangle + \rho^2 \|Av_1 - Av_2\|^2 \\ &\leq \left(1 - 2\rho\alpha + \rho^2 C^2\right) \|v_1 - v_2\|^2.\end{aligned}$$

If we pick $\rho > 0$ such that $\rho < 2\alpha/C^2$, then

$$k^2 = 1 - 2\rho\alpha + \rho^2 C^2 < 1,$$

and then

$$\|F(v_1) - F(v_2)\| \leq k \|v_1 - v_2\|, \tag{$*_7$}$$

with $0 \leq k < 1$, which shows that F is a contraction. By Theorem 49.5, the map F has a unique fixed point $u \in U$, which concludes the proof of the first statement. If a is also symmetric, then the second statement is just the first part of Theorem 49.4. ☐

Remark: Many physical problems can be expressed in terms of an unknown function u that satisfies some inequality

$$a(u, v - u) \geq h(v - u) \quad \text{for all } v \in U,$$

for some set U of "admissible" functions which is closed and convex. The bilinear form a and the linear form h are often given in terms of integrals. The above inequality is called a *variational inequality*.

In the special case where $U = V$ we obtain the Lax–Milgram theorem.

Theorem 49.7. *(Lax–Milgram's Theorem) Given a Hilbert space V, let $a\colon V \times V \to \mathbb{R}$ be a continuous bilinear form (not necessarily symmetric), let $h \in V'$ be a continuous linear form, and let J be given by*

$$J(v) = \frac{1}{2} a(v, v) - h(v), \quad v \in V.$$

If a is coercive, which means that there is some $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|^2 \quad \text{for all } v \in V,$$

then there is a unique $u \in V$ such that

$$a(u, v) = h(v) \quad \text{for all } v \in V.$$

If a is symmetric, then $u \in V$ is the unique element of V such that

$$J(u) = \inf_{v \in V} J(v).$$

The Lax–Milgram theorem plays an important role in solving linear elliptic partial differential equations; see Brezis [31].

We now consider various methods, known as gradient descents, to find minima of certain types of functionals.

49.4 Elliptic Functionals

We begin by defining the notion of an elliptic functional which generalizes the notion of a quadratic function defined by a symmetric positive definite matrix. Elliptic functionals are well adapted to the types of iterative methods described in this section and lend themselves well to an analysis of the convergence of these methods.

Definition 49.7. Given a Hilbert space V, a functional $J \colon V \to \mathbb{R}$ is said to be *elliptic* if it is continuously differentiable on V, and if there is some constant $\alpha > 0$ such that

$$\langle \nabla J_v - \nabla J_u, v - u \rangle \geq \alpha \left\| v - u \right\|^2 \quad \text{for all } u, v \in V.$$

The following proposition gathers properties of elliptic functionals that will be used later to analyze the convergence of various gradient descent methods.

Theorem 49.8. *Let V be a Hilbert space.*

(1) An elliptic functional $J \colon V \to \mathbb{R}$ is strictly convex and coercive. Furthermore, it satisfies the identity

$$J(v) - J(u) \geq \langle \nabla J_u, v - u \rangle + \frac{\alpha}{2} \left\| v - u \right\|^2 \quad \textit{for all } u, v \in V.$$

(2) If U is a nonempty, convex, closed subset of the Hilbert space V and if J is an elliptic functional, then Problem (P),

$$\begin{aligned} &\text{find} \quad u \\ &\text{such that} \ \ u \in U \ \textit{ and } J(u) = \inf_{v \in U} J(v) \end{aligned}$$

has a unique solution.

(3) Suppose the set U is convex and that the functional J is elliptic. Then an element $u \in U$ is a solution of Problem (P) if and only if it satisfies the condition

$$\langle \nabla J_u, v - u \rangle \geq 0 \quad \textit{for every } v \in U$$

in the general case, or

$$\nabla J_u = 0 \ \textit{ if } U = V.$$

(4) A functional J which is twice differentiable in V is elliptic if and only if

$$\langle \nabla^2 J_u(w), w \rangle \geq \alpha \left\| w \right\|^2 \quad \textit{for all } u, w \in V.$$

Proof. (1) Since J is a C^1-function, by Taylor's formula with integral remainder in the case $m = 0$ (Theorem 39.26), we get

$$\begin{aligned}
J(v) - J(u) &= \int_0^1 dJ_{u+t(v-u)}(v-u)dt \\
&= \int_0^1 \langle \nabla J_{u+t(v-u)}, v-u\rangle dt \\
&= \langle \nabla J_u, v-u\rangle + \int_0^1 \langle \nabla J_{u+t(v-u)} - \nabla J_u, v-u\rangle dt \\
&= \langle \nabla J_u, v-u\rangle + \int_0^1 \frac{\langle \nabla J_{u+t(v-u)} - \nabla J_u, t(v-u)\rangle}{t} dt \\
&\geq \langle \nabla J_u, v-u\rangle + \int_0^1 \alpha t \left\|v-u\right\|^2 dt \qquad \text{since } J \text{ is elliptic} \\
&= \langle \nabla J_u, v-u\rangle + \frac{\alpha}{2}\left\|v-u\right\|^2.
\end{aligned}$$

Using the inequality

$$J(v) - J(u) \geq \langle \nabla J_u, v-u\rangle + \frac{\alpha}{2}\left\|v-u\right\|^2 \quad \text{for all } u, v \in V,$$

by Proposition 40.11(2), since

$$J(v) > J(u) + \langle \nabla J_u, v-u\rangle \quad \text{for all } u, v \in V,\ v \neq u,$$

the function J is strictly convex. It is coercive because (using Cauchy–Schwarz)

$$\begin{aligned}
J(v) &\geq J(0) + \langle \nabla J_0, v\rangle + \frac{\alpha}{2}\left\|v\right\|^2 \\
&\geq J(0) - \left\|\nabla J_0\right\|\left\|v\right\| + \frac{\alpha}{2}\left\|v\right\|^2,
\end{aligned}$$

and the term $(-\left\|\nabla J_0\right\| + \frac{\alpha}{2}\left\|v\right\|)\left\|v\right\|$ goes to $+\infty$ when $\left\|v\right\|$ tends to $+\infty$.

(2) Since by (1) the functional J is coercive, by Theorem 49.2, Problem (P) has a solution. Since J is strictly convex, by Theorem 40.13(2), it has a unique minimum.

(3) These are just the conditions of Theorem 40.13(3, 4).

(4) If J is twice differentiable, we showed in Section 39.6 that we have

$$\mathrm{D}^2 J_u(w,w) = \mathrm{D}_w(\mathrm{D}J)(u) = \lim_{\theta \mapsto 0} \frac{\mathrm{D}J_{u+\theta w}(w) - \mathrm{D}J_u(w)}{\theta},$$

and since

$$\begin{aligned}
\mathrm{D}^2 J_u(w,w) &= \langle \nabla^2 J_u(w), w\rangle \\
\mathrm{D}J_{u+\theta w}(w) &= \langle \nabla J_{u+\theta w}, w\rangle \\
\mathrm{D}J_u(w) &= \langle \nabla J_u, w\rangle,
\end{aligned}$$

and since J is elliptic, for all $u, w \in V$ we can write

$$\begin{aligned}
\langle \nabla^2 J_u(w), w \rangle &= \lim_{\theta \mapsto 0} \frac{\langle \nabla J_{u+\theta w} - \nabla J_u, w \rangle}{\theta} \\
&= \lim_{\theta \mapsto 0} \frac{\langle \nabla J_{u+\theta w} - \nabla J_u, \theta w \rangle}{\theta^2} \\
&\geq \theta \left\| w \right\|^2 .
\end{aligned}$$

Conversely, assume that the condition

$$\langle \nabla^2 J_u(w), w \rangle \geq \alpha \left\| w \right\|^2 \quad \text{for all } u, w \in V$$

holds. If we define the function $g \colon V \to \mathbb{R}$ by

$$g(w) = \langle \nabla J_w, v - u \rangle = dJ_w(v - u) = \mathrm{D}_{v-u} J(w),$$

where u and v are fixed vectors in V, then we have

$$dg_{u+\theta(v-u)}(v-u) = \mathrm{D}_{v-u} g(u+\theta(v-u)) = \mathrm{D}_{v-u} \mathrm{D}_{v-u} J(u+\theta(v-u)) = \mathrm{D}^2 J_{u+\theta(v-u)}(v-u, v-u)$$

and we can apply the Taylor–MacLaurin formula (Theorem 39.25 with $m = 0$) to g, and we get

$$\begin{aligned}
\langle \nabla J_v - \nabla J_u, v - u \rangle &= g(v) - g(u) \\
&= dg_{u+\theta(v-u)}(v - u) \qquad (0 < \theta < 1) \\
&= \mathrm{D}^2 J_{u+\theta(v-u)}(v - u, v - u) \\
&= \langle \nabla^2 J_{u+\theta(v-u)}(v - u), v - u \rangle \\
&\geq \alpha \left\| v - u \right\|^2 ,
\end{aligned}$$

which shows that J is elliptic. $\square$

Corollary 49.9. *If $J \colon \mathbb{R}^n \to \mathbb{R}$ is a quadratic function given by*

$$J(v) = \frac{1}{2} \langle Av, v \rangle - \langle b, v \rangle$$

(where A is a symmetric $n \times n$ matrix and $\langle -, - \rangle$ is the standard Eucidean inner product), then J is elliptic iff A is positive definite.

This a consequence of Theorem 49.8 because

$$\langle \nabla^2 J_u(w), w \rangle = \langle Aw, w \rangle \geq \lambda_1 \left\| w \right\|^2$$

where λ_1 is the smallest eigenvalue of A; see Proposition 17.24 (Rayleigh–Ritz). Note that by Proposition 17.24 (Rayleigh–Ritz), we also have the folllowing corollary.

Corollary 49.10. *If $J\colon \mathbb{R}^n \to \mathbb{R}$ is a quadratic function given by*

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$$

then

$$\langle \nabla^2 J_u(w), w\rangle \leq \lambda_n \left\|w\right\|^2$$

where λ_n is the largest eigenvalue of A;

The above fact will be useful later on.

Similarly, given a quadratic functional J defined on a Hilbert space V, where

$$J(v) = \frac{1}{2}a(v, v) - h(v),$$

by Theorem 49.8 (4), the functional J is elliptic iff there is some $\alpha > 0$ such that

$$\langle \nabla^2 J_u(v), v\rangle = a(v, v) \geq \alpha \left\|v\right\|^2 \quad \text{for all } v \in V.$$

This is precisely the hypothesis $(*_\alpha)$ used in Theorem 49.4.

49.5 Iterative Methods for Unconstrained Problems

We will now describe methods for solving unconstrained minimization problems, that is, finding the minimum (or minima) of a functions J over the whole space V. These methods are *iterative*, which means that given some *initial* vector u_0, we construct a sequence $(u_k)_{k\geq 0}$ that converges to a minimum u of the function J.

The key step is define u_{k+1} from u_k, and a first idea is to reduce the problem to a simpler problem, namely the minimization of a function of a *single (real) variable*. For this, we need two perform two steps:

(1) Find a *descent direction* at u_k, which is a some nonzero vector d_k which is usually determined from the gradient of J at various points. The descent direction d_k must satisfy the inequality $\langle \nabla J_{u_k}, d_k\rangle < 0$.

(2) *Exact line search*: Find the minimum of the restriction of the function J along the line through u_k and parallel to the direction d_k. This means finding a real $\rho_k \in \mathbb{R}$ (depending on u_k and d_k) such that

$$J(u_k + \rho_k d_k) = \inf_{\rho\in\mathbb{R}} J(u_k + \rho d_k).$$

Typically, $\rho_k > 0$. This problem only succeeds if ρ_k is *unique*, in which case we set

$$u_{k+1} = u_k + \rho_k d_k.$$

This step is often called a *line search* or *line minimization*, and ρ_k is called the *stepsize* parameter. See Figure 49.1.

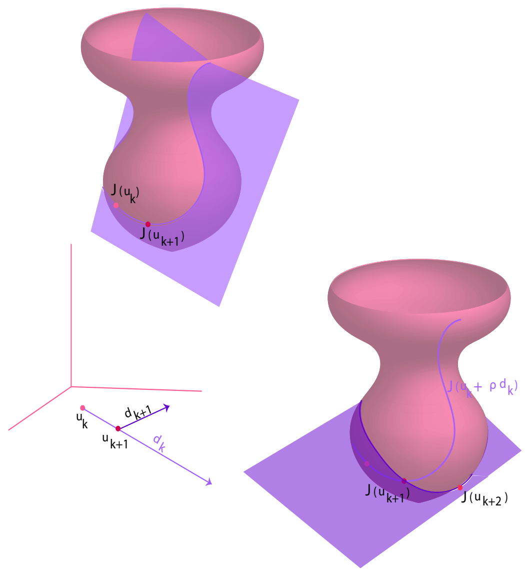


Figure 49.1: Let $J\colon \mathbb{R}^2 \to \mathbb{R}$ be the function whose graph is represented by the pink surface. Given a point u_k in the xy-plane, and a direction d_k, we calculate first u_{k+1} and then u_{k+2}.

Proposition 49.11. *If J is a quadratic elliptic functional of the form*

$$J(v) = \frac{1}{2}a(v, v) - h(v),$$

then given d_k, there is a unique ρ_k solving the line search in Step (2).

Proof. This is because, by Proposition 49.3, we have

$$J(u_k + \rho d_k) = \frac{\rho^2}{2}a(d_k, d_k) + \rho\langle \nabla J_{u_k}, d_k\rangle + J(u_k),$$

and since $a(d_k, d_k) > 0$ (because J is elliptic), the above function of ρ has a unique minimum when its derivative is zero, namely

$$\rho\, a(d_k, d_k) + \langle \nabla J_{u_k}, d_k\rangle = 0. \qquad \square$$

Since Step (2) is often too costly, an alternative is

(3) *Backtracking line search*: Pick two constants α and β such that $0 < \alpha < 1/2$ and $0 < \beta < 1$, and set $t = 1$. Given a descent direction d_k at $u_k \in \mathrm{dom}(J)$,

while $J(u_k + td_k) > J(u_k) + \alpha t\langle \nabla J_{u_k}, d_k\rangle$ **do** $t := \beta t$;
$\rho_k = t$; $u_{k+1} = u_k + \rho_k d_k$.

Since d_k is a descent direction, we must have $\langle \nabla J_{u_k}, d_k \rangle < 0$, so for t small enough the condition $J(u_k + td_k) \leq J(u_k) + \alpha t \langle \nabla J_{u_k}, d_k \rangle$ will hold and the search will stop. It can be shown that the exit inequality $J(u_k + td_k) \leq J(u_k) + \alpha t \langle \nabla J_{u_k}, d_k \rangle$ holds for all $t \in (0, t_0]$, for some $t_0 > 0$. Thus the backtracking line search stops with a step length ρ_k that satisfies $\rho_k = 1$ or $\rho_k \in (\beta t_0, t_0]$. Care has to be exercised so that $u_k + \rho_k d_k \in \mathrm{dom}(J)$. For more details, see Boyd and Vandenberghe [29] (Section 9.2).

We now consider one of the simplest methods for choosing the directions of descent in the case where $V = \mathbb{R}^n$, which is to pick the directions of the coordinate axes in a cyclic fashion. Such a method is called the *method of relaxation*.

If we write

$$u_k = (u_1^k, u_2^k, \ldots, u_n^k),$$

then the components u_i^{k+1} of u_{k+1} are computed in terms of u_k by solving from top down the following system of equations:

$$\begin{aligned}
J(\mathbf{u_1^{k+1}}, u_2^k, u_3^k, \ldots, u_n^k) &= \inf_{\lambda \in \mathbb{R}} J(\lambda, u_2^k, u_3^k, \ldots, u_n^k) \\
J(u_1^{k+1}, \mathbf{u_2^{k+1}}, u_3^k, \ldots, u_n^k) &= \inf_{\lambda \in \mathbb{R}} J(u_1^{k+1}, \lambda, u_3^k, \ldots, u_n^k) \\
&\vdots \\
J(u_1^{k+1}, \ldots, u_{n-1}^{k+1}, \mathbf{u_n^{k+1}}) &= \inf_{\lambda \in \mathbb{R}} J(u_1^{k+1}, \ldots, u_{n-1}^{k+1}, \lambda).
\end{aligned}$$

Another and more informative way to write the above system is to define the vectors $u_{k;i}$ by

$$\begin{aligned}
u_{k;0} &= (u_1^k, u_2^k, \ldots, u_n^k) \\
u_{k;1} &= (u_1^{k+1}, u_2^k, \ldots, u_n^k) \\
&\vdots \\
u_{k;i} &= (u_1^{k+1}, \ldots, u_i^{k+1}, u_{i+1}^k, \ldots, u_n^k) \\
&\vdots \\
u_{k;n} &= (u_1^{k+1}, u_2^{k+1}, \ldots, u_n^{k+1}).
\end{aligned}$$

Note that $u_{k;0} = u_k$ and $u_{k;n} = u_{k+1}$. Then our minimization problem can be written as

$$\begin{aligned}
J(u_{k;1}) &= \inf_{\lambda \in \mathbb{R}} J(u_{k;0} + \lambda e_1) \\
&\vdots \\
J(u_{k;i}) &= \inf_{\lambda \in \mathbb{R}} J(u_{k;i-1} + \lambda e_i) \\
&\vdots \\
J(u_{k;n}) &= \inf_{\lambda \in \mathbb{R}} J(u_{k;n-1} + \lambda e_n),
\end{aligned}$$

where e_i denotes the ith canonical basis vector in $\mathbb{R}^n$. If J is differentiable, necessary conditions for a minimum, which are also sufficient if J is convex, is that the directional derivatives $dJ_v(e_i)$ be all zero, that is,

$$\langle \nabla J_v, e_i \rangle = 0 \quad i = 0, \ldots, n.$$

The following result regarding the convergence of the mehod of relaxation is proven in Ciarlet [41] (Chapter 8, Theorem 8.4.2).

Proposition 49.12. *If the functional $J \colon \mathbb{R}^n \to \mathbb{R}$ is elliptic, then the relaxation method converges.*

Remarks: The proof of Proposition 49.12 uses Theorem 49.8. The finite dimensionality of $\mathbb{R}^n$ also plays a crucial role. The differentiability of the function J is also crucial. Examples where the method loops forever if J is not differentiable can be given; see Ciarlet [41] (Chapter 8, Section 8.4). The proof of Proposition 49.12 yields an *a priori* bound on the error $\|u - u_k\|$. If J is a quadratic functional

$$J(v) = \frac{1}{2} v^\top A v - b^\top v,$$

where A is a symmetric positive definite matrix, then $\nabla J_v = Av - b$, so the above method for solving for u_{k+1} in terms of u_k becomes the *Gauss–Seidel method* for solving a linear system; see Section 10.3.

We now discuss gradient methods.

49.6 Gradient Descent Methods for Unconstrained Problems

The intuition behind these methods is that the convergence of an iterative method ought to be better if the difference $J(u_k) - J(u_{k+1})$ is as large as possible during every iteration step. To achieve this, it is natural to pick the descent direction to be the one *in the opposite direction of the gradient vector* ∇J_{u_k}. This choice is justified by the fact that we can write

$$J(u_k + w) = J(u_k) + \langle \nabla J_{u_k}, w \rangle + \epsilon(w) \|w\|, \quad \text{with } \lim_{w \mapsto 0} \epsilon(w) = 0.$$

If $\nabla J_{u_k} \neq 0$, the first-order part of the variation of the function J is bounded in absolute value by $\|\nabla J_{u_k}\| \|w\|$ (by the Cauchy–Schwarz inequality), with equality if ∇J_{u_k} and w are collinear.

Gradient descent methods pick the direction of descent to be $d_k = -\nabla J_{u_k}$, so that we have

$$u_{k+1} = u_k - \rho_k \nabla J_{u_k},$$

where we put a negative sign in front of of the variable ρ_k as a reminder that the descent direction is *opposite* to that of the gradient; a positive value is expected for the scalar ρ_k.

There are four standard methods to pick ρ_k:

(1) *Gradient method with fixed stepsize parameter*. This is the simplest and cheapest method which consists of using the *same* constant $\rho_k = \rho$ for all iterations.

(2) *Gradient method with variable stepsize parameter*. In this method, the parameter ρ_k is adjusted in the course of iterations according to various criteria.

(3) *Gradient method with optimal stepsize parameter*, also called *steepest descent method for the Euclidean norm*. This is a version of Method 2 in which ρ_k is determined by the following line search:

$$J(u_k - \rho_k \nabla J_{u_k}) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho \nabla J_{u_k}).$$

This optimization problem only succeeds if the above minimization problem has a *unique* solution.

(4) *Gradient descent method with backtracking line search*. In this method, the step parameter is obtained by performing a backtracking line search.

We have the following useful result about the convergence of the gradient method with optimal parameter.

Proposition 49.13. *Let $J\colon \mathbb{R}^n \to \mathbb{R}$ be an elliptic functional. Then the gradient method with optimal stepsize parameter converges.*

Proof. Since J is elliptic, by Theorem 49.8(3), the functional J has a unique minimum u characterized by $\nabla J_u = 0$. Our goal is to prove that the sequence $(u_k)_{k \geq 0}$ constructed using the gradient method with optimal parameter converges to u, starting from any initial vector u_0. Without loss of generality we may assume that $u_{k+1} \neq u_k$ and $\nabla J_{u_k} \neq 0$ for all k, since otherwise the method converges in a finite number of steps.

Step 1. Show that any two consecutive descent directions are *orthogonal* and

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2} \left\| u_k - u_{k+1} \right\|^2.$$

Let $\varphi_k \colon \mathbb{R} \to \mathbb{R}$ be the function given by

$$\varphi_k(\rho) = J(u_k - \rho \nabla J_{u_k}).$$

Since the function φ_k is strictly convex and coercive, by Theorem 49.8(2), it has a unique minimum ρ_k which is the unique solution of the equation $\varphi_k'(\rho) = 0$. By the chain rule

$$\begin{aligned}
\varphi_k'(\rho) &= dJ_{u_k - \rho \nabla J_{u_k}}(-\nabla J_{u_k}) \\
&= -\langle \nabla J_{u_k - \rho \nabla J_{u_k}}, \nabla J_{u_k} \rangle,
\end{aligned}$$

and since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ we get

$$\langle \nabla J_{u_{k+1}}, \nabla J_{u_k} \rangle = 0,$$

which shows that two consecutive descent directions are orthogonal.

Since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ and we assumed that that $u_{k+1} \neq u_k$, we have $\rho_k \neq 0$, and we also get

$$\langle \nabla J_{u_{k+1}}, u_{k+1} - u_k \rangle = 0.$$

By the inequality of Theorem 49.8(1) we have

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2} \left\| u_k - u_{k+1} \right\|^2.$$

Step 2. Show that $\lim_{k \mapsto \infty} \| u_k - u_{k+1} \| = 0$.

It follows from the inequality proven in Step 1 that the sequence $(J(u_k))_{k \geq 0}$ is decreasing and bounded below (by $J(u)$, where u is the minimum of J), so it converges and we conclude that

$$\lim_{k \mapsto \infty} (J(u_k) - J(u_{k+1})) = 0,$$

which combined with the preceding inequality shows that

$$\lim_{k \mapsto \infty} \| u_k - u_{k+1} \| = 0.$$

Step 3. Show that $\| \nabla J_{u_k} \| \leq \left\| \nabla J_{u_k} - \nabla J_{u_{k+1}} \right\|$.

Using the orthogonality of consecutive descent directions, by Cauchy–Schwarz we have

$$\begin{aligned} \| \nabla J_{u_k} \|^2 &= \langle \nabla J_{u_k}, \nabla J_{u_k} - \nabla J_{u_{k+1}} \rangle \\ &\leq \| \nabla J_{u_k} \| \left\| \nabla J_{u_k} - \nabla J_{u_{k+1}} \right\|, \end{aligned}$$

so that

$$\| \nabla J_{u_k} \| \leq \left\| \nabla J_{u_k} - \nabla J_{u_{k+1}} \right\|.$$

Step 4. Show that $\lim_{k \mapsto \infty} \| \nabla J_{u_k} \| = 0$.

Since the sequence $(J(u_k))_{k \geq 0}$ is decreasing and the functional J is coercive, the sequence $(u_k)_{k \geq 0}$ must be bounded. By hypothesis, the derivative dJ is of J is continuous, so it is uniformly continuous over compact subsets of $\mathbb{R}^n$; here we are using the fact that $\mathbb{R}^n$ is finite dimensional. Hence, we deduce that for every $\epsilon > 0$, there is some $\delta > 0$ such that if $\| u_k - u_{k+1} \| < \delta$ then

$$\left\| dJ_{u_k} - dJ_{u_{k+1}} \right\|_2 < \epsilon.$$

But by definition of the operator norm and using the Cauchy–Schwarz inequality

$$\begin{aligned}\left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 &= \sup_{\|w\|=1} |dJ_{u_k}(w) - dJ_{u_{k+1}}(w)| \\ &= \sup_{\|w\|=1} |\langle \nabla J_{u_k} - \nabla J_{u_{k+1}}, w\rangle| \\ &\leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|.\end{aligned}$$

But we also have

$$\begin{aligned}\left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|^2 &= \langle \nabla J_{u_k} - \nabla J_{u_{k+1}}, \nabla J_{u_k} - \nabla J_{u_{k+1}}\rangle \\ &= dJ_{u_k}(\nabla J_{u_k} - \nabla J_{u_{k+1}}) - dJ_{u_{k+1}}(\nabla J_{u_k} - \nabla J_{u_{k+1}}) \\ &\leq \left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2^2,\end{aligned}$$

and so

$$\left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 = \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|.$$

It follows that since

$$\lim_{k\mapsto\infty} \|u_k - u_{k+1}\| = 0$$

then

$$\lim_{k\mapsto\infty} \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\| = \lim_{k\mapsto\infty} \left\|dJ_{u_k} - dJ_{u_{k+1}}\right\|_2 = 0,$$

and using the fact that

$$\|\nabla J_{u_k}\| \leq \left\|\nabla J_{u_k} - \nabla J_{u_{k+1}}\right\|,$$

we obtain

$$\lim_{k\mapsto\infty} \|\nabla J_{u_k}\| = 0.$$

Step 5. Finally we can prove the convergence of the sequence $(u_k)_{k\geq 0}$.

Since J is elliptic and since $\nabla J_u = 0$ (since u is the minimum of J over $\mathbb{R}^n$), we have

$$\begin{aligned}\alpha \|u_k - u\|^2 &\leq \langle \nabla J_{u_k} - \nabla J_u, u_k - u\rangle \\ &= \langle \nabla J_{u_k}, u_k - u\rangle \\ &\leq \|\nabla J_{u_k}\| \, \|u_k - u\|.\end{aligned}$$

Hence, we obtain

$$\|u_k - u\| \leq \frac{1}{\alpha} \|\nabla J_{u_k}\|, \tag{b}$$

and since we showed that

$$\lim_{k\mapsto\infty} \|\nabla J_{u_k}\| = 0,$$

we see that the sequence $(u_k)_{k\geq 0}$ converges to the mininum u. $\square$

Remarks: As with the previous proposition, the assumption of finite dimensionality is crucial. The proof provides an *a priori* bound on the error $\|u_k - u\|$.

If J is a an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

we can use the orthogonality of the descent directions ∇J_{u_k} and $\nabla J_{u_{k+1}}$ to compute ρ_k. Indeed, we have $\nabla J_v = Av - b$, so

$$0 = \langle \nabla J_{u_{k+1}}, \nabla J_{u_k}\rangle = \langle A(u_k - \rho_k(Au_k - b)) - b, Au_k - b\rangle,$$

which yields

$$\rho_k = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle}, \quad \text{with} \quad w_k = Au_k - b = \nabla J_{u_k}.$$

Consequently, a step of the iteration method takes the following form:

(1) Compute the vector

$$w_k = Au_k - b.$$

(2) Compute the scalar

$$\rho_k = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle}.$$

(3) Compute the next vector u_{k+1} by

$$u_{k+1} = u_k - \rho_k w_k.$$

This method is of particular interest when the computation of Aw for a given vector w is cheap, which is the case if A is sparse.

Example 49.1. For a particular illustration of this method, we turn to the example provided by Shewchuk, with $A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$ and $b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$, namely

$$\begin{aligned} J(x, y) &= \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 2 & -8 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \\ &= \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y. \end{aligned}$$

This quadratic ellipsoid, which is illustrated in Figure 49.2, has a unique minimum at $(2, -2)$. In order to find this minimum via the gradient descent with optimal step size

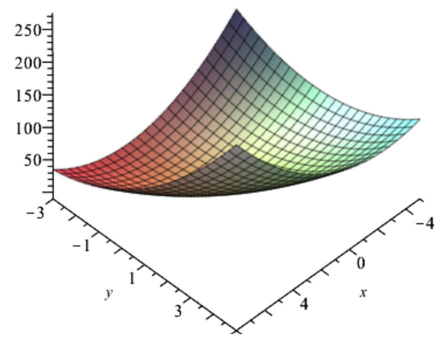


Figure 49.2: The ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$.

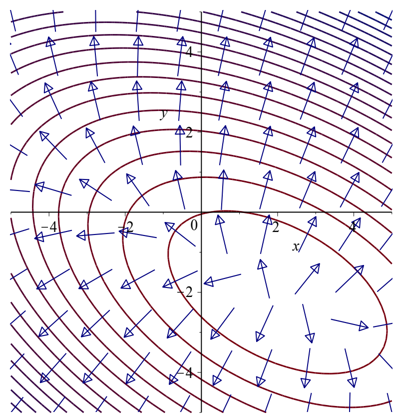


Figure 49.3: The level curves of $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ and the associated gradient vector field $\nabla J(x, y) = (3x + 2y - 2, 2x + 6y + 8)$.

parameter, we pick a starting point, say $u_k = (-2, -2)$, and calculate the search direction $w_k = \nabla J(-2, -2) = (-12, -8)$. Note that

$$\nabla J(x, y) = (3x + 2y - 2, 2x + 6y + 8) = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 2 \\ -8 \end{pmatrix}$$

is perpendicular to the appropriate elliptical level curve; see Figure 49.3. We next perform the line search along the line given by the equation $-8x + 12y = -8$ and determine ρ_k. See Figures 49.4 and 49.5.

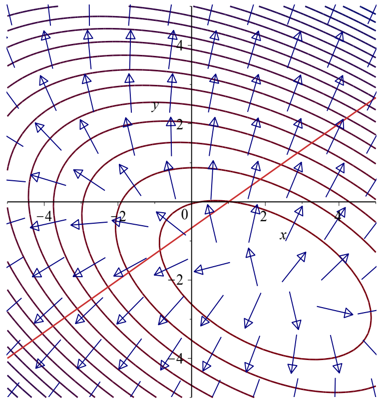


Figure 49.4: The level curves of $J(x,y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ and the red search line with direction $\nabla J(-2,-2) = (-12,-8)$

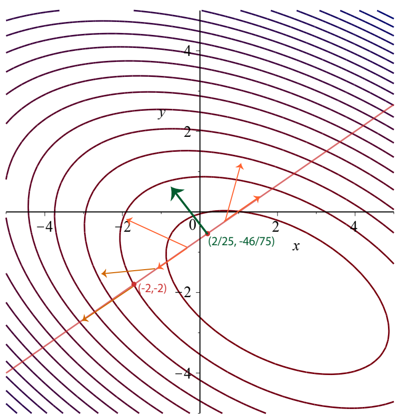


Figure 49.5: Let $u_k = (-2,-2)$. When traversing along the red search line, we look for the green perpendicular gradient vector. This gradient vector, which occurs at $u_{k+1} = (2/25, -46/75)$, provides a minimal ρ_k, since it has no nonzero projection on the search line.

In particular, we find that

$$\rho_k = \frac{\|w_k\|^2}{\langle A w_k, w_k\rangle} = \frac{13}{75}.$$

This in turn gives us the new point

$$u_{k+1} = u_k - \frac{13}{75} w_k = (-2,-2) - \frac{13}{75}(-12,-8) = \left(\frac{2}{25}, -\frac{46}{75}\right),$$

and we continue the procedure by searching along the gradient direction $\nabla J(2/25, -46/75) = (-224/75, 112/25)$. Observe that $u_{k+1} = (\frac{2}{25}, -\frac{46}{75})$ has a gradient vector which is perpendicular to the search line with direction vector $w_k = \nabla J(-2,-2) = (-12\ -8)$; see Figure

49.5. Geometrically this procedure corresponds to intersecting the plane $-8x + 12y = -8$ with the ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$ to form the parabolic curve $f(x) = 25/6x^2 - 2/3x - 4$, and then locating the x-coordinate of its apex which occurs when $f'(x) = 0$, i.e when $x = 2/25$; see Figure 49.6. After 31 iterations, this procedure stabi-

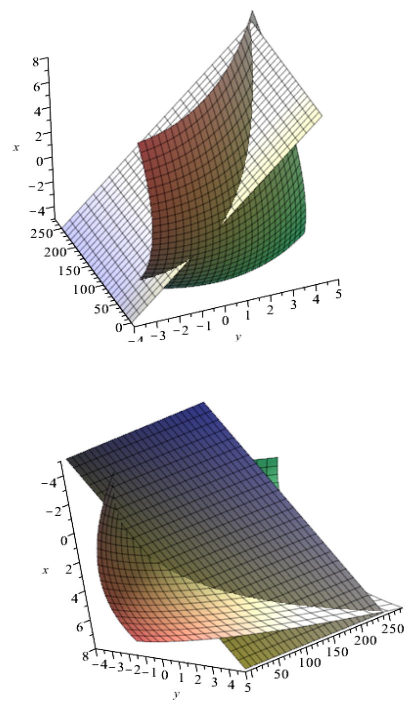


Figure 49.6: Two views of the intersection between the plane $-8x + 12y = -8$ and the ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$. The point u_{k+1} is the minimum of the parabolic intersection.

lizes to point $(2, -2)$, which as we know, is the unique minimum of the quadratic ellipsoid $J(x, y) = \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y$.

A proof of the convergence of the gradient method with backtracking line search, under the hypothesis that J is strictly convex, is given in Boyd and Vandenberghe[29] (Section 9.3.1). More details on this method and the steepest descent method for the Euclidean norm can also be found in Boyd and Vandenberghe [29] (Section 9.3).

49.7 Convergence of Gradient Descent with Variable Stepsize

We now give a sufficient condition for the gradient method with variable stepsize parameter to converge. In addition to requiring J to be an elliptic functional, we add a Lipschitz condition on the gradient of J. This time the space V can be infinite dimensional.

Proposition 49.14. *Let $J\colon V \to \mathbb{R}$ be a continuously differentiable functional defined on a Hilbert space V. Suppose there exists two constants $\alpha > 0$ and $M > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \left\|v - u\right\|^2 \quad \text{for all } u, v \in V,$$

and the Lipschitz condition

$$\left\|\nabla J_v - \nabla J_u\right\| \leq M \left\|v - u\right\| \quad \text{for all } u, v \in V.$$

If there exists two real numbers $a, b \in \mathbb{R}$ such that

$$0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2} \quad \text{for all } k \geq 0,$$

then the gradient method with variable stepsize parameter converges. Furthermore, there is some constant $\beta > 0$ (depending on α, M, a, b) such that

$$\beta < 1 \quad \text{and} \quad \left\|u_k - u\right\| \leq \beta^k \left\|u_0 - u\right\|,$$

where $u \in V$ is the unique minimum of J.

Proof. By hypothesis the functional J is elliptic, so by Theorem 49.8(2) it has a unique minimum u characterized by the fact that $\nabla J_u = 0$. Then since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$, we can write

$$u_{k+1} - u = (u_k - u) - \rho_k(\nabla J_{u_k} - \nabla J_u). \tag{$*$}$$

Using the inequalities

$$\langle \nabla J_{u_k} - \nabla J_u, u_k - u\rangle \geq \alpha \left\|u_k - u\right\|^2$$

and

$$\left\|\nabla J_{u_k} - \nabla J_u\right\| \leq M \left\|u_k - u\right\|,$$

and assuming that $\rho_k > 0$, it follows that

$$\begin{aligned} \left\|u_{k+1} - u\right\|^2 &= \left\|u_k - u\right\|^2 - 2\rho_k \langle \nabla J_{u_k} - \nabla J_u, u_k - u\rangle + \rho_k^2 \left\|\nabla J_{u_k} - \nabla J_u\right\|^2 \\ &\leq \Big(1 - 2\alpha\rho_k + M^2\rho_k^2\Big) \left\|u_k - u\right\|^2. \end{aligned}$$

Consider the function

$$T(\rho) = M^2\rho^2 - 2\alpha\rho + 1.$$

Its graph is a parabola intersecting the y-axis at $y = 1$ for $\rho = 0$, it has a minimum for $\rho = \alpha/M^2$, and it also has the value $y = 1$ for $\rho = 2\alpha/M^2$; see Figure 49.7. Therefore if we pick a, b and ρ_k such that

$$0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2},$$

we ensure that for $\rho \in [a, b]$ we have

$$T(\rho)^{1/2} = (M^2\rho^2 - 2\alpha\rho + 1)^{1/2} \leq (\max\{T(a), T(b)\})^{1/2} = \beta < 1.$$

Then by induction we get

$$\|u_{k+1} - u\| \leq \beta^{k+1} \|u_0 - u\|,$$

which proves convergence. $\square$

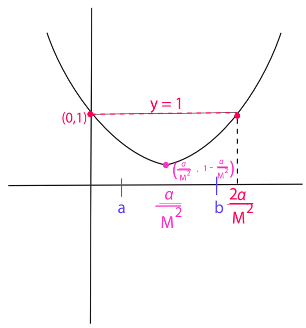


Figure 49.7: The parabola $T(\rho)$ used in the proof of Proposition 49.14.

Remarks: In the proof of Proposition 49.14, it is the fact that V is complete which plays a crucial role. If J is twice differentiable, the hypothesis

$$\|\nabla J_v - \nabla J_u\| \leq M \|v - u\| \quad \text{for all } u, v \in V$$

can be expressed as

$$\sup_{v \in V} \left\|\nabla^2 J_v\right\| \leq M.$$

In the case of a quadratic elliptic functional defined over $\mathbb{R}^n$,

$$J(v) = \langle Av, v\rangle - \langle b, v\rangle,$$

the upper bound $2\alpha/M^2$ can be improved. In this case we have

$$\nabla J_v = Av - b,$$

and we know that $\alpha = \lambda_1$ and $M = \lambda_n$ do the job, where λ_1 is the smallest eigenvalue of A and λ_n is the largest eigenvalue of A. Hence we can pick a, b such that

$$0 < a \leq \rho_k \leq b < \frac{2\lambda_1}{\lambda_n^2}.$$

Since $u_{k+1} = u_k - \rho_k \nabla J_{u_k}$ and $\nabla J_{u_k} = Au_k - b$, we have

$$u_{k+1} - u = (u_k - u) - \rho_k(Au_k - Au) = (I - \rho_k A)(u_k - u),$$

so we get

$$\|u_{k+1} - u\| \leq \|I - \rho_k A\|_2 \, \|u_k - u\| \, .$$

However, since $I - \rho_k A$ is a symmetric matrix, $\|I - \rho_k A\|_2$ is the largest absolute value of its eigenvalues, so

$$\|I - \rho_k A\|_2 \leq \max\{|1 - \rho_k \lambda_1|, |1 - \rho_k \lambda_n|\}.$$

The function

$$\mu(\rho) = \max\{|1 - \rho\lambda_1|, |1 - \rho\lambda_n|\}$$

is a piecewise affine function, and it is easy to see that if we pick a, b such that

$$0 < a \leq \rho_k \leq b < \frac{2}{\lambda_n},$$

then

$$\max_{\rho \in [a,b]} \mu(\rho) \leq \max\{\mu(a), \mu(b)\} < 1.$$

Therefore, the upper bound $2\lambda_1/\lambda_n^2$ can be replaced by $2/\lambda_n$, which is typically much larger. A "good" pick for ρ_k is $2/(\lambda_1 + \lambda_n)$ (as opposed to λ_1/λ_n^2 for the first version). In this case

$$|1 - \rho_k \lambda_1| = |1 - \rho_k \lambda_n| = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1},$$

so we get

$$\beta = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1},$$

where $\mathrm{cond}_2(A) = \lambda_n/\lambda_1$ is the condition number of the matrix A with respect to the spectral norm. Thus we see that the larger the condition number of A is, the slower the convergence of the method will be. This is not surprising since we already know that linear systems involving ill-conditioned matrices (matrices with a large condition number) are problematic and prone to numerical instability. One way to deal with this problem is to use a method known as preconditioning.

We only described the most basic gradient descent methods. There are numerous variants, and we only mention a few of these methods.

The method of *scaling* consists in using $-\rho_k D_k \nabla J_{u_k}$ as descent direction, where D_k is some suitably chosen symmetric positive definite matrix.

In the *gradient method with extrapolation*, u_{k+1} is determined by

$$u_{k+1} = u_k - \rho_k \nabla J_{u_k} + \beta_k(u_k - u_{k-1}).$$

Another rule for choosing the stepsize is *Armijo's rule*.

These methods, and others, are discussed in detail in Berstekas [17].

Boyd and Vandenberghe discuss steepest descent methods for various types of norms besides the Euclidean norm; see Boyd and Vandenberghe [29] (Section 9.4). Here is brief summary.

49.8 Steepest Descent for an Arbitrary Norm

The idea is to make $\langle \nabla J_{u_k}, d_k\rangle$ as negative as possible. To make the question sensible, we have to limit the size of d_k or normalize by the length of d_k.

Let $\| \ \|$ be any norm on $\mathbb{R}^n$. Recall from Section 14.7 that the *dual norm* is defined by

$$\|y\|^D = \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} |\langle x, y\rangle|.$$

Definition 49.8. A *normalized steepest descent direction* (with respect to the norm $\| \ \|$) is any unit vector $d_{\mathrm{nsd},k}$ which achieves the minimum of the set of reals

$$\{\langle \nabla J_{u_k}, d\rangle \mid \|d\| = 1\}.$$

By definition, $\|d_{\mathrm{nsd},k}\| = 1$.

A *unnormalized steepest descent direction* $d_{\mathrm{sd},k}$ is defined as

$$d_{\mathrm{sd},k} = \|\nabla J_{u_k}\|^D \, d_{\mathrm{nsd},k}.$$

It can be shown that

$$\langle \nabla J_{u_k}, d_{\mathrm{sd},k}\rangle = -(\|\nabla J_{u_k}\|^D)^2;$$

see Boyd and Vandenberghe [29] (Section 9.4).

The *steepest descent method (with respect to the norm $\| \ \|$)* consists of the following steps: Given a starting point $u_0 \in \mathrm{dom}(J)$ do:

repeat

(1) Compute the steepest descent direction $d_{\mathrm{sd},k}$.

(2) Line search. Perform an exact or backtracking line search to find ρ_k.

(3) Update. $u_{k+1} = u_k + \rho_k d_{\mathrm{sd},k}$.

until stopping criterion is satisfied.

If $\| \|$ is the ℓ^2-norm, then we see immediately that $d_{\mathrm{sd},k} = -\nabla J_{u_k}$, so in this case the method *coincides* with the steepest descent method for the Euclidean norm as defined at the beginning of Section 49.6 in (3) and (4).

If P is a symmetric positive definite matrix, it is easy to see that $\|z\|_P = (z^\top P z)^{1/2} = \left\|P^{1/2} z\right\|_2$ is a norm. Then it can be shown that the normalized steepest descent direction is

$$d_{\mathrm{nsd,k}} = -(\nabla J_{u_k}^\top P^{-1} \nabla J_{u_k})^{-1/2} P^{-1} \nabla J_{u_k},$$

the dual norm is $\|z\|^D = \left\|P^{-1/2} z\right\|_2$, and the steepest descent direction with respect to $\| \|_P$ is given by

$$d_{\mathrm{sd},k} = -P^{-1} \nabla J_{u_k}.$$

A judicious choice for P can speed up the rate of convergence of the gradient descent method; see see Boyd and Vandenberghe [29] (Section 9.4.1 and Section 9.4.4).

If $\| \|$ is the ℓ^1-norm, then it can be shown that $d_{\mathrm{nsd,k}}$ is determined as follows: let i be any index for which $\|\nabla J_{u_k}\|_\infty = |(\nabla J_{u_k})_i|$. Then

$$d_{\mathrm{nsd,k}} = -\mathrm{sign}\left(\frac{\partial J}{\partial x_i}(u_k)\right) e_i,$$

where e_i is the ith canonical basis vector, and

$$d_{\mathrm{sd,k}} = -\frac{\partial J}{\partial x_i}(u_k) e_i.$$

For more details, see Boyd and Vandenberghe [29] (Section 9.4.2 and Section 9.4.4). It is also shown in Boyd and Vandenberghe [29] (Section 9.4.3) that the steepest descent method converges for any norm $\| \|$ and any strictly convex function J.

One of the main goals in designing a gradient descent method is to ensure that the convergence factor is as small as possible, which means that the method converges as quickly as possible. Machine learning has been a catalyst for finding such methods. A method discussed in Strang [171] (Chapter VI, Section 4) consists in adding a *momentum term* to the gradient. In this method, u_{k+1} and d_{k+1} are determined by the following system of equations:

$$\begin{aligned} u_{k+1} &= u_k - \rho d_k \\ d_{k+1} - \nabla J_{u_{k+1}} &= \beta d_k. \end{aligned}$$

Of course the trick is to choose ρ and β in such a way that the convergence factor is as small as possible. If J is given by a quadratic functional, say $(1/2)u^\top A u - b^\top u$, then $\nabla J_{u_{k+1}} = A u_{k+1} - b$ so we obtain a linear system. It turns out that the rate of convergence of the method is determined by the largest and the smallest eigenvalues of A. Strang discusses this issue in the case of a 2×2 matrix. Convergence is significantly accelerated.

Another method is known as *Nesterov acceleration*. In this method,

$$u_{k+1} = u_k + \beta(u_k - u_{k-1}) - \rho \nabla J_{u_k + \gamma(u_k - u_{k-1})},$$

where β, ρ, γ are parameters. For details, see Strang [171] (Chapter VI, Section 4).

Lax also discusses other methods in which the step ρ_k is chosen using roots of Chebyshev polynomials; see Lax [113], Chapter 17, Sections 2–4.

A variant of Newton's method described in Section 41.2 can be used to find the minimum of a function belonging to a certain class of strictly convex functions. This method is the special case of the case where the norm is induced by a symmetric positive definite matrix P, namely $P = \nabla^2 J(x)$, the Hessian of J at x.

49.9 Newton's Method For Finding a Minimum

If $J\colon \Omega \to \mathbb{R}$ is a convex function defined on some open subset Ω of $\mathbb{R}^n$ which is twice differentiable and if its Hessian $\nabla^2 J(x)$ is symmetric positive definite for all $x \in \Omega$, then by Proposition 40.12(2), the function J is strictly convex. In this case, for any $x \in \Omega$, we have the quadratic norm induced by $P = \nabla^2 J(x)$ as defined in the previous section, given by

$$\|u\|_{\nabla^2 J(x)} = (u^\top \nabla^2 J(x)\, u)^{1/2}.$$

The steepest descent direction for this quadratic norm is given by

$$d_{\mathrm{nt}} = -(\nabla^2 J(x))^{-1} \nabla J_x.$$

The norm of d_{nt} for the the quadratic norm defined by $\nabla^2 J(x)$ is given by

$$\begin{aligned}
(d_{\mathrm{nt}}^\top \nabla^2 J(x)\, d_{\mathrm{nt}})^{1/2} &= \left(-(\nabla J_x)^\top (\nabla^2 J(x))^{-1} \nabla^2 J(x) (-(\nabla^2 J(x))^{-1} \nabla J_x)\right)^{1/2} \\
&= \left((\nabla J_x)^\top (\nabla^2 J(x))^{-1} \nabla J_x\right)^{1/2}.
\end{aligned}$$

Definition 49.9. Given a function $J\colon \Omega \to \mathbb{R}$ as above, for any $x \in \Omega$, the *Newton step* d_{nt} is defined by

$$d_{\mathrm{nt}} = -(\nabla^2 J(x))^{-1} \nabla J_x,$$

and the *Newton decrement* $\lambda(x)$ is defined by

$$\lambda(x) = \left((\nabla J_x)^\top (\nabla^2 J(x))^{-1} \nabla J_x\right)^{1/2}.$$

Observe that

$$\langle \nabla J_x, d_{\mathrm{nt}} \rangle = (\nabla J_x)^\top (-(\nabla^2 J(x))^{-1} \nabla J_x) = -\lambda(x)^2.$$

If $\nabla J_x \neq 0$, we have $\lambda(x) \neq 0$, so $\langle \nabla J_x, d_{\rm nt} \rangle < 0$, and $d_{\rm nt}$ is indeed a descent direction. The number $\langle \nabla J_x, d_{\rm nt} \rangle$ is the constant that shows up during a backtracking line search.

A nice feature of the Newton step and of the Newton decrement is that they are affine invariant. This means that if T is an invertible matrix and if we define g by $g(y) = J(Ty)$, if the Newton step associated with J is denoted by $d_{J,\rm nt}$ and similarly the Newton step associated with g is denoted by $d_{g,\rm nt}$, then it is shown in Boyd and Vandenberghe [29] (Section 9.5.1) that

$$d_{g,\rm nt} = T^{-1} d_{J,\rm nt},$$

and so

$$x + d_{J,\rm nt} = T(y + d_{g,\rm nt}).$$

A similar properties applies to the Newton decrement.

Newton's method consists of the following steps: Given a starting point $u_0 \in \mathrm{dom}(J)$ and a tolerance $\epsilon > 0$ do:

repeat

(1) Compute the Newton step and decrement
$d_{{\rm nt},k} = -(\nabla^2 J(u_k))^{-1} \nabla J_{u_k}$ and $\lambda(u_k)^2 = (\nabla J_{u_k})^\top (\nabla^2 J(u_k))^{-1} \nabla J_{u_k}$.

(2) Stopping criterion. **quit** if $\lambda(u_k)^2/2 \leq \epsilon$.

(3) Line Search. Perform an exact or backtracking line search to find ρ_k.

(4) Update. $u_{k+1} = u_k + \rho_k d_{{\rm nt},k}$.

Observe that this is essentially the descent procedure of Section 49.8 using the Newton step as search direction, except that the stopping criterion is checked just after computing the search direction, rather than after the update (a very minor difference).

The convergence of Newton's method is thoroughly analyzed in Boyd and Vandenberghe [29] (Section 9.5.3). This analysis is made under the following assumptions:

(1) The function $J \colon \Omega \to \mathbb{R}$ is a convex function defined on some open subset Ω of $\mathbb{R}^n$ which is twice differentiable and its Hessian $\nabla^2 J(x)$ is symmetric positive definite for all $x \in \Omega$. This implies that there are two constants $m > 0$ and $M > 0$ such that $mI \preceq \nabla^2 J(x) \preceq MI$ for all $x \in \Omega$, which means that the eigenvalues of $\nabla^2 J(x)$ belong to $[m, M]$.

(2) The Hessian is Lipschitzian, which means that there is some $L \geq 0$ such that

$$\left\| \nabla^2 J(x) - \nabla^2 J(y) \right\|_2 \leq L \left\| x, y \right\|_2 \quad \text{for all } x, y \in \Omega.$$

It turns out that the iterations of Newton's method fall into two phases, depending whether $\left\| \nabla J_{u_k} \right\|_2 \geq \eta$ or $\left\| \nabla J_{u_k} \right\|_2 < \eta$, where η is a number which depends on m, L, and the constant α used in the backtracking line search, and $\eta \leq m^2/L$.

(1) The first phase, called the *damped Newton phase*, occurs while $\|\nabla J_{u_k}\|_2 \geq \eta$. During this phase, the procedure can choose a step size $\rho_k = t < 1$, and there is some constant $\gamma > 0$ such that

$$J(u_{k+1}) - J(u_k) \leq -\gamma.$$

(2) The second phase, called the *quadratically convergent phase* or *pure Newton phase*, occurs while $\|\nabla J_{u_k}\|_2 < \eta$. During this phase, the step size $\rho_k = t = 1$ is always chosen, and we have

$$\frac{L}{2m^2}\left\|\nabla J_{u_{k+1}}\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla J_{u_k}\right\|_2\right)^2. \tag{$*_1$}$$

If we denote the minimal value of f by p^*, then the number of damped Newton steps is at most

$$\frac{J(u_0) - p^*}{\gamma}.$$

Equation ($*_1$) and the fact that $\eta \leq m^2/L$ shows that if $\|\nabla J_{u_k}\|_2 < \eta$, then $\left\|\nabla J_{u_{k+1}}\right\|_2 < \eta$. It follows by induction that for all $\ell \geq k$, we have

$$\frac{L}{2m^2}\left\|\nabla J_{u_{\ell+1}}\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla J_{u_\ell}\right\|_2\right)^2, \tag{$*_2$}$$

and thus (since $\eta \leq m^2/L$ and $\|\nabla J_{u_k}\|_2 < \eta$, we have $(L/m^2)\left\|\nabla J_{u_k}\right\|_2 < (L/m^2)\eta \leq 1$), so

$$\frac{L}{2m^2}\left\|\nabla J_{u_\ell}\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla J_{u_k}\right\|_2\right)^{2^{\ell-k}} \leq \left(\frac{1}{2}\right)^{2^{\ell-k}}, \quad \ell \geq k. \tag{$*_3$}$$

It is shown in Boyd and Vandenberghe [29] (Section 9.1.2) that the hypothesis $mI \preceq \nabla^2 J(x)$ implies that

$$J(x) - p^* \leq \frac{1}{2m}\left\|\nabla J_x\right\|_2^2 \quad x \in \Omega.$$

As a consequence, by ($*_3$), we have

$$J(u_\ell) - p^* \leq \frac{1}{2m}\left\|\nabla J_{u_\ell}\right\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}}. \tag{$*_4$}$$

Equation ($*_4$) shows that the convergence during the quadratically convergence phase is very fast. If we let

$$\epsilon_0 = \frac{2m^3}{L^2},$$

then Equation ($*_4$) implies that we must have $J(u_\ell) - p^* \leq \epsilon$ after no more than

$$\log_2 \log_2(\epsilon_0/\epsilon)$$

iterations. The term $\log_2 \log_2(\epsilon_0/\epsilon)$ grows *extremely slowly* as ϵ goes to zero, and for practical purposes it can be considered constant, say five or six (six iterations gives an accuracy of about $\epsilon \approx 5 \cdot 10^{-20}\epsilon_0$).

In summary, the number of Newton iterations required to find a minimum of J is approximately bounded by

$$\frac{J(u_0) - p^*}{\gamma} + 6.$$

Examples of the application of Newton's method and further discussion of its efficiency are given in Boyd and Vandenberghe [29] (Section 9.5.4). Basically, Newton's method has a faster convergence rate than gradient or steepest descent. Its main disadvantage is the cost for forming and storing the Hessian, and of computing the Newton step, which requires solving a linear system.

There are two major shortcomings of the convergence analysis of Newton's method as sketched above. The first is a pracical one. The complexity estimates involve the constants m, M, and L, which are almost never known in practice. As a result, the bound on the number of steps required is almost never known specifically.

The second shortcoming is that although Newton's method itself is affine invariant, the analysis of convergence is very much dependent on the choice of coordinate system. If the coordinate system is changed, the constants m, M, L also change. This can be viewed as an aesthetic problem, but it would be nice if an analysis of convergence independent of an affine change of coordinates could be given.

Nesterov and Nemirovski discovered a condition on functions that allows an affine-invariant convergence analysis. This property, called *self-concordance*, is unfortunately not very intuitive.

Definition 49.10. A (partial) convex function f defined on $\mathbb{R}$ is *self-concordant* if

$$|f'''(x)| \leq 2(f''(x))^{3/2} \quad \text{for all } x \in \mathbb{R}.$$

A (partial) convex function f defined on $\mathbb{R}^n$ is *self-concordant* if for every nonzero $v \in \mathbb{R}^n$ and all $x \in \mathbb{R}^n$, the function $t \mapsto J(x + tv)$ is self-concordant.

Affine and convex quadratic functions are obviously self-concordant, since $f''' = 0$. There are many more interesting self-concordant functions, for example, the function $X \mapsto -\log\det(X)$, where X is a symmetric positive definite matrix.

Self-concordance is discussed extensively in Boyd and Vandenberghe [29] (Section 9.6). The main point of self-concordance is that a coordinate system-invariant proof of convergence can be given for a certain class of strictly convex self-concordant functions. This proof is given in Boyd and Vandenberghe [29] (Section 9.6.4). Given a starting value u_0, we assume that the sublevel set $\{x \in \mathbb{R}^n \mid J(x) \leq J(u_0)\}$ is closed and that J is bounded below. Then there are two parameters η and γ as before, but *depending only on the parameters α, β involved in the line search*, such that:

(1) If $\lambda(u_k) > \eta$, then
$$J(u_{k+1}) - J(u_k) \leq -\gamma.$$

(2) If $\lambda(u_k) \leq \eta$, then the backtracking line search selects $t = 1$ and we have
$$2\lambda(u_{k+1}) \leq (2\lambda(u_k))^2.$$

As a consequence, for all $\ell \geq k$, we have
$$J(u_\ell) - p^* \leq \lambda(u_\ell)^2 \leq \left(\frac{1}{2}\right)^{2^{\ell-k+1}}.$$

In the end, accuracy $\epsilon > 0$ is achieved in at most
$$\frac{20 - 8\alpha}{\alpha\beta(1-2\alpha)^2}(J(u_0) - p^*) + \log_2 \log_2(1/\epsilon)$$
iterations, where α and β are the constants involved in the line search. This bound is obviously independent of the chosen coordinate system.

Contrary to intuition, the descent direction $d_k = -\nabla J_{u_k}$ given by the opposite of the gradient is *not* always optimal. In the next section we will see how a better direction can be picked; this is the method of *conjugate gradients*.

49.10 Conjugate Gradient Methods for Unconstrained Problems

The conjugate gradient method due to Hestenes and Stiefel (1952) is a gradient descent method that applies to an elliptic quadratic functional $J\colon \mathbb{R}^n \to \mathbb{R}$ given by
$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$
where A is an $n \times n$ symmetric positive definite matrix. Although it is presented as an iterative method, it terminates in at most n steps.

As usual, the conjugate gradient method starts with some arbitrary initial vector u_0 and proceeds through a sequence of iteration steps generating (better and better) approximations u_k of the optimal vector u minimizing J. During an iteration step, two vectors need to be determined:

(1) The descent direction d_k.

(2) The next approximation u_{k+1}. To find u_{k+1}, we need to find the stepsize $\rho_k > 0$ and then
$$u_{k+1} = u_k - \rho_k d_k.$$
Typically, ρ_k is found by performing a line search along the direction d_k, namely we find ρ_k as the real number such that the function $\rho \mapsto J(u_k - \rho d_k)$ is minimized.

We saw in Proposition 49.13 that during execution of the gradient method with optimal stepsize parameter that any two consecutive descent directions are orthogonal. The new twist with the conjugate gradient method is that given $u_0, u_1, \ldots, u_k$, the next approximation u_{k+1} is obtained as the solution of the problem which consists in minimizing J over the affine subspace $u_k + \mathcal{G}_k$, where $\mathcal{G}_k$ is the subspace of $\mathbb{R}^n$ spanned by the gradients

$$\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_k}.$$

We may assume that $\nabla J_{u_\ell} \neq 0$ for $\ell = 0, \ldots, k$, since the method terminates as soon as $\nabla J_{u_k} = 0$. A priori the subspace $\mathcal{G}_k$ has dimension $\leq k+1$, but we will see that in fact it has dimension $k+1$. Then we have

$$u_k + \mathcal{G}_k = \left\{ u_k + \sum_{i=0}^{k} \alpha_i \nabla J_{u_i} \;\middle|\; \alpha_i \in \mathbb{R},\ 0 \leq i \leq k \right\},$$

and our minimization problem is to find u_{k+1} such that

$$u_{k+1} \in u_k + \mathcal{G}_k \quad \text{and} \quad J(u_{k+1}) = \inf_{v \in u_k + \mathcal{G}_k} J(v).$$

In the gradient method with optimal stepsize parameter the descent direction d_k is proportional to the gradient ∇J_{u_k}, but in the conjugate gradient method, d_k is equal to ∇J_{u_k} corrected by some multiple of d_{k-1}.

The conjugate gradient method is superior to the gradient method with optimal stepsize parameter for the following reasons proved correct later:

(a) The gradients ∇J_{u_i} and ∇J_{u_j} are orthogonal for all i, j with $0 \leq i \neq j \leq k$. This implies that if $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, then the vectors ∇J_{u_i} are linearly independent, so the method stops in at most n steps.

(b) If we write $\Delta_\ell = u_{\ell+1} - u_\ell = -\rho_\ell d_\ell$, the second remarkable fact about the conjugate gradient method is that the vectors Δ_ℓ satisfy the following conditions:

$$\langle A\Delta_\ell, \Delta_i \rangle = 0 \quad 0 \leq i < \ell \leq k.$$

The vectors Δ_ℓ and Δ_i are said to be *conjugate* with respect to the matrix A (or *A-conjugate*). As a consequence, if $\Delta_\ell \neq 0$ for $\ell = 0, \ldots, k$, then the vectors Δ_ℓ are linearly independent.

(c) There is a simple formula to compute d_{k+1} from d_k, and to compute ρ_k.

We now prove the above facts. We begin with (a).

Proposition 49.15. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$. Then the minimization problem, find u_{k+1} such that*

$$u_{k+1} \in u_k + \mathcal{G}_k \quad \textit{and} \quad J(u_{k+1}) = \inf_{v \in u_k + \mathcal{G}_k} J(v),$$

has a unique solution, and the gradients ∇J_{u_i} and ∇J_{u_j} are orthogonal for all i, j with $0 \leq i \neq j \leq k+1$.

Proof. The affine space $u_\ell + \mathcal{G}_\ell$ is closed and convex, and since J is a quadratic elliptic functional it is coercive and strictly convex, so by Theorem 49.8(2) it has a unique minimum in $u_\ell + \mathcal{G}_\ell$. This minimum $u_{\ell+1}$ is also the minimum of the problem, find $u_{\ell+1}$ such that

$$u_{\ell+1} \in u_\ell + \mathcal{G}_\ell \quad \text{and} \quad J(u_{\ell+1}) = \inf_{v \in \mathcal{G}_\ell} J(u_\ell + v),$$

and since $\mathcal{G}_\ell$ is a subspace, by Corollary 40.10 we must have

$$dJ_{u_{\ell+1}}(w) = 0 \quad \text{for all } w \in \mathcal{G}_\ell,$$

that is

$$\langle \nabla J_{u_{\ell+1}}, w \rangle = 0 \quad \text{for all } w \in \mathcal{G}_\ell.$$

Since $\mathcal{G}_\ell$ is spanned by $(\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_\ell})$, we obtain

$$\langle \nabla J_{u_{\ell+1}}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq j \leq \ell,$$

and since this holds for $\ell = 0, \ldots, k$, we get

$$\langle \nabla J_{u_i}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq i \neq j \leq k+1,$$

which shows the second part of the proposition. ∎

As a corollary of Proposition 49.15, if $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, then the vectors ∇J_{u_i} are linearly independent and $\mathcal{G}_k$ has dimension $k+1$. Therefore, the conjugate gradient method terminates in at most n steps. Here is an example of a problem for which the gradient descent with optimal stepsize parameter does not converge in a finite number of steps.

Example 49.2. Let $J \colon \mathbb{R}^2 \to \mathbb{R}$ be the function given by

$$J(v_1, v_2) = \frac{1}{2}(\alpha_1 v_1^2 + \alpha_2 v_2^2),$$

where $0 < \alpha_1 < \alpha_2$. The minimum of J is attained at $(0,0)$. Unless the initial vector $u_0 = (u_1^0, u_2^0)$ has the property that either $u_1^0 = 0$ or $u_2^0 = 0$, we claim that the gradient descent with optimal stepsize parameter does not converge in a finite number of steps. Observe that

$$\nabla J_{(v_1,v_2)} = \begin{pmatrix} \alpha_1 v_1 \\ \alpha_2 v_2 \end{pmatrix}.$$

As a consequence, given u_k, the line search for finding ρ_k and u_{k+1} yields $u_{k+1} = (0,0)$ iff there is some $\rho \in \mathbb{R}$ such that

$$u_1^k = \rho \alpha_1 u_1^k \quad \text{and} \quad u_2^k = \rho \alpha_2 u_2^k.$$

Since $\alpha_1 \neq \alpha_2$, this is only possible if either $u_1^k = 0$ or $u_2^k = 0$. The formulae given just before Proposition 49.14 yield

$$u_1^{k+1} = \frac{\alpha_2^2(\alpha_2 - \alpha_1) u_1^k (u_2^k)^2}{\alpha_1^3 (u_1^k)^2 + \alpha_2^3 (u_2^k)^2}, \quad u_2^{k+1} = \frac{\alpha_1^2(\alpha_1 - \alpha_2) u_2^k (u_1^k)^2}{\alpha_1^3 (u_1^k)^2 + \alpha_2^3 (u_2^k)^2},$$

which implies that if $u_1^k \neq 0$ and $u_2^k \neq 0$, then $u_1^{k+1} \neq 0$ and $u_2^{k+1} \neq 0$, so the method runs forever from any initial vector $u_0 = (u_1^0, u_2^0)$ such that $u_1^0 \neq 0$ and, $u_2^0 \neq 0$.

We now prove (b).

Proposition 49.16. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, and let $\Delta_\ell = u_{\ell+1} - u_\ell$, for $\ell = 0, \ldots, k$. Then $\Delta_\ell \neq 0$ for $\ell = 0, \ldots, k$, and*

$$\langle A\Delta_\ell, \Delta_i \rangle = 0, \quad 0 \leq i < \ell \leq k.$$

The vectors $\Delta_0, \ldots, \Delta_k$ are linearly independent.

Proof. Since J is a quadratic functional we have

$$\nabla J_{v+w} = A(v+w) - b = Av - b + Aw = \nabla J_v + Aw.$$

It follows that

$$\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell + \Delta_\ell} = \nabla J_{u_\ell} + A\Delta_\ell, \quad 0 \leq \ell \leq k. \tag{$*_1$}$$

By Proposition 49.15, since

$$\langle \nabla J_{u_i}, \nabla J_{u_j} \rangle = 0, \quad 0 \leq i \neq j \leq k,$$

we get

$$0 = \langle \nabla J_{u_\ell+1}, \nabla J_{u_\ell} \rangle = \left\| \nabla J_{u_\ell} \right\|^2 + \langle A\Delta_\ell, \nabla J_{u_\ell} \rangle, \quad \ell = 0, \ldots, k,$$

and since by hypothesis $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$, we deduce that

$$\Delta_\ell \neq 0, \quad \ell = 0, \ldots, k.$$

If $k \geq 1$, for $i = 0, \ldots, \ell - 1$ and $\ell \leq k$ we also have

$$\begin{aligned} 0 = \langle \nabla J_{u_{\ell+1}}, \nabla J_{u_i} \rangle &= \langle \nabla J_{u_\ell}, \nabla J_{u_i} \rangle + \langle A\Delta_\ell, \nabla J_{u_i} \rangle \\ &= \langle A\Delta_\ell, \nabla J_{u_i} \rangle. \end{aligned}$$

Since $\Delta_j = u_{j+1} - u_j \in \mathcal{G}_j$ and $\mathcal{G}_j$ is spanned by $(\nabla J_{u_0}, \nabla J_{u_1}, \ldots, \nabla J_{u_j})$, we obtain

$$\langle A\Delta_\ell, \Delta_j \rangle = 0, \quad 0 \leq j < \ell \leq k.$$

For the last statement of the proposition, let $w_0, w_1, \ldots, w_k$ be any $k+1$ nonzero vectors such that

$$\langle Aw_i, w_j \rangle = 0, \quad 0 \leq i < j \leq k.$$

We claim that $w_0, w_1, \ldots, w_k$ are linearly independent.

If we have a linear dependence $\sum_{i=0}^k \lambda_i w_i = 0$, then we have

$$0 = \left\langle A\left(\sum_{i=0}^k \lambda_i w_i\right), w_j \right\rangle = \sum_{i=0}^k \lambda_i \langle Aw_i, w_j \rangle = \lambda_j \langle Aw_j, w_j \rangle,$$

where we form these inner products for $j = 0, \ldots, k$, in that order. Since A is symmetric positive definite (because J is a quadratic elliptic functional) and $w_j \neq 0$, we have $\langle Aw_j, w_j \rangle > 0$, and so $\lambda_j = 0$ for $j = 0, \ldots, k$. Therefore the vectors $w_0, w_1, \ldots, w_k$ are linearly independent. $\square$

Remarks:

(1) Since A is symmetric positive definite, the bilinear map $(u, v) \mapsto \langle Au, v\rangle$ is an inner product $\langle -, -\rangle_A$ on $\mathbb{R}^n$. Consequently, two vectors u, v are *conjugate* with respect to the matrix A (or A-*conjugate*), which means that $\langle Au, v\rangle = 0$, iff u and v are orthogonal with respect to the inner product $\langle -, -\rangle_A$.

(2) By picking the descent direction to be $-\nabla J_{u_k}$, the gradient descent method with optimal stepsize parameter treats the level sets $\{u \mid J(u) = J(u_k)\}$ as if they were spheres. The conjugate gradient method is more subtle, and takes the "geometry" of the level set $\{u \mid J(u) = J(u_k)\}$ into account, through the notion of conjugate directions.

(3) The notion of conjugate direction has its origins in the theory of projective conics and quadrics where A is a 2×2 or a 3×3 matrix and where u and v are conjugate iff $u^\top Av = 0$.

(4) The terminology conjugate gradient is somewhat misleading. It is not the gradients who are conjugate directions, but the descent directions.

By definition of the vectors $\Delta_\ell = u_{\ell+1} - u_\ell$, we can write

$$\Delta_\ell = \sum_{i=0}^{\ell} \delta_i^\ell \, \nabla J_{u_i}, \quad 0 \leq \ell \leq k. \tag{$*_2$}$$

In matrix form, we can write

$$\begin{pmatrix} \Delta_0 & \Delta_1 & \cdots & \Delta_k \end{pmatrix} = \begin{pmatrix} \nabla J_{u_0} & \nabla J_{u_1} & \cdots & \nabla J_{u_k} \end{pmatrix} \begin{pmatrix} \delta_0^0 & \delta_0^1 & \cdots & \delta_0^{k-1} & \delta_0^k \\ 0 & \delta_1^1 & \cdots & \delta_1^{k-1} & \delta_1^k \\ 0 & 0 & \cdots & \delta_2^{k-1} & \delta_2^k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \delta_k^k \end{pmatrix},$$

which implies that $\delta_\ell^\ell \neq 0$ for $\ell = 0, \ldots, k$.

In view of the above fact, since Δ_ℓ and d_ℓ are collinear, it is convenient to write the descent direction d_ℓ as

$$d_\ell = \sum_{i=0}^{\ell-1} \lambda_i^\ell \, \nabla J_{u_i} + \nabla J_{u_\ell}, \quad 0 \leq \ell \leq k. \tag{$*_3$}$$

Our next goal is to compute u_{k+1}, assuming that the coefficients λ_i^k are known for $i = 0, \ldots, k$, and then to find simple formulae for the λ_i^k.

The problem reduces to finding ρ_k such that

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k),$$

and then $u_{k+1} = u_k - \rho_k d_k$. In fact, by $(*_2)$, since

$$\Delta_k = \sum_{i=0}^{k} \delta_i^k \, \nabla J_{u_i} = \delta_k^k \left(\sum_{i=0}^{k-1} \frac{\delta_i^k}{\delta_k^k} \, \nabla J_{u_i} + \nabla J_{u_k} \right),$$

we must have

$$\Delta_k = \delta_k^k d_k \quad \text{and} \quad \rho_k = -\delta_k^k. \tag{$*_4$}$$

Remarkably, the coefficients λ_i^k and the descent directions d_k can be computed easily using the following formulae.

Proposition 49.17. *Assume that $\nabla J_{u_i} \neq 0$ for $i = 0, \ldots, k$. If we write*

$$d_\ell = \sum_{i=0}^{\ell-1} \lambda_i^\ell \, \nabla J_{u_i} + \nabla J_{u_\ell}, \quad 0 \leq \ell \leq k,$$

then we have

$$(\dagger) \quad \begin{cases} \lambda_i^k = \dfrac{\left\| \nabla J_{u_k} \right\|^2}{\left\| \nabla J_{u_i} \right\|^2}, \quad 0 \leq i \leq k-1, \\ d_0 = \nabla J_{u_0} \\ d_\ell = \nabla J_{u_\ell} + \dfrac{\left\| \nabla J_{u_\ell} \right\|^2}{\left\| \nabla J_{u_{\ell-1}} \right\|^2} \, d_{\ell-1}, \quad 1 \leq \ell \leq k. \end{cases}$$

Proof. Since by $(*_4)$ we have $\Delta_k = \delta_k^k d_k$, $\delta_k^k \neq 0$, (by Proposition 49.16) we have

$$\langle A d_\ell, \Delta_i \rangle = 0, \quad 0 \leq i < \ell \leq k.$$

By $(*_1)$ we have $\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell} + A\Delta_\ell$, and since A is a symmetric matrix, we have

$$0 = \langle A d_k, \Delta_\ell \rangle = \langle d_k, A\Delta_\ell \rangle = \langle d_k, \nabla J_{u_{\ell+1}} - \nabla J_{u_\ell} \rangle,$$

for $\ell = 0, \ldots, k-1$. Since

$$d_k = \sum_{i=0}^{k-1} \lambda_i^k \, \nabla J_{u_i} + \nabla J_{u_k},$$

we have

$$\left\langle \sum_{i=0}^{k-1} \lambda_i^k \, \nabla J_{u_i} + \nabla J_{u_k}, \nabla J_{u_{\ell+1}} - \nabla J_{u_\ell} \right\rangle = 0, \quad 0 \leq \ell \leq k-1.$$

Since by Proposition 49.15 the gradients ∇J_{u_i} are pairwise orthogonal, the above equations yield

$$\begin{aligned} -\lambda_{k-1}^k \left\| \nabla J_{u_{k-1}} \right\|^2 + \left\| \nabla J_{u_k} \right\|^2 &= 0 \\ -\lambda_\ell^k \left\| \nabla J_{u_\ell} \right\|^2 + \lambda_{\ell+1}^k \left\| \nabla J_{u_{\ell+1}} \right\|^2 &= 0, \quad 0 \leq \ell \leq k-2, \quad k \geq 2, \end{aligned}$$

and an easy induction yields

$$\lambda_i^k = \frac{\|\nabla J_{u_k}\|^2}{\|\nabla J_{u_i}\|^2}, \quad 0 \leq i \leq k-1.$$

Consequently, using $(*_3)$ we have

$$\begin{aligned}
d_k &= \sum_{i=0}^{k-1} \frac{\|\nabla J_{u_k}\|^2}{\|\nabla J_{u_i}\|^2} \nabla J_{u_i} + \nabla J_{u_k} \\
&= \nabla J_{u_k} + \frac{\|\nabla J_{u_k}\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2} \left(\sum_{i=0}^{k-2} \frac{\left\|\nabla J_{u_{k-1}}\right\|^2}{\|\nabla J_{u_i}\|^2} \nabla J_{u_i} + \nabla J_{u_{k-1}} \right) \\
&= \nabla J_{u_k} + \frac{\|\nabla J_{u_k}\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2} d_{k-1},
\end{aligned}$$

which concludes the proof. $\square$

It remains to compute ρ_k, which is the solution of the line search

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k).$$

Since J is a quadratic functional, a basic computation left to the reader shows that the function to be minimized is

$$\rho \mapsto \frac{\rho^2}{2} \langle A d_k, d_k \rangle - \rho \langle \nabla J_{u_k}, d_k \rangle + J(u_k),$$

whose mininum is obtained when its derivative is zero, that is,

$$\rho_k = \frac{\langle \nabla J_{u_k}, d_k \rangle}{\langle A d_k, d_k \rangle}. \tag{$*_5$}$$

In summary, the conjugate gradient method finds the minimum u of the elliptic quadratic functional

$$J(v) = \frac{1}{2} \langle Av, v \rangle - \langle b, v \rangle$$

by computing the sequence of vectors $u_1, d_1, \ldots, u_{k-1}, d_{k-1}, u_k$, starting from any vector u_0, with

$$d_0 = \nabla J_{u_0}.$$

If $\nabla J_{u_0} = 0$, then the algorithm terminates with $u = u_0$. Otherwise, for $k \geq 0$, assuming that $\nabla J_{u_i} \neq 0$ for $i = 1, \ldots, k$, compute

$$(*_6) \quad \begin{cases} \rho_k = \dfrac{\langle \nabla J_{u_k}, d_k \rangle}{\langle A d_k, d_k \rangle} \\ u_{k+1} = u_k - \rho_k d_k \\ d_{k+1} = \nabla J_{u_{k+1}} + \dfrac{\left\|\nabla J_{u_{k+1}}\right\|^2}{\|\nabla J_{u_k}\|^2} d_k. \end{cases}$$

If $\nabla J_{u_{k+1}} = 0$, then the algorihm terminates with $u = u_{k+1}$.

As we showed before, the algorithm terminates in at most n iterations.

Example 49.3. Let us take the example of Section 49.6 and apply the conjugate gradient procedure. Recall that

$$\begin{aligned} J(x,y) &= \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 2 & -8 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \\ &= \frac{3}{2}x^2 + 2xy + 3y^2 - 2x + 8y. \end{aligned}$$

Note that $\nabla J_v = (3x + 2y - 2, 2x + 6y + 8)$,
Initialize the procedure by setting

$$u_0 = (-2,-2), \qquad d_0 = \nabla J_{u_0} = (-12,-8)$$

Step 1 involves calculating

$$\begin{aligned} \rho_0 &= \frac{\langle \nabla J_{u_0}, d_0\rangle}{\langle Ad_0, d_0\rangle} = \frac{13}{75} \\ u_1 &= u_0 - \rho_0 d_0 = (-2,-2) - \frac{13}{75}(-12,-8) = \left(\frac{2}{25}, -\frac{46}{75}\right) \\ d_1 &= \nabla J_{u_1} + \frac{||\nabla J_{u_1}||^2}{||\nabla J_{u_0}||^2} d_0 = \left(-\frac{2912}{625}, \frac{18928}{5625}\right). \end{aligned}$$

Observe that ρ_0 and u_1 are precisely the *same* as in the case the case of gradient descent with optimal step size parameter. The difference lies in the calculation of d_1. As we will see, this change will make a *huge* difference in the convergence to the unique minimum $u = (2, -2)$.

We continue with the conjugate gradient procedure and calculate Step 2 as

$$\begin{aligned} \rho_1 &= \frac{\langle \nabla J_{u_1}, d_1\rangle}{\langle Ad_1, d_1\rangle} = \frac{75}{82} \\ u_2 &= u_1 - \rho_1 d_1 = \left(\frac{2}{25}, -\frac{46}{75}\right) - \frac{75}{82}\left(-\frac{2912}{625}, \frac{18928}{5625}\right) = (2,-2) \\ d_2 &= \nabla J_{u_2} + \frac{||\nabla J_{u_2}||^2}{||\nabla J_{u_1}||^2} d_1 = (0,0). \end{aligned}$$

Since $\nabla J_{u_2} = 0$, the procedure terminates in *two* steps, as opposed to the 31 steps needed for gradient descent with optimal step size parameter.

Hestenes and Stiefel realized that Equations $(*_6)$ can be modified to make the computations more efficient, by having only one evaluation of the matrix A on a vector, namely d_k. The idea is to compute ∇_{u_k} inductively.

Since by $(*_1)$ and $(*_4)$ we have $\nabla J_{u_{\ell+1}} = \nabla J_{u_\ell} + A\Delta_\ell = \nabla J_{u_\ell} - \rho_\ell A d_\ell$, the gradient $\nabla J_{u_{\ell+1}}$ can be computed iteratively:

$$\begin{aligned}\nabla J_0 &= Au_0 - b\\ \nabla J_{u_{\ell+1}} &= \nabla J_{u_\ell} - \rho_\ell A d_\ell.\end{aligned}$$

Since by Proposition 49.17 we have

$$d_k = \nabla J_{u_k} + \frac{\left\|\nabla J_{u_k}\right\|^2}{\left\|\nabla J_{u_{k-1}}\right\|^2}\, d_{k-1}$$

and since d_{k-1} is a linear combination of the gradients ∇J_{u_i} for $i = 0, \ldots, k-1$, which are all orthogonal to ∇J_{u_k}, we have

$$\rho_k = \frac{\langle \nabla J_{u_k}, d_k\rangle}{\langle A d_k, d_k\rangle} = \frac{\left\|\nabla J_{u_k}\right\|^2}{\langle A d_k, d_k\rangle}.$$

It is customary to introduce the term r_k defined as

$$r_k = \nabla J_{u_k} = Au_k - b \tag{$*_7$}$$

and to call it the *residual*. Then the conjugate gradient method consists of the following steps. We intitialize the method starting from any vector u_0 and set

$$d_0 = r_0 = Au_0 - b.$$

The main iteration step is $(k \geq 0)$:

$$(*_8) \qquad \begin{cases} \rho_k = \dfrac{\left\|r_k\right\|^2}{\langle A d_k, d_k\rangle}\\ u_{k+1} = u_k - \rho_k d_k\\ r_{k+1} = r_k - \rho_k A d_k\\ \beta_{k+1} = \dfrac{\left\|r_{k+1}\right\|^2}{\left\|r_k\right\|^2}\\ d_{k+1} = r_{k+1} + \beta_{k+1}\, d_k. \end{cases}$$

Beware that some authors define the residual r_k as $r_k = b - Au_k$ and the descent direction d_k as $-d_k$. In this case, the second equation becomes

$$u_{k+1} = u_k + \rho_k d_k.$$

Since $d_0 = r_0$, the equations

$$\begin{aligned} r_{k+1} &= r_k - \rho_k A d_k\\ d_{k+1} &= r_{k+1} + \beta_{k+1} d_k\end{aligned}$$

imply by induction that the subspace $\mathcal{G}_k$ is spanned by $(r_0, r_1, \ldots, r_k)$ and $(d_0, d_1, \ldots, d_k)$ is the subspace spanned by

$$(r_0, Ar_0, A^2r_0, \ldots, A^kr_0).$$

Such a subspace is called a *Krylov subspace*.

If we define the *error* e_k as $e_k = u_k - u$, then $e_0 = u_0 - u$ and $Ae_0 = Au_0 - Au = Au_0 - b = d_0 = r_0$, and then because

$$u_{k+1} = u_k - \rho_k d_k$$

we see that

$$e_{k+1} = e_k - \rho_k d_k.$$

Since d_k belongs to the subspace spanned by $(r_0, Ar_0, A^2r_0, \ldots, A^kr_0)$ and $r_0 = Ae_0$, we see that d_k belongs to the subspace spanned by $(Ae_0, A^2e_0, A^3e_0, \ldots, A^{k+1}e_0)$, and then by induction we see that e_{k+1} belongs to the subspace spanned by $(e_0, Ae_0, A^2e_0, A^3e_0, \ldots, A^{k+1}e_0)$. This means that there is a polynomial P_k of degree $\leq k$ such that $P_k(0) = 1$ and

$$e_k = P_k(A)e_0.$$

This is an important fact because it allows for an analysis of the convergence of the conjugate gradient method; see Trefethen and Bau [176] (Lecture 38). For this, since A is symmetric positive definite, we know that $\langle u, v\rangle_A = \langle Av, u\rangle$ is an inner product on $\mathbb{R}^n$ whose associated norm is denoted by $\|v\|_A$. Then observe that if $e(v) = v - u$, then

$$\begin{aligned}
\|e(v)\|_A^2 &= \langle Av - Au, v - u\rangle \\
&= \langle Av, v\rangle - 2\langle Au, v\rangle + \langle Au, u\rangle \\
&= \langle Av, v\rangle - 2\langle b, v\rangle + \langle b, u\rangle \\
&= 2J(v) + \langle b, u\rangle.
\end{aligned}$$

It follows that $v = u_k$ minimizes $\|e(v)\|_A$ on $u_{k-1} + \mathcal{G}_{k-1}$ since u_k minimizes J on $u_{k-1} + \mathcal{G}_{k-1}$. Since $e_k = P_k(A)e_0$ for some polynomial P_k of degree $\leq k$ such that $P_k(0) = 1$, if we let $\mathcal{P}_k$ be the set of polynomials $P(t)$ of degree $\leq k$ such that $P(0) = 1$, then we have

$$\|e_k\|_A = \inf_{P \in \mathcal{P}_k} \|P(A)e_0\|_A.$$

Since A is a symmetric positive definite matrix it has real positive eigenvalues $\lambda_1, \ldots, \lambda_n$ and there is an orthonormal basis of eigenvectors $h_1, \ldots, h_n$ so that if we write $e_0 = \sum_{j=1}^n a_j h_j$. then we have

$$\|e_0\|_A^2 = \langle Ae_0, e_0\rangle = \left\langle \sum_{i=1}^n a_i\lambda_i h_i, \sum_{j=1}^n a_j h_j \right\rangle = \sum_{j=1}^n a_j^2 \lambda_j$$

and

$$\|P(A)e_0\|_A^2 = \langle AP(A)e_0, P(A)e_0\rangle = \left\langle \sum_{i=1}^n a_i\lambda_i P(\lambda_i) h_i, \sum_{j=1}^n a_j P(\lambda_j) h_j \right\rangle = \sum_{j=1}^n a_j^2 \lambda_j (P(\lambda_j))^2.$$

These equations imply that

$$\|e_k\|_A \leq \left(\inf_{P \in \mathcal{P}_k} \max_{1 \leq i \leq n} |P(\lambda_i)| \right) \|e_0\|_A .$$

It can be shown that the conjugate gradient method requires of the order of

n^3 additions,

n^3 multiplications,

$2n$ divisions.

In theory, this is worse than the number of elementary operations required by the Cholesky method. Even though the conjugate gradient method does not seem to be the best method for *full* matrices, it usually outperforms other methods for *sparse* matrices. The reason is that the matrix A only appears in the computation of the vector Ad_k. If the matrix A is banded (for example, tridiagonal), computing Ad_k is very cheap and there is no need to store the entire matrix A, in which case the conjugate gradient method is fast. Also, although in theory, up to n iterations may be required, in practice, convergence may occur after a much smaller number of iterations.

Using the inequality

$$\|e_k\|_A \leq \left(\inf_{P \in \mathcal{P}_k} \max_{1 \leq i \leq n} |P(\lambda_i)| \right) \|e_0\|_A ,$$

by choosing P to be a shifted Chebyshev polynomial, it can be shown that

$$\|e_k\|_A \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|e_0\|_A ,$$

where $\kappa = \mathrm{cond}_2(A)$; see Trefethen and Bau [176] (Lecture 38, Theorem 38.5). Thus the rate of convergence of the conjugate gradient method is governed by the ratio

$$\frac{\sqrt{\mathrm{cond}_2(A)} - 1}{\sqrt{\mathrm{cond}_2(A)} + 1},$$

where $\mathrm{cond}_2(A) = \lambda_n/\lambda_1$ is the condition number of the matrix A. Since A is positive definite, λ_1 is its smallest eigenvalue and λ_n is its largest eigenvalue.

The above fact leads to the process of *preconditioning*, a method which consists in replacing the matrix of a linear system $Ax = b$ by an "equivalent" one for example $M^{-1}A$ (since M is invertible, the system $Ax = b$ is equivalent to the system $M^{-1}Ax = M^{-1}b$), where M is chosen so that $M^{-1}A$ is still symmetric positive definite and has a smaller condition number than A; see Trefethen and Bau [176] (Lecture 40) and Demmel [48] (Section 6.6.5).

The method of conjugate gradients can be generalized to functionals that are not necessarily quadratic. The stepsize parameter ρ_k is still determined by a line search which consists in finding ρ_k such that

$$J(u_k - \rho_k d_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho d_k).$$

This is more difficult than in the quadratic case and in general there is no guarantee that ρ_k is unique, so some criterion to pick ρ_k is needed. Then

$$u_{k+1} = u_k - \rho_k d_k,$$

and the next descent direction can be chosen in two ways:

(1) (*Polak–Ribière*)

$$d_k = \nabla J_{u_k} + \frac{\langle \nabla J_{u_k}, \nabla J_{u_k} - \nabla J_{u_{k-1}} \rangle}{\left\| \nabla J_{u_{k-1}} \right\|^2} d_{k-1},$$

(2) (*Fletcher–Reeves*)

$$d_k = \nabla J_{u_k} + \frac{\left\| \nabla J_{u_k} \right\|^2}{\left\| \nabla J_{u_{k-1}} \right\|^2} d_{k-1}.$$

Consecutive gradients are no longer orthogonal so these methods may run forever. There are various sufficient criteria for convergence. In practice, the Polak–Ribière method converges faster. There is no longer any guarantee that these methods converge to a global minimum.

49.11 Gradient Projection Methods for Constrained Optimization

We now consider the problem of finding the minimum of a convex functional $J\colon V \to \mathbb{R}$ over a nonempty, convex, closed subset U of a Hilbert space V. By Theorem 40.13(3), the functional J has a minimum at $u \in U$ iff

$$dJ_u(v - u) \geq 0 \quad \text{for all } v \in U,$$

which can be expressed as

$$\langle \nabla J_u, v - u \rangle \geq 0 \quad \text{for all } v \in U.$$

On the other hand, by the projection lemma (Proposition 48.5), the condition for a vector $u \in U$ to be the projection of an element $w \in V$ onto U is

$$\langle u - w, v - u \rangle \geq 0 \quad \text{for all } v \in U.$$

These conditions are obviously analogous, and we can make this analogy more precise as follows. If $p_U\colon V \to U$ is the projection map onto U, we have the following chain of equivalences:

$$\begin{aligned}
&u \in U \quad \text{and} \quad J(u) = \inf_{v\in U} J(v) \quad \text{iff} \\
&u \in U \quad \text{and} \quad \langle \nabla J_u, v - u\rangle \geq 0 \quad \text{for every } v \in U, \text{ iff} \\
&u \in U \quad \text{and} \quad \langle u - (u - \rho\nabla J_u), v - u\rangle \geq 0 \quad \text{for every } v \in U \text{ and every } \rho > 0, \text{ iff} \\
&u = p_U(u - \rho\nabla J_u) \quad \text{for every } \rho > 0.
\end{aligned}$$

In other words, for every $\rho > 0$, $u \in V$ is a *fixed-point* of the function $g\colon V \to U$ given by

$$g(v) = p_U(v - \rho\nabla J_v).$$

The above suggests finding u by the method of successive approximations for finding the fixed-point of a contracting mapping, namely given any initial $u_0 \in V$, to define the sequence $(u_k)_{k\geq 0}$ such that

$$u_{k+1} = p_U(u_k - \rho_k\nabla J_{u_k}),$$

where the parameter $\rho_k > 0$ is chosen at each step. This method is called the *projected-gradient method with variable stepsize parameter*. Observe that if $U = V$, then this is just the gradient method with variable stepsize. We have the following result about the convergence of this method.

Proposition 49.18. *Let $J\colon V \to \mathbb{R}$ be a continuously differentiable functional defined on a Hilbert space V, and let U be nonempty, convex, closed subset of V. Suppose there exists two constants $\alpha > 0$ and $M > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \left\|v - u\right\|^2 \quad \textit{for all } u, v \in V,$$

and

$$\left\|\nabla J_v - \nabla J_u\right\| \leq M \left\|v - u\right\| \quad \textit{for all } u, v \in V.$$

If there exists two real nunbers $a, b \in \mathbb{R}$ such that

$$0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2} \quad \textit{for all } k \geq 0,$$

then the projected-gradient method with variable stepsize parameter converges. Furthermore, there is some constant $\beta > 0$ (depending on α, M, a, b) such that

$$\beta < 1 \quad \textit{and} \quad \left\|u_k - u\right\| \leq \beta^k \left\|u_0 - u\right\|,$$

where $u \in M$ is the unique minimum of J.

Proof. For every $\rho_k \geq 0$, define the function $g_k \colon V \to U$ by

$$g_k(v) = p_U(v - \rho_k \nabla J_v).$$

By Proposition 48.6, the projection map p_U has Lipschitz constant 1, so using the inequalities assumed to hold in the proposition, we have

$$\begin{aligned}
\|g_k(v_1) - g_k(v_2)\|^2 &= \|p_U(v_1 - \rho_k \nabla J_{v_1}) - p_U(v_2 - \rho_k \nabla J_{v_2})\|^2 \\
&\leq \|(v_1 - v_2) - \rho_k(\nabla J_{v_1} - \nabla J_{v_2})\|^2 \\
&= \|v_1 - v_2\|^2 - 2\rho_k \langle \nabla J_{v_1} - \nabla J_{v_2}, v_1 - v_2 \rangle + \rho_k^2 \|\nabla J_{v_1} - \nabla J_{v_2}\|^2 \\
&\leq \left(1 - 2\alpha\rho_k + M^2\rho_k^2\right) \|v_1 - v_2\|^2 .
\end{aligned}$$

As in the proof of Proposition 49.14, we know that if a and b satisfy the conditions $0 < a \leq \rho_k \leq b \leq \frac{2\alpha}{M^2}$, then there is some β such that

$$\left(1 - 2\alpha\rho_k + M^2\rho_k^2\right)^{1/2} \leq \beta < 1 \quad \text{for all } k \geq 0.$$

Since the minimizing point $u \in U$ is a fixed point of g_k for all k, by letting $v_1 = u_k$ and $v_2 = u$, we get

$$\|u_{k+1} - u\| = \|g_k(u_k) - g_k(u)\| \leq \beta \|u_k - u\| ,$$

which proves the convergence of the sequence $(u_k)_{k \geq 0}$. ☐

In the case of an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, a \rangle - \langle b, v \rangle$$

defined on $\mathbb{R}^n$, the reasoning just after the proof of Proposition 49.14 can be immediately adapted to show that convergence takes place as long as a, b and ρ_k are chosen such that

$$0 < a \leq \rho_k \leq b \leq \frac{2}{\lambda_n}.$$

In theory, Proposition 49.18 gives a guarantee of the convergence of the projected-gradient method. Unfortunately, because computing the projection $p_U(v)$ effectively is generally impossible, the range of practical applications of Proposition 49.18 is rather limited. One exception is the case where U is a product $\prod_{i=1}^m [a_i, b_i]$ of closed intervals (where $a_i = -\infty$ or $b_i = +\infty$ is possible). In this case, it is not hard to show that

$$p_U(w)_i = \begin{cases} a_i & \text{if } w_i < a_i \\ w_i & \text{if } a_i \leq w_i \leq b_i \\ b_i & \text{if } b_i < w_i. \end{cases}$$

In particular, this is the case if

$$U = \mathbb{R}^n_+ = \{v \in \mathbb{R}^n \mid v \geq 0\}$$

and if

$$J(v) = \frac{1}{2}\langle Av, a\rangle - \langle b, v\rangle$$

is an elliptic quadratic functional on $\mathbb{R}^n$. Then the vector $u_{k+1} = (u_1^{k+1}, \ldots, u_n^{k+1})$ is given in terms of $u_k = (u_1^k, \ldots, u_n^k)$ by

$$u_i^{k+1} = \max\{u_i^k - \rho_k(Au_k - b)_i, 0\}, \quad 1 \leq i \leq n.$$

49.12 Penalty Methods for Constrained Optimization

In the case where $V = \mathbb{R}^n$, another method to deal with constrained optimization is to incorporate the domain U into the objective function J by adding a penalty function.

Definition 49.11. Given a nonempty closed convex subset U of $\mathbb{R}^n$, a function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is called a *penalty function* for U if ψ is convex and continuous and if the following conditions hold:

$$\psi(v) \geq 0 \quad \text{for all} \quad v \in \mathbb{R}^n, \quad \text{and} \quad \psi(v) = 0 \quad \text{iff} \quad v \in U.$$

The following proposition shows that the use of penalty functions reduces a constrained optimization problem to a sequence of unconstrained optimization problems.

Proposition 49.19. *Let $J\colon \mathbb{R}^n \to \mathbb{R}$ be a continuous, coercive, strictly convex function, U be a nonempty, convex, closed subset of $\mathbb{R}^n$, $\psi\colon \mathbb{R}^n \to \mathbb{R}$ be a penalty function for U, and let $J_\epsilon\colon \mathbb{R}^n \to \mathbb{R}$ be the penalized objective function given by*

$$J_\epsilon(v) = J(v) + \frac{1}{\epsilon}\,\psi(v) \quad \textit{for all } v \in \mathbb{R}^n.$$

Then for every $\epsilon > 0$, there exists a unique element $u_\epsilon \in \mathbb{R}^n$ such that

$$J_\epsilon(u_\epsilon) = \inf_{v\in\mathbb{R}^n} J_\epsilon(v).$$

Furthermore, if $u \in U$ is the unique minimizer of J over U, so that $J(u) = \inf_{v\in U} J(v)$, then

$$\lim_{\epsilon\mapsto 0} u_\epsilon = u.$$

Proof. Observe that since J is coercive, since $\psi(v) \geq 0$ for all $v \in \mathbb{R}^n$, and $J_\epsilon = J + (1/\epsilon)\psi$, we have $J_\epsilon(v) \geq J(v)$ for all $v \in \mathbb{R}^n$, so J_ϵ is also coercive. Since J is strictly convex and $(1/\epsilon)\psi$ is convex, it is immediately checked that $J_\epsilon = J + (1/\epsilon)\psi$ is also strictly convex. Then by Proposition 49.1 (and the fact that J and J_ϵ are strictly convex), J has a unique minimizer $u \in U$, and J_ϵ has a unique minimizer $u_\epsilon \in \mathbb{R}^n$.

Since $\psi(u) = 0$ iff $u \in U$, and $\psi(v) \geq 0$ for all $v \in \mathbb{R}^n$, we have $J_\epsilon(u) = J(u)$, and since u_ϵ is the minimizer of J_ϵ we have $J_\epsilon(u_\epsilon) \leq J_\epsilon(u)$, so we obtain

$$J(u_\epsilon) \leq J(u_\epsilon) + \frac{1}{\epsilon}\,\psi(u_\epsilon) = J_\epsilon(u_\epsilon) \leq J_\epsilon(u) = J(u),$$

that is,

$$J_\epsilon(u_\epsilon) \leq J(u). \tag{$*_1$}$$

Since J is coercive, the family $(u_\epsilon)_{\epsilon>0}$ is bounded. By compactness (since we are in $\mathbb{R}^n$), there exists a subsequence $(u_{\epsilon(i)})_{i\geq 0}$ with $\lim_{i\mapsto\infty} \epsilon(i) = 0$ and some element $u' \in \mathbb{R}^n$ such that

$$\lim_{i\mapsto\infty} u_{\epsilon(i)} = u'.$$

From the inequality $J(u_\epsilon) \leq J(u)$ proven in $(*_1)$ and the continuity of J, we deduce that

$$J(u') = \lim_{i\mapsto\infty} J(u_{\epsilon(i)}) \leq J(u). \tag{$*_2$}$$

By definition of $J_\epsilon(u_\epsilon)$ and $(*_1)$, we have

$$0 \leq \psi(u_{\epsilon(i)}) \leq \epsilon(i)(J(u) - J(u_{\epsilon(i)})),$$

and since the sequence $(u_{\epsilon(i)})_{i\geq 0}$ converges, the numbers $J(u) - J(u_{\epsilon(i)})$ are bounded independently of i. Consequently, since $\lim_{i\mapsto\infty} \epsilon(i) = 0$ and since the function ψ is continuous, we have

$$0 = \lim_{i\mapsto\infty} \psi(u_{\epsilon(i)}) = \psi(u'),$$

which shows that $u' \in U$. Since by $(*_2)$ we have $J(u') \leq J(u)$, and since both $u, u' \in U$ and u is the unique minimizer of J over U we must have $u' = u$. Therfore u' is the unique minimizer of J over U. But then the whole family $(u_\epsilon)_{\epsilon>0}$ converges to u since we can use the same argument as above for *every* subsequence of $(u_\epsilon)_{\epsilon>0}$. $\square$

Note that a convex function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is automatically continuous, so the assumption of continuity is redundant.

As an application of Proposition 49.19, if U is given by

$$U = \{v \in \mathbb{R}^n \mid \varphi_i(v) \leq 0,\ i = 1, \ldots, m\},$$

where the functions $\varphi_i\colon \mathbb{R}^n \to \mathbb{R}$ are convex, we can take ψ to be the function given by

$$\psi(v) = \sum_{i=1}^m \max\{\varphi_i(v), 0\}.$$

In practice, the applicability of the penalty-function method is limited by the difficulty to construct effectively "good" functions ψ, for example, differentiable ones. Note that in the above example the function ψ is not diferentiable. A better penalty function is

$$\psi(v) = \sum_{i=1}^{m} (\max\{\varphi_i(v), 0\})^2.$$

Another way to deal with constrained optimization problems is to use *duality*. This approach is investigated in Chapter 50.

49.13 Summary

The main concepts and results of this chapter are listed below:

- Minimization, minimizer.
- Coercive functions.
- Minima of quadratic functionals.
- The theorem of Lions and Stampacchia.
- Lax–Milgram's theorem.
- Elliptic functionals.
- Descent direction, exact line search, backtracking line search.
- Method of relaxation.
- Gradient descent.
- Gradient descent method with fixed stepsize parameter.
- Gradient descent method with variable stepsize parameter.
- Steepest descent method for the Euclidean norm.
- Gradient descent method with backtracking line search.
- Normalized steepest descent direction.
- Unormalized steepest descent direction.
- Steepest descent method (with respect to the norm $\| \, \|$).
- Momentum term.

- Newton's method.
- Newton step.
- Newton decrement.
- Damped Newton phase.
- Quadratically convergent phase.
- Self-concordant functions.
- Conjugate gradient method.
- Projected gradient methods.
- Penalty methods.

49.14 Problems

Problem 49.1. Consider the function $J\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle + g(v),$$

where A is a real $n \times n$ symmetric positive definite matrix, $b \in \mathbb{R}^n$, and $g\colon \mathbb{R}^n \to \mathbb{R}$ is a continuous (not necessarily differentiable) convex function such that $g(v) \geq 0$ for all $v \in \mathbb{R}^n$, and let U be a nonempty, bounded, closed, convex subset of $\mathbb{R}^n$.

(1) Prove that there is a unique element $u \in U$ such that

$$J(u) = \inf_{v \in U} J(v).$$

Hint. Prove that J is strictly convex on $\mathbb{R}^n$.

(2) Check that

$$J(v) - J(u) = \langle Au - b, v - u\rangle + g(v) - g(u) + \frac{1}{2}\langle A(v-u), v-u\rangle.$$

Prove that an element $u \in U$ minimizes J in U iff

$$\langle Au - b, v - u\rangle + g(v) - g(u) \geq 0 \quad \text{for all } v \in U.$$

Problem 49.2. Consider n piecewise C^1 functions $\varphi_i\colon [0,1] \to \mathbb{R}$ and assume that these functions are linearly independent and that

$$\sum_{i=1}^{n} \varphi_i(x) = 1 \quad \text{for all } x \in [0,1].$$

Let $J\colon \mathbb{R}^n \to \mathbb{R}$ be the function given by

$$J(v) = \sum_{i,j=1}^{n} a_{iij} v_i v_j + \sum_{i=1}^{n} b_i v_i,$$

where $v = (v_1, \ldots, v_n)$ and

$$a_{ij} = \int_0^1 \varphi_i'(t)\varphi_j'(t)dt, \quad b_i = \int_0^1 \varphi_i(t)dt.$$

(1) Let U_1 be the subset of $\mathbb{R}^n$ given by

$$U_1 = \left\{ v \in \mathbb{R}^n \mid \sum_{i=1}^{n} b_i v_i = 0 \right\}.$$

Consider the problem of finding a minimum of J over U_1. Prove that the Lagrange multiplier λ for which the Lagrangian has a critical point is $\lambda = -1$.

(2) Prove that the map defined on U_1 by

$$\|v\| = \left(\int_0^1 \left(\sum_{i=1}^{n} v_i \varphi_i'(x) \right)^2 dx \right)^{1/2}$$

is a norm. Prove that J is elliptic on U_1 with this norm. Prove that J has a unique minimum on U_1.

(3) Consider the the subset of $\mathbb{R}^n$ given by

$$U_2 = \left\{ v \in \mathbb{R}^n \mid \sum_{i=1}^{n} (\varphi_i(1) + \varphi_i(0)) v_i = 0 \right\}.$$

Consider the problem of finding a minimum of J over U_2. Prove that the Lagrange multiplier λ for which the Lagrangian has a critical point is $\lambda = -1/2$. Prove that J is elliptic on U_2 with the same norm as in (2). Prove that J has a unique minimum on U_2.

(4) Consider the the subset of $\mathbb{R}^n$ given by

$$U_3 = \left\{ v \in \mathbb{R}^n \mid \sum_{i=1}^{n} (\varphi_i(1) - \varphi_i(0)) v_i = 0 \right\}.$$

This time show that the necessary condition for having a minimum on U_3 yields the equation $1 + \lambda(1 - 1) = 0$. Conclude that J does not have a minimum on U_3.

Problem 49.3. Let A be a real $n \times n$ symmetric positive definite matrix and let $b \in \mathbb{R}^n$.

(1) Prove that if we apply the steepest descent method (for the Euclidean norm) to

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

and if we define the norm $\|v\|_A$ by

$$\|v\|_A = \langle Av, v\rangle^{1/2},$$

we get the inequality

$$\|u_{k+1} - u\|_A^2 \leq \|u_k - u\|_A^2 \left(1 - \frac{\|A(u_k - u)\|_2^4}{\|u_k - u\|_A^2 \, \|A(u_k - u)\|_A^2}\right).$$

(2) Using a diagonalization of A, where the eigenvalues of A are denoted $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, prove that

$$\|u_{k+1} - u\|_A \leq \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1} \|u_k - u\|_A,$$

where $\mathrm{cond}_2(A) = \lambda_n/\lambda_1$, and thus

$$\|u_k - u\|_A \leq \left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}\right)^k \|u_0 - u\|_A.$$

(3) Prove that when $\mathrm{cond}_2(A) = 1$, then $A = I$ and the method converges in one step.

Problem 49.4. Prove that the method of Polak–Ribière converges if $J \colon \mathbb{R}^n \to \mathbb{R}$ is elliptic and a C^2 function.

Problem 49.5. Prove that the backtracking line search method described in Section 49.5 has the property that for t small enough the condition $J(u_k + td_k) \leq J(u_k) + \alpha t\langle \nabla J_{u_k}, d_k\rangle$ will hold and the search will stop. Prove that the exit inequality $J(u_k + td_k) \leq J(u_k) + \alpha t\langle \nabla J_{u_k}, d_k\rangle$ holds for all $t \in (0, t_0]$, for some $t_0 > 0$, so the backtracking line search stops with a step length ρ_k that satisfies $\rho_k = 1$ or $\rho_k \in (\beta t_0, t_0]$.

Problem 49.6. Let $d_{\mathrm{nsd},k}$ and $d_{\mathrm{sd},k}$ be the normalized and unnormalized descent directions of the steepest descent method for an arbitrary norm (see Section 49.8). Prove that

$$\begin{aligned}
\langle \nabla J_{u_k}, d_{\mathrm{nsd},k}\rangle &= -\|\nabla J_{u_k}\|^D \\
\langle \nabla J_{u_k}, d_{\mathrm{sd},k}\rangle &= -(\|\nabla J_{u_k}\|^D)^2 \\
d_{\mathrm{sd},k} &= \arg\min_v \left(\langle \nabla J_{u_k}, v\rangle + \frac{1}{2}\|v\|^2\right).
\end{aligned}$$

Problem 49.7. If P is a symmetric positive definite matrix, prove that $\|z\|_P = (z^\top P z)^{1/2} = \left\|P^{1/2}z\right\|_2$ is a norm. Prove that the normalized steepest descent direction is

$$d_{\mathrm{nsd,k}} = -(\nabla J_{u_k}^\top P^{-1} \nabla J_{u_k})^{-1/2} P^{-1} \nabla J_{u_k},$$

the dual norm is $\|z\|^D = \left\|P^{-1/2}z\right\|_2$, and the steepest descent direction with respect to $\| \, \|_P$ is given by

$$d_{\mathrm{sd},k} = -P^{-1}\nabla J_{u_k}.$$

Problem 49.8. If $\| \, \|$ is the ℓ^1-norm, then show that $d_{\mathrm{nsd,k}}$ is determined as follows: let i be any index for which $\|\nabla J_{u_k}\|_\infty = |(\nabla J_{u_k})_i|$. Then

$$d_{\mathrm{nsd,k}} = -\mathrm{sign}\left(\frac{\partial J}{\partial x_i}(u_k)\right) e_i,$$

where e_i is the ith canonical basis vector, and

$$d_{\mathrm{sd,k}} = -\frac{\partial J}{\partial x_i}(u_k) e_i.$$

Problem 49.9. (From Boyd and Vandenberghe [29], Problem 9.12). If $\nabla^2 f(x)$ is singular (or very ill-conditioned), the Newton step $d_{\mathrm{nt}} = -(\nabla^2 J(x))^{-1}\nabla J_x$ is not well defined. Instead we can define a search direction d_{tr} as the solution of the problem

$$\begin{aligned} &\text{minimize} && (1/2)\langle Hv, v\rangle + \langle g, v\rangle \\ &\text{subject to} && \|v\|_2 \leq \gamma, \end{aligned}$$

where $H = \nabla^2 f_x$, $g = \nabla f_x$, and γ is some positive constant. The idea is to use a *trust region*, which is the closed ball $\{v \mid \|v\|_2 \leq \gamma\}$. The point $x + d_{\mathrm{tr}}$ minimizes the second-order approximation of f at x, subject to the constraint that

$$\|x + d_{\mathrm{tr}} - x\|_2 \leq \gamma.$$

The parameter γ, called the *trust parameter*, reflects our confidence in the second-order approximation.

Prove that d_{tr} minimizes

$$\frac{1}{2}\langle Hv, v\rangle + \langle g, v\rangle + \widehat{\beta}\,\|v\|_2^2,$$

for some $\widehat{\beta}$.

Problem 49.10. (From Boyd and Vandenberghe [29], Problem 9.9). Prove that the Newton decrement $\lambda(x)$ is given by

$$\lambda(x) = \sup_{v \neq 0} -\frac{\langle \nabla J_x, v\rangle}{(\langle \nabla^2 J_x v, v\rangle)^{1/2}}.$$

Problem 49.11. Show that the function f given by $f(x) = \log(e^x + e^{-x})$ has a unique minimum for $x^* = 0$. Run Newton's method with fixed step size $t = 1$, starting with $x_0 = 1$, and then $x_0 = 1.1$. What do you observe?

Problem 49.12. Write a `Matlab` program implementing the conjugate gradient method. Test your program with the $n \times n$ matrix

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

and various right-hand sides, for various values of n. Verify that the running time is $O(n^{3/2})$.

Chapter 50

Introduction to Nonlinear Optimization

This chapter contains the most important results of nonlinear optimization theory.

In Chapter 40 we investigated the problem of determining when a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum in a subset U of Ω defined by equational constraints, namely

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable). Theorem 40.2 gave a necessary condition in terms of the Lagrange multipliers. In Section 40.3 we assumed that U was a convex subset of Ω; then Theorem 40.9 gave us a necessary condition for the function $J\colon \Omega \to \mathbb{R}$ to have a local minimum at u with respect to U if dJ_u exists, namely

$$dJ_u(v - u) \geq 0 \quad \text{for all } v \in U.$$

Our first goal is to find a necessary criterion for a function $J\colon \Omega \to \mathbb{R}$ to have a minimum on a subset U, even is this subset is *not* convex. This can be done by introducing a notion of "tangent cone" at a point $u \in U$. We define the cone of feasible directions and then state a necessary condition for a function to have local minimum on a set U that is not necessarily convex in terms of the cone of feasible directions. The cone of feasible directions is not always convex, but it is if the constraints are inequality constraints. An inequality constraint $\varphi(u) \leq 0$ is said to be *active* if $\varphi(u) = 0$. One can also define the notion of *qualified constraint*. Theorem 50.5 gives necessary conditions for a function J to have a minimum on a subset U defined by qualified inequality constraints in terms of the Karush–Kuhn–Tucker conditions (for short KKT conditions), which involve nonnegative Lagrange multipliers. The proof relies on a version of the Farkas–Minkowski lemma. Some of the KTT conditions assert that $\lambda_i \varphi_i(u) = 0$, where $\lambda_i \geq 0$ is the Lagrange multiplier associated with the constraint $\varphi_i \leq 0$. To some extent, this implies that active constaints are more important than inactive constraints, since if $\varphi_i(u) < 0$ is an inactive constraint, then $\lambda_i = 0$. In general, the KKT conditions are useless unlesss the constraints are convex. In this case, there is a manageable notion of qualified constraint given by Slater's conditions. Theorem 50.6 gives necessary conditions for a function J to have a minimum on a subset U defined by convex inequality constraints in terms of the Karush–Kuhn–Tucker conditions. Furthermore, if J is also convex and if the KKT conditions hold, then J has a global minimum.

In Section 50.4, we apply Theorem 50.6 to the special case where the constraints are equality constraints, which can be expressed as $Ax = b$. In the special case where the convex objective function J is a convex quadratic functional of the form

$$J(x) = \frac{1}{2}x^\top Px + q^\top x + r,$$

where P is a $n \times n$ symmetric positive semidefinite matrix, the necessary and sufficient conditions for having a minimum are expressed by a linear system involving a matrix called the KKT matrix. We discuss conditions that guarantee that the KKT matrix is invertible, and how to solve the KKT system. We also briefly discuss variants of Newton's method dealing with equality constraints.

We illustrate the KKT conditions on an interesting example, the so-called hard margin support vector machine; see Sections 50.5 and 50.6. The problem is a classification problem, or more accurately a separation problem. Suppose we have two nonempty disjoint finite sets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$ in $\mathbb{R}^n$. Our goal is to find a hyperplane H of equation $w^\top x - b = 0$ (where $w \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$), such that all the blue points u_i are in one of the two open half-spaces determined by H, and all the red points v_j are in the other open half-space determined by H.

If the two sets are indeed separable, then in general there are infinitely many hyperplanes separating them. Vapnik had the idea to find a hyperplane that maximizes the smallest distance between the points and the hyperplane. Such a hyperplane is indeed unique and is called a maximal hard margin hyperplane, or hard margin support vector machine. The support vectors are those for which the constraints are active.

Section 50.7 contains the most important results of the chapter. The notion of Lagrangian duality is presented. Given a primal optimization problem (P) consisting in minimizing an objective function $J(v)$ with respect to some inequality constraints $\varphi_i(v) \leq 0$, $i = 1, \ldots, m$, we define the *dual function* $G(\mu)$ as the result of minimizing the Lagrangian

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v)$$

with respect to v, with $\mu \in \mathbb{R}^m_+$. The dual program (D) is then to maximize $G(\mu)$ with respect to $\mu \in \mathbb{R}^m_+$. It turns out that G is a concave function, and the dual program is an unconstrained maximization. This is actually a misleading statement because G is generally a partial function, so maximizing $G(\mu)$ is equivalent to a constrained maximization problem in which the constraints specify the domain of G, but in many cases, we obtain a dual program simpler than the primal program. If d^* is the optimal value of the dual program and if p^* is the optimal value of the primal program, we always have

$$d^* \leq p^*,$$

which is known as *weak duality*. Under certain conditions, $d^* = p^*$, that is, the duality gap is zero, in which case we say that *strong duality* holds. Also, under certain conditions, a solution of the dual yields a solution of the primal, and if the primal has an optimal solution, then the dual has an optimal solution, but beware that the converse is generally false (see Theorem 50.17). We also show how to deal with equality constraints, and discuss the use of conjugate functions to find the dual function. Our coverage of Lagrangian duality is quite thorough, but we do not discuss more general orderings such as the semidefinite ordering. For these topics which belong to convex optimization, the reader is referred to Boyd and Vandenberghe [29].

Our approach in this chapter is very much inspired by Ciarlet [41] because we find it one of the more direct, and it is general enough to accomodate Hilbert spaces. The field of nonlinear optimization and convex optimization is vast and there are many books on the subject. Among those we recommend (in alphabetic order) Bertsekas [16, 17, 18], Bertsekas, Nedić, and Ozdaglar [19], Boyd and Vandenberghe [29], Luenberger [116], and Luenberger and Ye [117].

50.1 The Cone of Feasible Directions

Let V be a normed vector space and let U be a nonempty subset of V. For any point $u \in U$, consider any converging sequence $(u_k)_{k\geq 0}$ of vectors $u_k \in U$ having u as their limit, with $u_k \neq u$ for all $k \geq 0$, and look at the sequence of "unit chords,"

$$\frac{u_k - u}{\|u_k - u\|}.$$

This sequence could oscillate forever, or it could have a limit, some unit vector $\widehat{w} \in V$. In the second case, all nonzero vectors $\lambda\widehat{w}$ for all $\lambda > 0$, belong to an object called the cone of feasible directions at u. First, we need to define the notion of cone.

Definition 50.1. Given a (real) vector space V, a nonempty subset $C \subseteq V$ is a *cone with apex* 0 (for short, a *cone*), if for any $v \in V$, if $v \in C$, then $\lambda v \in C$ for all $\lambda > 0$ ($\lambda \in \mathbb{R}$). For any $u \in V$, a *cone with apex* u is any nonempty subset of the form $u + C = \{u + v \mid v \in C\}$, where C is a cone with apex 0; see Figure 50.1.

Observe that a cone with apex 0 (or u) is not necessarily convex, and that 0 does not necessarily belong to C (resp. u does not necessarily belong to $u + C$) (although in the case of the cone of feasible directions $C(u)$ we have $0 \in C(u)$). The condition for being a cone only asserts that if a nonzero vector v belongs to C, then the open ray $\{\lambda v \mid \lambda > 0\}$ (resp. the affine open ray $u + \{\lambda v \mid \lambda > 0\}$) also belongs to C.

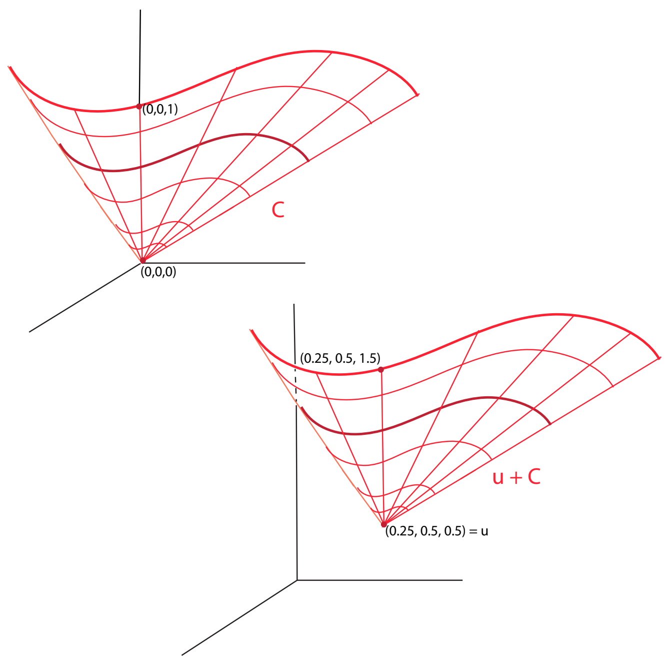


Figure 50.1: Let C be the cone determined by the bold orange curve through $(0, 0, 1)$ in the plane $z = 1$. Then $u+C$, where $u = (0.25, 0.5, 0.5)$, is the affine translate of C via the vector u.

Definition 50.2. Let V be a normed vector space and let U be a nonempty subset of V. For any point $u \in U$, the *cone $C(u)$ of feasible directions at u* is the union of $\{0\}$ and the set of all nonzero vectors $w \in V$ for which there exists some convergent sequence $(u_k)_{k\geq 0}$ of vectors such that

(1) $u_k \in U$ and $u_k \neq u$ for all $k \geq 0$, and $\lim_{k\mapsto\infty} u_k = u$.

(2) $\lim_{k\mapsto\infty} \dfrac{u_k - u}{\|u_k - u\|} = \dfrac{w}{\|w\|}$, with $w \neq 0$.

Condition (2) can also be expressed as follows: there is a sequence $(\delta_k)_{k\geq 0}$ of vectors $\delta_k \in V$ such that

$$u_k = u + \|u_k - u\| \frac{w}{\|w\|} + \|u_k - u\| \delta_k, \quad \lim_{k\mapsto\infty} \delta_k = 0, \ w \neq 0.$$

Figure 50.2 illustrates the construction of w in $C(u)$.

Clearly, the cone $C(u)$ of feasible directions at u is a cone with apex 0, and $u+C(u)$ is a cone with apex u. Obviously, it would be desirable to have conditions on U that imply that $C(u)$ is a convex cone. Such conditions will be given later on.

Observe that the cone $C(u)$ of feasible directions at u contains the velocity vectors at u of all curves γ in U through u. If $\gamma\colon (-1, 1) \to U$ is such a curve with $\gamma(0) = u$, and if $\gamma'(u) \neq 0$

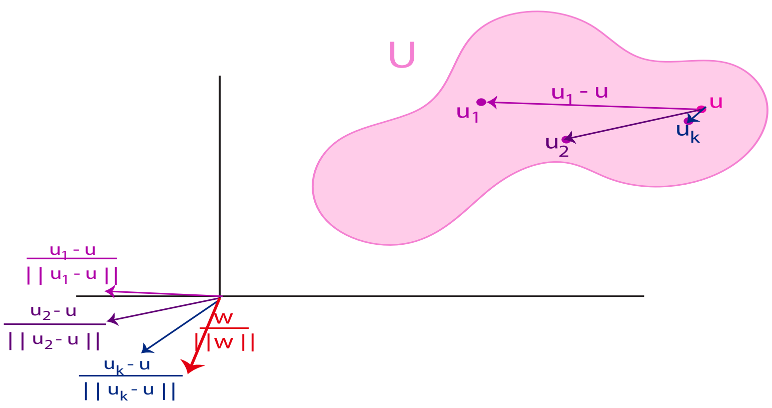


Figure 50.2: Let U be the pink region in $\mathbb{R}^2$ with fuchsia point $u \in U$. For any sequence $(u_k)_{k\geq 0}$ of points in U which converges to u, form the chords $u_k - u$ and take the limit to construct the red vector w.

exists, then there is a sequence $(u_k)_{k\geq 0}$ of vectors in U converging to u as in Definition 50.2, with $u_k = \gamma(t_k)$ for some sequence $(t_k)_{k\geq 0}$ of reals $t_k > 0$ such that $\lim_{k\mapsto\infty} t_k = 0$, so that

$$u_k - u = t_k\gamma'(0) + t_k\epsilon_k, \quad \lim_{k\mapsto\infty} \epsilon_k = 0,$$

and we get

$$\lim_{k\mapsto\infty} \frac{u_k - u}{\|u_k - u\|} = \frac{\gamma'(0)}{\|\gamma'(0)\|}.$$

For an illustration of this paragraph in $\mathbb{R}^2$, see Figure 50.3.

Example 50.1. In $V = \mathbb{R}^2$, let φ_1 and φ_2 be given by

$$\begin{aligned}\varphi_1(u_1, u_2) &= -u_1 - u_2 \\ \varphi_2(u_1, u_2) &= u_1(u_1^2 + u_2^2) - (u_1^2 - u_2^2),\end{aligned}$$

and let

$$U = \{(u_1, u_2) \in \mathbb{R}^2 \mid \varphi_1(u_1, u_2) \leq 0,\ \varphi_2(u_1, u_2) \leq 0\}.$$

The region U is shown in Figure 50.4 and is bounded by the curve given by the equation $\varphi_1(u_1, u_2) = 0$, that is, $-u_1 - u_2 = 0$, the line of slope -1 through the origin, and the curve given by the equation $u_1(u_1^2 + u_2^2) - (u_1^2 - u_2^2) = 0$, a nodal cubic through the origin. We obtain a parametric definition of this curve by letting $u_2 = tu_1$, and we find that

$$u_1(t) = \frac{u_1^2(t) - u_2^2(t)}{u_1^2(t) + u_2^2(t)} = \frac{1 - t^2}{1 + t^2}, \quad u_2(t) = \frac{t(1 - t^2)}{1 + t^2}.$$

The tangent vector at t is given by $(u_1'(t), u_2'(t))$ with

$$u_1'(t) = \frac{-2t(1 + t^2) - (1 - t^2)2t}{(1 + t^2)^2} = \frac{-4t}{(1 + t^2)^2}$$

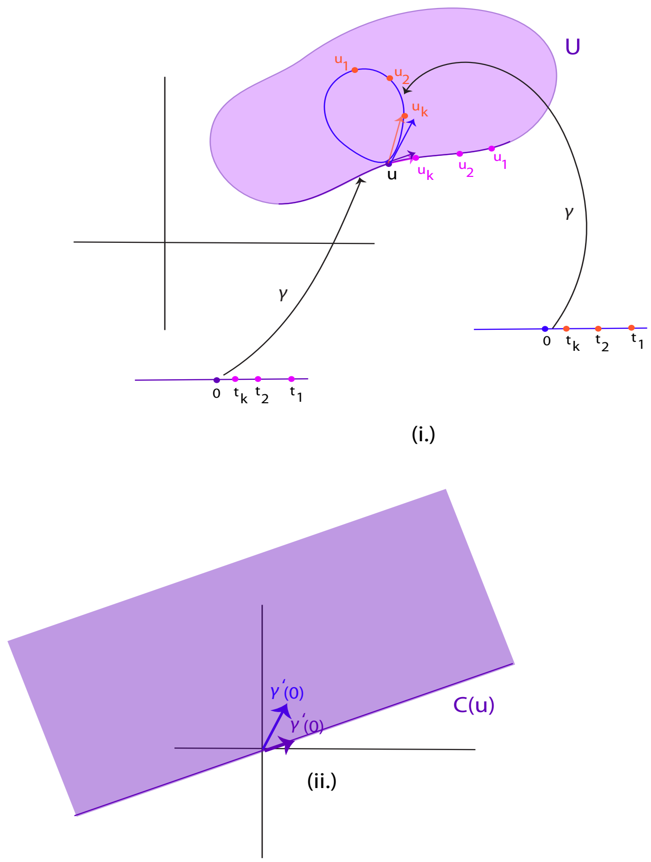


Figure 50.3: Let U be purple region in $\mathbb{R}^2$ and u be the designated point on the boundary of U. Figure (i.) illustrates two curves through u and two sequences $(u_k)_{k\geq 0}$ converging to u. The limit of the chords $u_k - u$ corresponds to the tangent vectors for the appropriate curve. Figure (ii.) illustrates the half plane $C(u)$ of feasible directions.

and

$$u_2'(t) = \frac{(1-3t^2)(1+t^2) - (t-t^3)2t}{(1+t^2)^2} = \frac{1-2t^2-3t^4-2t^2+2t^4}{(1+t^2)^2} = \frac{1-4t^2-t^4}{(1+t^2)^2}.$$

The nodal cubic passes through the origin for $t = \pm 1$, and for $t = -1$ the tangent vector is $(1, -1)$, and for $t = 1$ the tangent vector is $(-1, -1)$. The cone of feasible directions $C(0)$ at the origin is given by

$$C(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 + u_2 \geq 0,\ |u_1| \geq |u_2|\}.$$

This is not a convex cone since it contains the sector delineated by the lines $u_2 = u_1$ and $u_2 = -u_1$, but also the ray supported by the vector $(-1, 1)$.

The two crucial properties of the cone of feasible directions are shown in the following proposition.

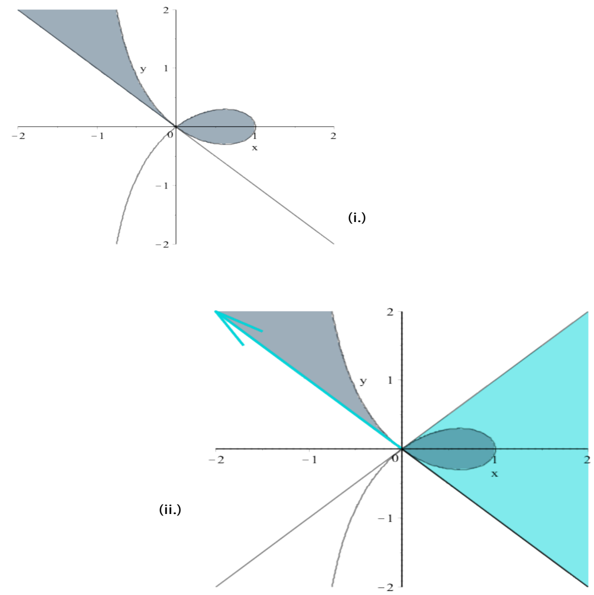


Figure 50.4: Figure (i.) illustrates U as the shaded gray region which lies between the line $y = -x$ and nodal cubic. Figure (ii.) shows the cone of feasible directions, $C(0)$, as the union of the turquoise triangular cone and the turquoise directional ray $(-1, 1)$.

Proposition 50.1. *Let U be any nonempty subset of a normed vector space V.*

(1) For any $u \in U$, the cone $C(u)$ of feasible directions at u is closed.

(2) Let $J \colon \Omega \to \mathbb{R}$ be a function defined on an open subset Ω containing U. If J has a local minimum with respect to the set U at a point $u \in U$, and if J'_u exists at u, then

$$J'_u(v - u) \geq 0 \quad \text{for all } v \in u + C(u).$$

Proof. (1) Let $(w_n)_{n\geq 0}$ be a sequence of vectors $w_n \in C(u)$ converging to a limit $w \in V$. We may assume that $w \neq 0$, since $0 \in C(u)$ by definition, and thus we may also assume that $w_n \neq 0$ for all $n \geq 0$. By definition, for every $n \geq 0$, there is a sequence $(u^n_k)_{k\geq 0}$ of vectors in V and some $w_n \neq 0$ such that

(1) $u^n_k \in U$ and $u^n_k \neq u$ for all $k \geq 0$, and $\lim_{k \mapsto \infty} u^n_k = u$.

(2) There is a sequence $(\delta^n_k)_{k\geq 0}$ of vectors $\delta^n_k \in V$ such that

$$u^n_k = u + \|u^n_k - u\| \frac{w_n}{\|w_n\|} + \|u^n_k - u\| \, \delta^n_k, \quad \lim_{k \mapsto \infty} \delta^n_k = 0, \;\; w_n \neq 0.$$

Let $(\epsilon_n)_{n\geq 0}$ be a sequence of real numbers $\epsilon_n > 0$ such that $\lim_{n\mapsto\infty} \epsilon_n = 0$ (for example, $\epsilon_n = 1/(n + 1)$). Due to the convergence of the sequences (u^n_k) and (δ^n_k) for every fixed n, there exist an integer $k(n)$ such that

$$\left\|u^n_{k(n)} - u\right\| \leq \epsilon_n, \quad \left\|\delta^n_{k(n)}\right\| \leq \epsilon_n.$$

Consider the sequence $(u^n_{k(n)})_{n\geq 0}$. We have

$$u^n_{k(n)} \in U, \;\; u^n_{k(n)} \neq 0, \;\; \text{for all } n \geq 0, \;\; \lim_{n\mapsto\infty} u^n_{k(n)} = u,$$

and we can write

$$u^n_{k(n)} = u + \left\|u^n_{k(n)} - u\right\| \frac{w}{\|w\|} + \left\|u^n_{k(n)} - u\right\| \left(\delta^n_{k(n)} + \left(\frac{w_n}{\|w_n\|} - \frac{w}{\|w\|}\right)\right).$$

Since $\lim_{k\mapsto\infty}(w_n/\left\|w_n\right\|) = w/\left\|w\right\|$, we conclude that $w \in C(u)$. See Figure 50.5.

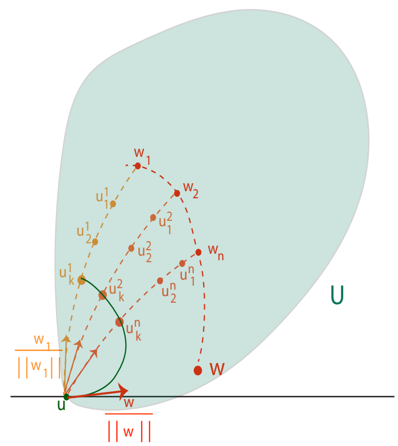


Figure 50.5: Let U be the mint green region in $\mathbb{R}^2$ with $u = (0,0)$. Let $(w_n)_{n\geq 0}$ be a sequence of vectors (points) along the upper dashed curve which converge to w. By following the dashed orange longitudinal curves, and selecting an appropriate vector(point), we construct the dark green curve in U, which passes through u, and at u has tangent vector proportional to w.

(2) Let $w = v - u$ be any nonzero vector in the cone $C(u)$, and let $(u_k)_{k\;\geq 0}$ be a sequence of vectors in $U - \{u\}$ such that

(1) $\lim_{k\mapsto\infty} u_k = u$.

(2) There is a sequence $(\delta_k)_{k\geq 0}$ of vectors $\delta_k \in V$ such that

$$u_k - u = \|u_k - u\| \frac{w}{\|w\|} + \|u_k - u\|\, \delta_k, \quad \lim_{k\mapsto\infty} \delta_k = 0, \;\; w \neq 0,$$

(3) $J(u) \leq J(u_k)$ for all $k \geq 0$.

Since J is differentiable at u, we have

$$0 \leq J(u_k) - J(u) = J'_u(u_k - u) + \|u_k - u\| \, \epsilon_k, \tag{$*$}$$

for some sequence $(\epsilon_k)_{k \geq 0}$ such that $\lim_{k \mapsto \infty} \epsilon_k = 0$. Since J'_u is linear and continuous, and since

$$u_k - u = \|u_k - u\| \frac{w}{\|w\|} + \|u_k - u\| \, \delta_k, \quad \lim_{k \mapsto \infty} \delta_k = 0, \;\; w \neq 0,$$

$(*)$ implies that

$$0 \leq \frac{\|u_k - u\|}{\|w\|} \, (J'_u(w) + \eta_k),$$

with

$$\eta_k = \|w\| \, (J'_u(\delta_k) + \epsilon_k).$$

Since J'_u is continuous, we have $\lim_{k \mapsto \infty} \eta_k = 0$. But then $J'_u(w) \geq 0$, since if $J'_u(w) < 0$, then for k large enough the expression $J'_u(w) + \eta_k$ would be negative, and since $u_k \neq u$, the expression $(\|u_k - u\| \, / \, \|w\|)(J'_u(w) + \eta_k)$ would also be negative, a contradiction. $\square$

From now on we assume that U is defined by a set of inequalities, that is

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \;\; 1 \leq i \leq m\},$$

where the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable). As we explained earlier, an equality constraint $\varphi_i(x) = 0$ is treated as the conjunction of the two inequalities $\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$. Later on we will see that when the functions φ_i are convex, since $-\varphi_i$ is not necessarily convex, it is desirable to treat equality constraints separately, but for the time being we won't.

50.2 Active Constraints and Qualified Constraints

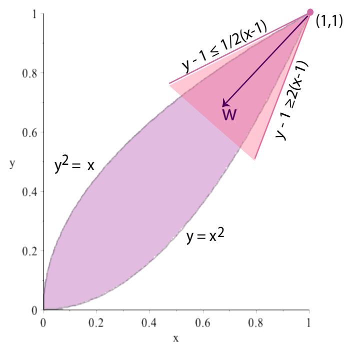
Our next goal is find sufficient conditions for the cone $C(u)$ to be convex, for any $u \in U$. For this we assume that the functions φ_i are differentiable at u. It turns out that the constraints φ_i that matter are those for which $\varphi_i(u) = 0$, namely the constraints that are tight, or as we say, active.

Definition 50.3. Given m functions $\varphi_i \colon \Omega \to \mathbb{R}$ defined on some open subset Ω of some vector space V, let U be the set defined by

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \;\; 1 \leq i \leq m\}.$$

For any $u \in U$, a constraint φ_i is said to be *active* at u if $\varphi_i(u) = 0$, else *inactive* at u if $\varphi_i(u) < 0$.

If a constraint φ_i is active at u, this corresponds to u being on a piece of the boundary of U determined by some of the equations $\varphi_i(u) = 0$; see Figure 50.6.



(i.)

(ii.)

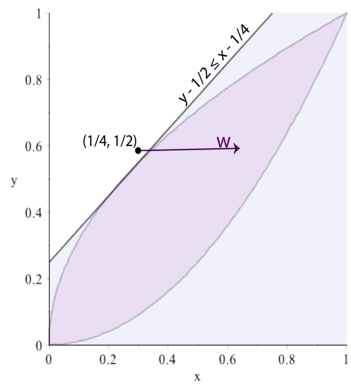


Figure 50.6: Let U be the light purple planar region which lies between the curves $y = x^2$ and $y^2 = x$. Figure (i.) illustrates the boundary point $(1, 1)$ given by the equalities $y-x^2 = 0$ and $y^2-x = 0$. The affine translate of cone of feasible directions, $C(1, 1)$, is illustrated by the pink triangle whose sides are the tangent lines to the boundary curves. Figure (ii.) illustrates the boundary point $(1/4, 1/2)$ given by the equality $y^2 - x = 0$. The affine translate of $C(1/4, 1/2)$ is the lilac half space bounded by the tangent line to $y^2 = x$ through $(1/4, 1/2)$.

Definition 50.4. For any $u \in U$, with

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

we define $I(u)$ as the set of indices

$$I(u) = \{i \in \{1, \ldots, m\} \mid \varphi_i(u) = 0\}$$

where the constraints are active. We define the set $C^*(u)$ as

$$C^*(u) = \{v \in V \mid (\varphi_i')_u(v) \leq 0,\ \ i \in I(u)\}.$$

Since each $(\varphi_i')_u$ is a linear form, the subset

$$C^*(u) = \{v \in V \mid (\varphi_i')_u(v) \leq 0,\ \ i \in I(u)\}$$

is the intersection of half spaces passing through the origin, so it is a convex set, and obviously it is a cone. If $I(u) = \emptyset$, then $C^*(u) = V$.

The special kinds of $\mathcal{H}$-polyhedra of the form $C^*(u)$ cut out by hyperplanes through the origin are called *$\mathcal{H}$-cones*. It can be shown that every $\mathcal{H}$-cone is a polyhedral cone (also called a *$\mathcal{V}$-cone*), and conversely. The proof is nontrivial; see Gallier [73] and Ziegler [195].

We will prove shortly that we always have the inclusion

$$C(u) \subseteq C^*(u).$$

However, the inclusion can be strict, as in Example 50.1. Indeed for $u = (0, 0)$ we have $I(0, 0) = \{1, 2\}$ and since

$$(\varphi_1')_{(u_1,u_2)} = (-1 \;\; -1), \quad (\varphi_2')_{(u_1,u_2)} = (3u_1^2 + u_2^2 - 2u_1 \;\; 2u_1u_2 + 2u_2),$$

we have $(\varphi_2')_{(0,0)} = (0 \;\; 0)$, and thus $C^*(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 + u_2 \geq 0\}$ as illustrated in Figure 50.7.

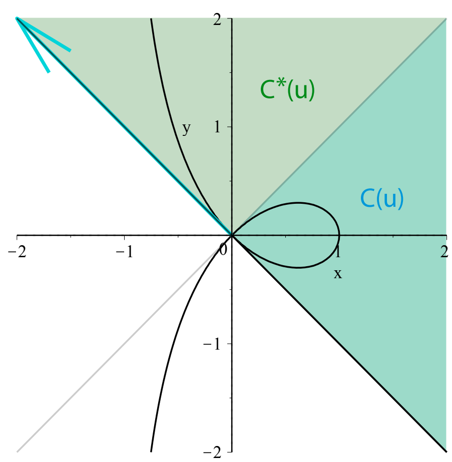


Figure 50.7: For $u = (0, 0)$, $C^*(u)$ is the sea green half space given by $u_1 + u_2 \geq 0$. This half space strictly contains $C(u)$, namely the union of the turquoise triangular cone and the directional ray $(-1, 1)$.

The conditions stated in the following definition are sufficient conditions that imply that $C(u) = C^*(u)$, as we will prove next.

Definition 50.5. For any $u \in U$, with

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \;\; 1 \leq i \leq m\},$$

if the functions φ_i are differentiable at u (in fact, we only this for $i \in I(u)$), we say that the constraints are *qualified* at u if the following conditions hold:

(a) Either the constraints φ_i are affine for all $i \in I(u)$, or

(b) There is some nonzero vector $w \in V$ such that the following conditions hold for all $i \in I(u)$:

  (i) $(\varphi_i')_u(w) \leq 0$.

  (ii) If φ_i is not affine, then $(\varphi_i')_u(w) < 0$.

Condition (b)(ii) implies that u is not a critical point of φ_i for every $i \in I(u)$, so there is no singularity at u in the zero locus of φ_i. Intuitively, if the constraints are qualified at u then the boundary of U near u behaves "nicely."

The boundary points illustrated in Figure 50.6 are qualified. Observe that $U = \{x \in \mathbb{R}^2 \mid \varphi_1(x, y) = y^2 - x \leq 0,\ \varphi_2(x, y) = x^2 - y \leq 0\}$. For $u = (1, 1)$, $I(u) = \{1, 2\}$, $(\varphi_1')_{(1,1)} = (-1\ \ 2)$, $(\varphi_2')_{(1,1)} = (2\ \ -1)$, and $w = (-1, -1)$ ensures that $(\varphi_1')_{(1,1)}$ and $(\varphi_1')_{(1,1)}$ satisfy Condition (b) of Definition 50.5. For $u = (1/4, 1/2)$, $I(u) = \{1\}$, $(\varphi_1')_{(1,1)} = (-1\ \ 1)$, and $w = (1, 0)$ will satisfy Condition (b).

In Example 50.1, the constraint $\varphi_2(u_1, u_2) = 0$ is not qualified at the origin because $(\varphi_2')_{(0,0)} = (0, 0)$; in fact, the origin is a self-intersection. In the example below, the origin is also a singular point, but for a different reason.

Example 50.2. Consider the region $U \subseteq \mathbb{R}^2$ determined by the two curves given by

$$\begin{aligned}\varphi_1(u_1, u_2) &= u_2 - \max(0, u_1^3)\\ \varphi_2(u_1, u_2) &= u_1^4 - u_2.\end{aligned}$$

We have $I(0, 0) = \{1, 2\}$, and since $(\varphi_1)'_{(0,0)}(w_1, w_2) = (0\ 1)\binom{w_1}{w_2} = w_2$ and $(\varphi_2')_{(0,0)}(w_1, w_2) = (0\ \ -1)\binom{w_1}{w_2} = -w_2$, we have $C^*(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_2 = 0\}$, but the constraints are not qualified at $(0, 0)$ since it is impossible to have simultaneously $(\varphi_1')_{(0,0)}(w_1, w_2) < 0$ and $(\varphi_2')_{(0,0)}(w_1, w_2) < 0$, so in fact $C(0) = \{(u_1, u_2) \in \mathbb{R}^2 \mid u_1 \geq 0,\ u_2 = 0\}$ is strictly contained in $C^*(0)$; see Figure 50.8.

Proposition 50.2. *Let u be any point of the set*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ 1 \leq i \leq m\},$$

where Ω is an open subset of the normed vector space V, and assume that the functions φ_i are differentiable at u (in fact, we only this for $i \in I(u)$). Then the following facts hold:

(1) The cone $C(u)$ of feasible directions at u is contained in the convex cone $C^*(u)$; that is,

$$C(u) \subseteq C^*(u) = \{v \in V \mid (\varphi_i')_u(v) \leq 0,\ i \in I(u)\}.$$

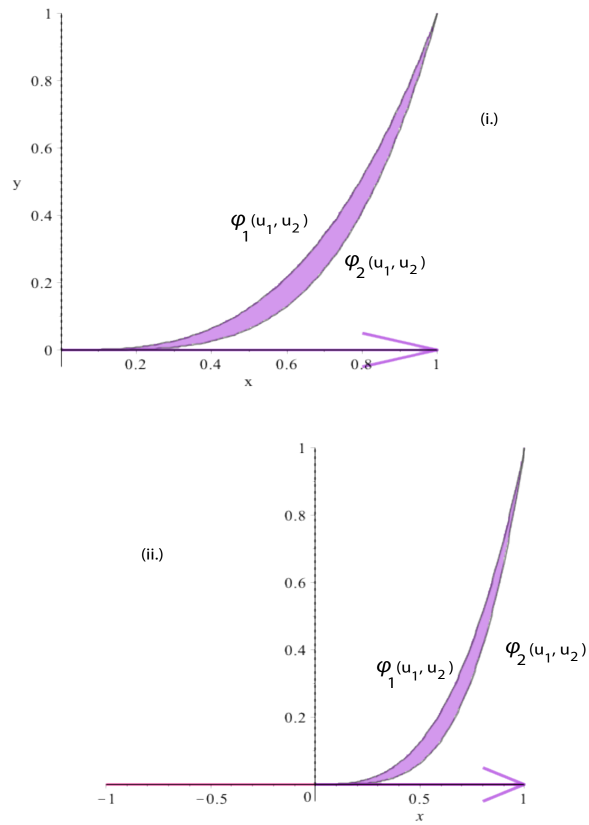


Figure 50.8: Figures (i.) and (ii.) illustrate the purple moon shaped region associated with Example 50.2. Figure (i.) also illustrates $C(0)$, the cone of feasible directions, while Figure (ii.) illustrates the strict containment of $C(0)$ in $C^*(0)$.

(2) If the constraints are qualified at u (and the functions φ_i are continuous at u for all $i \notin I(u)$ if we only assume φ_i differentiable at u for all $i \in I(u)$), then

$$C(u) = C^*(u).$$

Proof. (1) For every $i \in I(u)$, since $\varphi_i(v) \leq 0$ for all $v \in U$ and $\varphi_i(u) = 0$, the function $-\varphi_i$ has a local minimum at u with respect to U, so by Proposition 50.1(2), we have

$$(-\varphi_i')_u(v) \geq 0 \quad \text{for all } v \in C(u),$$

which is equivalent to $(\varphi_i')_u(v) \leq 0$ for all $v \in C(u)$ and for all $i \in I(u)$, that is, $u \in C^*(u)$.

(2)(a) First, let us assume that φ_i is affine for every $i \in I(u)$. Recall that φ_i must be given by $\varphi_i(v) = h_i(v) + c_i$ for all $v \in V$, where h_i is a linear form and $c_i \in \mathbb{R}$. Since the derivative of a linear map at any point is itself,

$$(\varphi_i')_u(v) = h_i(v) \quad \text{for all } v \in V.$$

Pick any nonzero $w \in C^*(u)$, which means that $(\varphi_i')_u(w) \leq 0$ for all $i \in I(u)$. For any sequence $(\epsilon_k)_{k\geq 0}$ of reals $\epsilon_k > 0$ such that $\lim_{k\mapsto\infty} \epsilon_k = 0$, let $(u_k)_{k\geq 0}$ be the sequence of vectors in V given by

$$u_k = u + \epsilon_k w.$$

We have $u_k - u = \epsilon_k w \neq 0$ for all $k \geq 0$ and $\lim_{k\mapsto\infty} u_k = u$. Furthermore, since the functions φ_i are continuous for all $i \notin I$, we have

$$0 > \varphi_i(u) = \lim_{k\mapsto\infty} \varphi_i(u_k),$$

and since φ_i is affine and $\varphi_i(u) = 0$ for all $i \in I$, we have $\varphi_i(u) = h_i(u) + c_i = 0$, so

$$\varphi_i(u_k) = h_i(u_k) + c_i = h_i(u_k) - h_i(u) = h_i(u_k - u) = (\varphi_i')_u(u_k - u) = \epsilon_k(\varphi_i')_u(w) \leq 0, \quad (*_0)$$

which implies that $u_k \in U$ for all k large enough. Since

$$\frac{u_k - u}{\|u_k - u\|} = \frac{w}{\|w\|} \quad \text{for all } k \geq 0,$$

we conclude that $w \in C(u)$. See Figure 50.9.

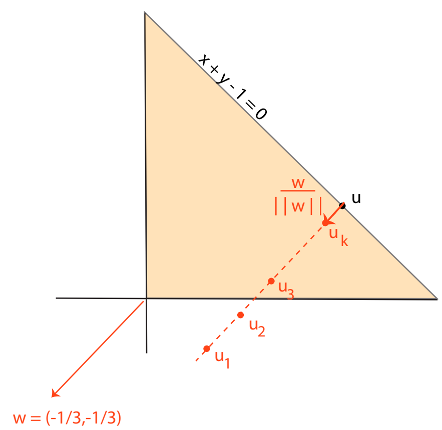


Figure 50.9: Let U be the peach triangle bounded by the lines $y = 0$, $x = 0$, and $y = -x + 1$. Let u satisfy the affine constraint $\varphi(x, y) = y + x - 1$. Since $\varphi'_{(x,y)} = (1\ 1)$, set $w = (-1, -1)$ and approach u along the line $u + tw$.

(2)(b) Let us now consider the case where some function φ_i is not affine for some $i \in I(u)$. Let $w \neq 0$ be some vector in V such that Condition (b) of Definition 50.5 holds, namely: for all $i \in I(u)$, we have

(i) $(\varphi_i')_u(w) \leq 0$.

(ii) If φ_i is not affine, then $(\varphi_i')_u(w) < 0$.

Pick any nonzero vector $v \in C^*(u)$, which means that $(\varphi'_i)_u(v) \leq 0$ for all $i \in I(u)$, and let $\delta > 0$ be any positive real number such that $v + \delta w \neq 0$. For any sequence $(\epsilon_k)_{k\geq 0}$ of reals $\epsilon_k > 0$ such that $\lim_{k\mapsto\infty} \epsilon_k = 0$, let $(u_k)_{k\geq 0}$ be the sequence of vectors in V given by

$$u_k = u + \epsilon_k(v + \delta w).$$

We have $u_k - u = \epsilon_k(v + \delta w) \neq 0$ for all $k \geq 0$ and $\lim_{k\mapsto\infty} u_k = u$. Furthermore, since the functions φ_i are continuous for all $i \notin I(u)$, we have

$$0 > \varphi_i(u) = \lim_{k\mapsto\infty} \varphi_i(u_k) \quad \text{for all } i \notin I(u). \tag{$*_1$}$$

Equation $(*_0)$ of the previous case shows that for all $i \in I(u)$ such that φ_i is affine, since $(\varphi'_i)_u(v) \leq 0$, $(\varphi'_i)_u(w) \leq 0$, and $\epsilon_k, \delta > 0$, we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w)) \leq 0 \quad \text{for all } i \in I(u) \text{ and } \varphi_i \text{ affine}. \tag{$*_2$}$$

Furthermore, since φ_i is differentiable and $\varphi_i(u) = 0$ for all $i \in I(u)$, if φ_i is not affine we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w)) + \epsilon_k \|u_k - u\| \, \eta_k(u_k - u)$$

with $\lim_{\|u_k - u\|\mapsto 0} \eta_k(u_k - u) = 0$, so if we write $\alpha_k = \|u_k - u\| \, \eta_k(u_k - u)$, we have

$$\varphi_i(u_k) = \epsilon_k((\varphi'_i)_u(v) + \delta(\varphi'_i)_u(w) + \alpha_k)$$

with $\lim_{k\mapsto\infty} \alpha_k = 0$, and since $(\varphi'_i)_u(v) \leq 0$, we obtain

$$\varphi_i(u_k) \leq \epsilon_k(\delta(\varphi'_i)_u(w) + \alpha_k) \quad \text{for all } i \in I(u) \text{ and } \varphi_i \text{ not affine}. \tag{$*_3$}$$

Equations $(*_1), (*_2), (*_3)$ show that $u_k \in U$ for k sufficiently large, where in $(*_3)$, since $(\varphi'_i)_u(w) < 0$ and $\delta > 0$, even if $\alpha_k > 0$, when $\lim_{k\mapsto\infty} \alpha_k = 0$, we will have $\delta(\varphi'_i)_u(w) + \alpha_k < 0$ for k large enough, and thus $\epsilon_k(\delta(\varphi'_i)_u(w) + \alpha_k) < 0$ for k large enough.

Since

$$\frac{u_k - u}{\|u_k - u\|} = \frac{v + \delta w}{\|v + \delta w\|}$$

for all $k \geq 0$, we conclude that $v + \delta w \in C(u)$ for $\delta > 0$ small enough. But now the sequence $(v_n)_{n\geq 0}$ given by

$$v_n = v + \epsilon_n w$$

converges to v, and for n large enough, $v_n \in C(u)$. Since by Proposition 50.1(1), the cone $C(u)$ is closed, we conclude that $v \in C(u)$. See Figure 50.10.

In all cases, we proved that $C^*(u) \subseteq C(u)$, as claimed. $\square$

In the case of m affine constraints $a_i x \leq b_i$, for some linear forms a_i and some $b_i \in \mathbb{R}$, for any point $u \in \mathbb{R}^n$ such that $a_i u = b_i$ for all $i \in I(u)$, the cone $C(u)$ consists of all $v \in \mathbb{R}^n$ such that $a_i v \leq 0$, so $u + C(u)$ consists of all points $u + v$ such that

$$a_i(u + v) \leq b_i \quad \text{for all } i \in I(u),$$

which is the cone cut out by the hyperplanes determining some face of the polyhedron defined by the m constraints $a_i x \leq b_i$.

We are now ready to prove one of the most important results of nonlinear optimization.

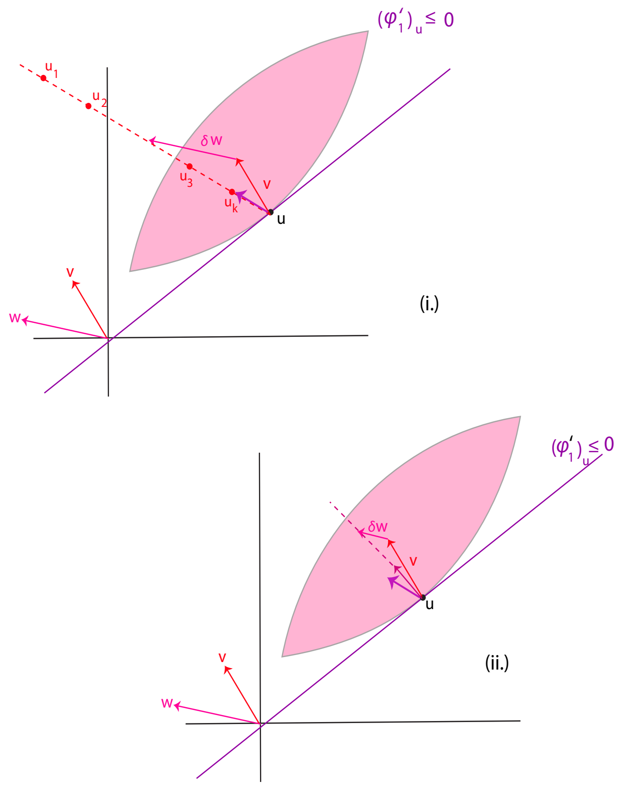


Figure 50.10: Let U be the pink lounge in $\mathbb{R}^2$. Let u satisfy the non-affine constraint $\varphi_1(u)$. Choose vectors v and w in the half space $(\varphi_1')_u \leq 0$. Figure (i.) approaches u along the line $u + t(\delta w + v)$ and shows that $v + \delta w \in C(u)$ for fixed δ. Figure (ii.) varies δ in order that the purple vectors approach v as $\delta \to \infty$.

50.3 The Karush–Kuhn–Tucker Conditions

If the domain U is defined by inequality constraints satisfying mild differentiability conditions and if the constraints at u are qualified, then there is a necessary condition for the function J to have a local minimum at $u \in U$ involving generalized Lagrange multipliers. The proof uses a version of Farkas lemma. In fact, the necessary condition stated next holds for infinite-dimensional vector spaces because there a version of Farkas lemma holding for *real* Hilbert spaces, but we will content ourselves with the version holding for finite dimensional normed vector spaces. For the more general version, see Theorem 48.12 (or Ciarlet [41], Chapter 9).

We will be using the following version of Farkas lemma.

Proposition 50.3. *(Farkas Lemma, Version I) Let A be a real $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero* *linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

We will use the version of Farkas lemma obtained by taking a contrapositive, namely: *if $yA \geq 0_n^\top$ implies $yb \geq 0$ for all linear forms $y \in (\mathbb{R}^m)^*$, then the linear system $Ax = b$ has some solution $x \geq 0$.*

Actually, it is more convenient to use a version of Farkas lemma applying to a Euclidean vector space (with an inner product denoted $\langle -, - \rangle$). This version also applies to an infinite dimensional real Hilbert space; see Theorem 48.12. Recall that in a Euclidean space V the inner product induces an isomorphism between V and V', the space of continuous linear forms on V. In our case, we need the isomorphism $\sharp$ from V' to V defined such that for every linear form $\omega \in V'$, the vector $\omega^\sharp \in V$ is uniquely defined by the equation

$$\omega(v) = \langle v, \omega^\sharp \rangle \quad \text{for all } v \in V.$$

In $\mathbb{R}^n$, the isomorphism between $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$ amounts to *transposition*: if $y \in (\mathbb{R}^n)^*$ is a linear form and $v \in \mathbb{R}^n$ is a vector, then

$$yv = v^\top y^\top.$$

The version of the Farkas–Minskowski lemma in term of an inner product is as follows.

Proposition 50.4. *(Farkas–Minkowski) Let V be a Euclidean space of finite dimension with inner product $\langle -, - \rangle$ (more generally, a Hilbert space). For any finite family $(a_1, \ldots, a_m)$ of m vectors $a_i \in V$ and any vector $b \in V$, for any $v \in V$,*

$$\textit{if } \langle a_i, v \rangle \geq 0 \textit{ for } i = 1, \ldots, m \textit{ implies that } \langle b, v \rangle \geq 0,$$

then there exist $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ such that

$$\lambda_i \geq 0 \textit{ for } i = 1, \ldots, m, \textit{ and } b = \sum_{i=1}^{m} \lambda_i a_i,$$

that is, b belong to the polyhedral cone $\mathrm{cone}(a_1, \ldots, a_m)$.

Proposition 50.4 is the special case of Theorem 48.12 which holds for real Hilbert spaces.

We can now prove the following theorem.

Theorem 50.5. *Let $\varphi_i \colon \Omega \to \mathbb{R}$ be m constraints defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), let $J \colon \Omega \to \mathbb{R}$ be some function, and let U be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0, \ 1 \leq i \leq m\}.$$

For any $u \in U$, let

$$I(u) = \{i \in \{1, \ldots, m\} \mid \varphi_i(u) = 0\},$$

and assume that the functions φ_i are differentiable at u for all $i \in I(u)$ and continuous at u for all $i \notin I(u)$. If J is differentiable at u, has a local minimum at u with respect to U, and if the constraints are qualified at u, then there exist some scalars $\lambda_i(u) \in \mathbb{R}$ for all $i \in I(u)$, such that

$$J'_u + \sum_{i \in I(u)} \lambda_i(u)(\varphi'_i)_u = 0, \quad \textit{and} \quad \lambda_i(u) \geq 0 \textit{ for all } i \in I(u).$$

The above conditions are called the Karush–Kuhn–Tucker optimality conditions. Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i \in I(u)} \lambda_i(u)\nabla(\varphi_i)_u = 0, \quad \textit{and} \quad \lambda_i(u) \geq 0 \textit{ for all } i \in I(u).$$

Proof. By Proposition 50.1(2), we have

$$J'_u(w) \geq 0 \quad \text{for all } w \in C(u), \tag{$*_1$}$$

and by Proposition 50.2(2), we have $C(u) = C^*(u)$, where

$$C^*(u) = \{v \in V \mid (\varphi'_i)_u(v) \leq 0,\ \ i \in I(u)\}, \tag{$*_2$}$$

so $(*_1)$ can be expressed as: for all $w \in V$,

$$\text{if } w \in C^*(u) \text{ then } J'_u(w) \geq 0,$$

or

$$\text{if } -(\varphi'_i)_u(w) \geq 0 \text{ for all } i \in I(u), \text{ then } J'_u(w) \geq 0. \tag{$*_3$}$$

Under the isomorphism $\sharp$, the vector $(J'_u)^\sharp$ is the gradient ∇J_u, so that

$$J'_u(w) = \langle w, \nabla J_u\rangle, \tag{$*_4$}$$

and the vector $((\varphi'_i)_u)^\sharp$ is the gradient $\nabla(\varphi_i)_u$, so that

$$(\varphi'_i)_u(w) = \langle w, \nabla(\varphi_i)_u\rangle. \tag{$*_5$}$$

Using Equations $(*_4)$ and $(*_5)$, Equation $(*_3)$ can be written as: for all $w \in V$,

$$\text{if } \langle w, -\nabla(\varphi_i)_u\rangle \geq 0 \text{ for all } i \in I(u), \text{ then } \langle w, \nabla J_u\rangle \geq 0. \tag{$*_6$}$$

By the Farkas–Minkowski proposition (Proposition 50.4), there exist some sacalars $\lambda_i(u)$ for all $i \in I(u)$, such that $\lambda_i(u) \geq 0$ and

$$\nabla J_u = \sum_{i \in I(u)} \lambda_i(u)(-\nabla(\varphi_i)_u),$$

that is

$$\nabla J_u + \sum_{i \in I(u)} \lambda_i(u)\nabla(\varphi_i)_u = 0,$$

and using the inverse of the isomorphism $\sharp$ (which is linear), we get

$$J'_u + \sum_{i\in I(u)} \lambda_i(u)(\varphi'_i)_u = 0,$$

as claimed. $\square$

Since the constraints are inequalities of the form $\varphi_i(x) \leq 0$, there is a way of expressing the Karush–Kuhn–Tucker optimality conditions, often abbreviated as *KKT conditions*, in a way that does not refer explicitly to the index set $I(u)$:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u)(\varphi'_i)_u = 0, \tag{KKT$_1$}$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m. \tag{KKT$_2$}$$

Indeed, if we have the strict inequality $\varphi_i(u) < 0$ (the constraint φ_i is inactive at u), since all the terms $\lambda_i(u)\varphi_i(u)$ are nonpositive, we must have $\lambda_i(u) = 0$; that is, we only need to consider the $\lambda_i(u)$ for all $i \in I(u)$. Yet another way to express the conditions in (KKT$_2$) is

$$\lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m. \tag{KKT$'_2$}$$

In other words, for any $i \in \{1, \ldots, m\}$, if $\varphi_i(u) < 0$, then $\lambda_i(u) = 0$; that is,

- *if the constraint φ_i is inactive at u, then $\lambda_i(u) = 0$.*

By contrapositive, if $\lambda_i(u) \neq 0$, then $\varphi_i(u) = 0$; that is,

- *if $\lambda_i(u) \neq 0$, then the constraint φ_i is active at u.*

The conditions in (KKT$'_2$) are referred to as *complementary slackness* conditions.

The scalars $\lambda_i(u)$ are often called *generalized Lagrange multipliers*. If $V = \mathbb{R}^n$, the necessary conditions of Theorem 50.5 are expressed as the following system of equations and inequalities in the unknowns $(u_1, \ldots, u_n) \in \mathbb{R}^n$ and $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_+$:

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) &= 0\\
\vdots \qquad\qquad & \vdots\\
\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_n}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) &= 0\\
\lambda_1\varphi_1(u) + \cdots + \lambda_m\varphi_m(u) &= 0\\
\varphi_1(u) &\leq 0\\
\vdots \quad & \vdots\\
\varphi_m(u) &\leq 0\\
\lambda_1, \ldots, \lambda_m &\geq 0.
\end{aligned}$$

Example 50.3. Let J, φ_1 and φ_2 be the functions defined on $\mathbb{R}$ by

$$\begin{aligned} J(x) &= x \\ \varphi_1(x) &= -x \\ \varphi_2(x) &= x - 1. \end{aligned}$$

In this case

$$U = \{x \in \mathbb{R} \mid -x \leq 0,\ x - 1 \leq 0\} = [0, 1].$$

Since the constraints are affine, they are automatically qualified for any $u \in [0, 1]$. The system of equations and inequalities shown above becomes

$$\begin{aligned} 1 - \lambda_1 + \lambda_2 &= 0 \\ -\lambda_1 x + \lambda_2(x - 1) &= 0 \\ -x &\leq 0 \\ x - 1 &\leq 0 \\ \lambda_1, \lambda_2 &\geq 0. \end{aligned}$$

The first equality implies that $\lambda_1 = 1 + \lambda_2$. The second equality then becomes

$$-(1 + \lambda_2)x + \lambda_2(x - 1) = 0,$$

which implies that $\lambda_2 = -x$. Since $0 \leq x \leq 1$, or equivalently $-1 \leq -x \leq 0$, and $\lambda_2 \geq 0$, we conclude that $\lambda_2 = 0$ and $\lambda_1 = 1$ is the solution associated with $x = 0$, the minimum of $J(x) = x$ over $[0, 1]$. Observe that the case $x = 1$ corresponds to the maximum and not a minimum of $J(x) = x$ over $[0, 1]$.

Remark: Unless the linear forms $(\varphi'_i)_u$ for $i \in I(u)$ are linearly independent, the $\lambda_i(u)$ are generally not unique. Also, if $I(u) = \emptyset$, then the KKT conditions reduce to $J'_u = 0$. This is not surprising because in this case u belongs to the relative interior of U.

If the constraints are all affine equality constraints, then the KKT conditions are a bit simpler. We will consider this case shortly.

The conditions for the qualification of nonaffine constraints are hard (if not impossible) to use in practice, because they depend on $u \in U$ and on the derivatives $(\varphi'_i)_u$. Thus it is desirable to find simpler conditions. Fortunately, this is possible if the nonaffine functions φ_i are *convex*.

Definition 50.6. Let $U \subseteq \Omega \subseteq V$ be given by

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where Ω is an open subset of the Euclidean vector space V. If the functions $\varphi_i \colon \Omega \to \mathbb{R}$ are convex, we say that the constraints are *qualified* if the following conditions hold:

(a) Either the constraints φ_i are affine for all $i = 1, \ldots, m$ and $U \neq \emptyset$, or

(b) There is some vector $v \in \Omega$ such that the following conditions hold for $i = 1, \ldots, m$:

  (i) $\varphi_i(v) \leq 0$.

  (ii) If φ_i is not affine, then $\varphi_i(v) < 0$.

The above qualification conditions are known as *Slater's conditions*.

Condition (b)(i) also implies that U has nonempty relative interior. If Ω is convex, then U is also convex. This is because for all $u, v \in \Omega$, if $u \in U$ and $v \in U$, that is $\varphi_i(u) \leq 0$ and $\varphi_i(v) \leq 0$ for $i = 1, \ldots, m$, since the functions φ_i are convex, for all $\theta \in [0, 1]$ we have

$$\begin{aligned} \varphi_i((1-\theta)u + \theta v) &\leq (1-\theta)\varphi_i(u) + \theta\varphi_i(v) && \text{since } \varphi_i \text{ is convex} \\ &\leq 0 && \text{since } 1-\theta \geq 0, \theta \geq 0, \varphi_i(u) \leq 0, \varphi_i(v) \leq 0, \end{aligned}$$

and any intersection of convex sets is convex.

It is important to observe that a *nonaffine equality constraint* $\varphi_i(u) = 0$ is *never* qualified.

Indeed, $\varphi_i(u) = 0$ is equivalent to $\varphi_i(u) \leq 0$ and $-\varphi_i(u) \leq 0$, so if these constraints are qualified and if φ_i is not affine then there is some nonzero vector $v \in \Omega$ such that both $\varphi_i(v) < 0$ and $-\varphi_i(v) < 0$, which is impossible. For this reason, *equality constraints are often assumed to be affine*.

The following theorem yields a more flexible version of Theorem 50.5 for constraints given by convex functions. If in addition, the function J is also *convex*, then the KKT conditions are also a *sufficient* condition for a local minimum.

Theorem 50.6. *Let $\varphi_i \colon \Omega \to \mathbb{R}$ be m convex constraints defined on some open convex subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), let $J \colon \Omega \to \mathbb{R}$ be some function, let U be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

and let $u \in U$ be any point such that the functions φ_i and J are differentiable at u.

(1) If J has a local minimum at u with respect to U, and if the constraints are qualified, then there exist some scalars $\lambda_i(u) \in \mathbb{R}$, such that the KKT condition hold:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u)(\varphi'_i)_u = 0$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m.$$

Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i=1}^{m} \lambda_i(u) \nabla(\varphi_i)_u = 0,$$

and

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m.$$

(2) *Conversely, if the restriction of J to U is convex and if there exist scalars $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_+$ such that the KKT conditions hold, then the function J has a (global) minimum at u with respect to U.*

Proof. (1) It suffices to prove that if the convex constraints are qualified according to Definition 50.6, then they are qualified according to Definition 50.5, since in this case we can apply Theorem 50.5.

If $v \in \Omega$ is a vector such that Condition (b) of Definition 50.6 holds and if $v \neq u$, for any $i \in I(u)$, since $\varphi_i(u) = 0$ and since φ_i is convex, by Proposition 40.11(1),

$$\varphi_i(v) \geq \varphi_i(u) + (\varphi_i')_u(v - u) = (\varphi_i')_u(v - u),$$

so if we let $w = v - u$ then

$$(\varphi_i')_u(w) \leq \varphi_i(v),$$

which shows that the nonaffine constraints φ_i for $i \in I(u)$ are qualified according to Definition 50.5, by Condition (b) of Definition 50.6.

If $v = u$, then the constraints φ_i for which $\varphi_i(u) = 0$ must be affine (otherwise, Condition (b)(ii) of Definition 50.6 would be false), and in this case we can pick $w = 0$.

(2) Let v be any arbitrary point in the convex subset U. Since $\varphi_i(v) \leq 0$ and $\lambda_i \geq 0$ for $i = 1, \ldots, m$, we have $\sum_{i=1}^m \lambda_i \varphi_i(v) \leq 0$, and using the fact that

$$\sum_{i=1}^{m} \lambda_i(u)\varphi_i(u) = 0, \quad \lambda_i(u) \geq 0, \quad i = 1, \ldots, m,$$

we have $\lambda_i = 0$ if $i \notin I(u)$ and $\varphi_i(u) = 0$ if $i \in I(u)$, so we have

$$\begin{aligned}
J(u) &\leq J(u) - \sum_{i=1}^{m} \lambda_i \varphi_i(v) \\
&\leq J(u) - \sum_{i \in I(u)} \lambda_i(\varphi_i(v) - \varphi_i(u)) && \lambda_i = 0 \text{ if } i \notin I(u),\ \varphi_i(u) = 0 \text{ if } i \in I(u) \\
&\leq J(u) - \sum_{i \in I(u)} \lambda_i(\varphi_i')_u(v - u) && \text{(by Proposition 40.11)(1)} \\
&\leq J(u) + J_u'(v - u) && \text{(by the KKT conditions)} \\
&\leq J(v) && \text{(by Proposition 40.11)(1)},
\end{aligned}$$

and this shows that u is indeed a (global) minimum of J over U. ∎

It is important to note that when *both* the constraints, the domain of definition Ω, *and* the objective function J are *convex*, if the KKT conditions hold for some $u \in U$ and some $\lambda \in \mathbb{R}^m_+$, then Theorem 50.6 implies that J has a (global) minimum at u with respect to U, *independently* of any assumption on the qualification of the constraints.

The above theorem suggests introducing the function $L\colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \sum_{i=1}^{m} \lambda_i \varphi_i(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$. The function L is called the *Lagrangian* of the *Minimization Problem* (P):

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

The KKT conditions of Theorem 50.6 imply that for any $u \in U$, if the vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ is known and if u is a minimum of J on U, then

$$\begin{aligned} \frac{\partial L}{\partial u}(u) &= 0 \\ J(u) &= L(u, \lambda). \end{aligned}$$

The Lagrangian technique "absorbs" the constraints into the new objective function L and reduces the problem of finding a constrained minimum of the function J, to the problem of finding an unconstrained minimum of the function $L(v, \lambda)$. This is the main point of Lagrangian duality which will be treated in the next section.

A case that arises often in practice is the case where the constraints φ_i are affine. If so, the m constraints $a_i x \leq b_i$ can be expressed in matrix form as $Ax \leq b$, where A is an $m \times n$ matrix whose ith row is the row vector a_i. The KKT conditions of Theorem 50.6 yield the following corollary.

Proposition 50.7. *If U is given by*

$$U = \{x \in \Omega \mid Ax \leq b\},$$

where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist some vector $\lambda \in \mathbb{R}^m$, such that

$$\begin{aligned} &\nabla J_u + A^\top \lambda = 0 \\ &\lambda_i \geq 0 \;\; \textit{and} \quad \textit{if } a_i u < b_i, \textit{ then } \lambda_i = 0, \; i = 1, \ldots, m. \end{aligned}$$

If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Another case of interest is the generalization of the minimization problem involving the affine constraints of a linear program in standard form, that is, equality constraints $Ax = b$ with $x \geq 0$, where A is an $m \times n$ matrix. In our formalism, this corresponds to the $2m + n$ constraints

$$\begin{aligned} a_i x - b_i &\leq 0, & i &= 1, \ldots, m \\ -a_i x + b_i &\leq 0, & i &= 1, \ldots, m \\ -x_j &\leq 0, & i &= 1, \ldots, n. \end{aligned}$$

In matrix form, they can be expressed as

$$\begin{pmatrix} A \\ -A \\ -I_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \leq \begin{pmatrix} b \\ -b \\ 0_n \end{pmatrix}.$$

If we introduce the generalized Lagrange multipliers λ_i^+ and λ_i^- for $i = 1, \ldots, m$ and μ_j for $j = 1, \ldots, n$, then the KKT conditions are

$$\nabla J_u + \begin{pmatrix} A^\top & -A^\top & -I_n \end{pmatrix} \begin{pmatrix} \lambda^+ \\ \lambda^- \\ \mu \end{pmatrix} = 0_n,$$

that is,

$$\nabla J_u + A^\top \lambda^+ - A^\top \lambda^- - \mu = 0,$$

and $\lambda^+, \lambda^-, \mu \geq 0$, and if $a_i u < b_i$, then $\lambda_i^+ = 0$, if $-a_i u < -b_i$, then $\lambda_i^- = 0$, and if $-u_j < 0$, then $\mu_j = 0$. But the constraints $a_i u = b_i$ hold for $i = 1, \ldots, m$, so this places no restriction on the λ_i^+ and λ_i^-, and if we write $\lambda_i = \lambda_i^+ - \lambda_i^-$, then we have

$$\nabla J_u + A^\top \lambda = \mu,$$

with $\mu_j \geq 0$, and if $u_j > 0$ then $\mu_j = 0$, for $j = 1, \ldots, n$.

Thus we proved the following proposition (which is slight generalization of Proposition 8.7.2 in Matousek and Gardner [123]).

Proposition 50.8. *If U is given by*

$$U = \{x \in \Omega \mid Ax = b, \ x \geq 0\},$$

where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist two vectors $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$, such that

$$\nabla J_u + A^\top \lambda = \mu,$$

with $\mu_j \geq 0$, and if $u_j > 0$ then $\mu_j = 0$, for $j = 1, \ldots, n$. Equivalently, there exists a vector $\lambda \in \mathbb{R}^m$ such that

$$(\nabla J_u)_j + (A^j)^\top \lambda \quad \begin{cases} = 0 & \text{if } u_j > 0 \\ \geq 0 & \text{if } u_j = 0, \end{cases}$$

where A^j is the jth column of A. If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Yet another special case that arises frequently in practice is the minimization problem involving the affine equality constraints $Ax = b$, where A is an $m \times n$ matrix, with no restriction on x. Reviewing the proof of Proposition 50.8, we obtain the following proposition.

Proposition 50.9. *If U is given by*

$$U = \{x \in \Omega \mid Ax = b\},$$

where Ω is an open convex subset of $\mathbb{R}^n$ and A is an $m \times n$ matrix, and if J is differentiable at u and J has a local minimum at u, then there exist some vector $\lambda \in \mathbb{R}^m$ such that

$$\nabla J_u + A^\top \lambda = 0.$$

Equivalently, there exists a vector $\lambda \in \mathbb{R}^m$ such that

$$(\nabla J_u)_j + (A^j)^\top \lambda = 0,$$

where A^j is the jth column of A. If the function J is convex, then the above conditions are also sufficient for J to have a minimum at $u \in U$.

Observe that in Proposition 50.9, the λ_i are just standard Lagrange multipliers, with no restriction of positivity. Thus, Proposition 50.9 is a slight generalization of Theorem 40.2 that requires A to have rank m, but in the case of equational affine constraints, this assumption is unnecessary.

Here is an application of Proposition 50.9 to the *interior point method* in linear programming.

Example 50.4. In linear programming, the interior point method using a central path uses a logarithmic barrier function to keep the solutions $x \in \mathbb{R}^n$ of the equation $Ax = b$ away from boundaries by forcing $x > 0$, which means that $x_i > 0$ for all i; see Matousek and Gardner [123] (Section 7.2). Write

$$\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n \mid x_i > 0,\ i = 1, \ldots, n\}.$$

Observe that $\mathbb{R}^n_{++}$ is open and convex. For any $\mu > 0$, we define the function f_μ defined on $\mathbb{R}^n_{++}$ by

$$f_\mu(x) = c^\top x + \mu \sum_{i=1}^{n} \ln x_i,$$

where $c \in \mathbb{R}^n$.

We would like to find necessary conditions for f_μ to have a maximum on

$$U = \{x \in \mathbb{R}^n_{++} \mid Ax = b\},$$

or equivalently to solve the following problem:

$$\begin{aligned} &\text{maximize} && f_\mu(x) \\ &\text{subject to} && \\ & && Ax = b \\ & && x > 0. \end{aligned}$$

Since maximizing f_μ is equivalent to minimizing $-f_\mu$, by Proposition 50.9, if x is an optimal of the above problem then there is some $y \in \mathbb{R}^m$ such that

$$-\nabla f_\mu(x) + A^\top y = 0.$$

Since

$$\nabla f_\mu(x) = \begin{pmatrix} c_1 + \frac{\mu}{x_1} \\ \vdots \\ c_n + \frac{\mu}{x_n} \end{pmatrix},$$

we obtain the equation

$$c + \mu \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix} = A^\top y.$$

To obtain a more convenient formulation, we define $s \in \mathbb{R}^n_{++}$ such that

$$s = \mu \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix}$$

which implies that

$$\begin{pmatrix} s_1 x_1 & \cdots & s_n x_n \end{pmatrix} = \mu \mathbf{1}_n^\top,$$

and we obtain the following necessary conditions for f_μ to have a maximum:

$$\begin{aligned} Ax &= b \\ A^\top y - s &= c \\ \begin{pmatrix} s_1 x_1 & \cdots & s_n x_n \end{pmatrix} &= \mu \mathbf{1}_n^\top \\ s, x &> 0. \end{aligned}$$

It is not hard to show that if the primal linear program with objective function $c^\top x$ and equational constraints $Ax = b$ and the dual program with objective function $b^\top y$ and inequality constraints $A^\top y \geq c$ have interior feasible points x and y, which means that $x > 0$ and $s > 0$ (where $s = A^\top y - c$), then the above system of equations has a unique solution such that x is the unique maximizer of f_μ on U; see Matousek and Gardner [123] (Section 7.2, Lemma 7.2.1).

A particularly important application of Proposition 50.9 is the situation where $\Omega = \mathbb{R}^n$.

50.4 Equality Constrained Minimization

In this section we consider the following Program (P):

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && Av = b,\ v \in \mathbb{R}^n, \end{aligned}$$

where J is a convex differentiable function and A is an $m \times n$ matrix of rank $m < n$ (the number of equality constraints is less than the number of variables, and these constraints are independent), and $b \in \mathbb{R}^m$.

According to Proposition 50.9 (with $\Omega = \mathbb{R}^n$), Program (P) has a minimum at $x \in \mathbb{R}^n$ if and only if there exist some Lagrange multipliers $\lambda \in \mathbb{R}^m$ such that the following equations hold:

$$\begin{aligned} Ax &= b && \text{(pfeasibilty)} \\ \nabla J_x + A^\top \lambda &= 0. && \text{(dfeasibility)} \end{aligned}$$

The set of linear equations $Ax = b$ is called the *primal feasibility equations* and the set of (generally nonlinear) equations $\nabla J_x + A^\top \lambda = 0$ is called the set of *dual feasibility equations*.

In general, it is impossible to solve these equations analytically, so we have to use numerical approximation procedures, most of which are variants of Newton's method. In special cases, for example if J is a quadratic functional, the dual feasibility equations are also linear, a case that we consider in more detail.

Suppose J is a convex quadratic functional of the form

$$J(x) = \frac{1}{2} x^\top P x + q^\top x + r,$$

where P is a $n \times n$ symmetric positive semidefinite matrix, $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$. In this case

$$\nabla J_x = Px + q,$$

so the feasibility equations become

$$\begin{aligned} Ax &= b \\ Px + q + A^\top \lambda &= 0, \end{aligned}$$

which in matrix form become

$$\begin{pmatrix} P & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}. \tag{KKT-eq}$$

The matrix of the linear system is usually called the *KKT-matrix*. Observe that the KKT matrix was already encountered in Proposition 42.3 with a different notation; there we had $P = A^{-1}$, $A = B^\top$, $q = b$, and $b = f$.

If the KKT matrix is invertible, then its unique solution (x^*, λ^*) yields a unique minimum x^* of Problem (P). If the KKT matrix is singular but the System (KKT-eq) is solvable, then *any solution* (x^*, λ^*) yields a minimum x^* of Problem (P).

Proposition 50.10. *If the System (KKT-eq) is not solvable, then Program (P) is unbounded below.*

Proof. We use the fact shown in Section 30.7, that a linear system $Bx = c$ has no solution iff there is some y that $B^\top y = 0$ and $y^\top c \neq 0$. By changing y to $-y$ if necessary, we may assume that $y^\top c > 0$. We apply this fact to the linear system (KKT-eq), so B is the KKT-matrix, which is symmetric, and we obtain the condition that there exist $v \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$ such that

$$Pv + A^\top \lambda = 0, \quad Av = 0, \quad -q^\top v + b^\top \lambda > 0.$$

Since the $m \times n$ matrix A has rank m and $b \in \mathbb{R}^m$, the system $Ax = b$, is solvable, so for any feasible x_0 (which means that $Ax_0 = b$), since $Av = 0$, the vector $x = x_0 + tv$ is also a feasible solution for all $t \in \mathbb{R}$. Using the fact that $Pv = -A^\top \lambda$, $v^\top P = -\lambda^\top A$, $Av = 0$, $x_0^\top A^\top = b^\top$, and P is symmetric, we have

$$\begin{aligned} J(x_0 + tv) &= J(x_0) + (v^\top P x_0 + q^\top v)t + (1/2)(v^\top P v)t^2 \\ &= J(x_0) + (x_0^\top P v + q^\top v)t - (1/2)(\lambda^\top A v)t^2 \\ &= J(x_0) + (-x_0^\top A^\top \lambda + q^\top v)t \\ &= J(x_0) - (b^\top \lambda - q^\top v)t, \end{aligned}$$

and since $-q^\top v + b^\top \lambda > 0$, the above expression goes to $-\infty$ when t goes to $+\infty$. $\square$

It is obviously important to have criteria to decide whether the KKT-matrix is invertible. There are indeed such criteria, as pointed in Boyd and Vandenberghe [29] (Chapter 10, Exercise 10.1).

Proposition 50.11. *The invertibility of the KKT-matrix*

$$\begin{pmatrix} P & A^\top \\ A & 0 \end{pmatrix}$$

is equivalent to the following conditions:

(1) For all $x \in \mathbb{R}^n$, if $Ax = 0$ with $x \neq 0$, then $x^\top P x > 0$; that is, P is positive definite on the kernel of A.

(2) The kernels of A and P only have 0 in common ($(\mathrm{Ker}\, A) \cap (\mathrm{Ker}\, P) = \{0\}$).

(3) There is some $n \times (n-m)$ matrix F such that $\mathrm{Im}(F) = \mathrm{Ker}\,(A)$ and $F^\top P F$ is symmetric positive definite.

(4) There is some symmetric positive semidefinite matrix Q such that $P + A^\top Q A$ is symmetric positive definite. In fact, $Q = I$ works.

Proof sketch. Recall from Proposition 6.19 that a square matrix B is invertible iff its kernel is reduced to $\{0\}$; equivalently, for all x, if $Bx = 0$, then $x = 0$. Assume that Condition (1) holds. We have

$$\begin{pmatrix} P & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

iff

$$Pv + A^\top w = 0, \quad Av = 0. \tag{$*$}$$

We deduce that

$$v^\top P v + v^\top A^\top w = 0,$$

and since

$$v^\top A^\top w = (Av)^\top w = 0w = 0,$$

we obtain $v^\top P v = 0$. Since Condition (1) holds, because $v \in \mathrm{Ker}\, A$, we deduce that $v = 0$. Then $A^\top w = 0$, but since the $m \times n$ matrix A has rank m, the $n \times m$ matrix $A^\top$ also has rank m, so its columns are linearly independent, and so $w = 0$. Therefore the KKT-matrix is invertible.

Conversely, assume that the KKT-matrix is invertible, yet the assumptions of Condition (1) fail. This means there is some $v \neq 0$ such that $Av = 0$ and $v^\top P v = 0$. We claim that $Pv = 0$. This is because if P is a symmetric positive semidefinite matrix, then for any v, we have $v^\top P v = 0$ iff $Pv = 0$.

If $Pv = 0$, then obviously $v^\top P v = 0$, so assume the converse, namely $v^\top P v = 0$. Since P is a symmetric positive semidefinite matrix, it can be diagonalized as

$$P = R^\top \Sigma R,$$

where R is an orthogonal matrix and Σ is a diagonal matrix

$$\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_s, 0, \ldots, 0),$$

where s is the rank of P and $\lambda_1 \geq \cdots \geq \lambda_s > 0$. Then $v^\top P v = 0$ is equivalent to

$$v^\top R^\top \Sigma R v = 0,$$

equivalently

$$(Rv)^\top \Sigma R v = 0.$$

If we write $Rv = y$, then we have

$$0 = (Rv)^\top \Sigma R v = y^\top \Sigma y = \sum_{i=1}^{s} \lambda_i y_i^2,$$

and since $\lambda_i > 0$ for $i = 1, \ldots, s$, this implies that $y_i = 0$ for $i = 1, \ldots, s$. Consequently, $\Sigma y = \Sigma R v = 0$, and so $Pv = R^\top \Sigma R v = 0$, as claimed. Since $v \neq 0$, the vector $(v, 0)$ is a nontrivial solution of Equations $(*)$, a contradiction of the invertibility assumption of the KKT-matrix.

Observe that we proved that $Av = 0$ and $Pv = 0$ iff $Av = 0$ and $v^\top P v = 0$, so we easily obtain the fact that Condition (2) is equivalent to the invertibility of the KKT-matrix. Parts (3) and (4) are left as an exercise. $\square$

In particular, if P is positive definite, then Proposition 50.11(4) applies, as we already know from Proposition 42.3. In this case, we can solve for x by elimination. We get

$$x = -P^{-1}(A^\top \lambda + q), \quad \text{where} \quad \lambda = -(AP^{-1}A^\top)^{-1}(b + AP^{-1}q).$$

In practice, we do not invert P and $AP^{-1}A^\top$. Instead, we solve the linear systems

$$\begin{aligned} Pz &= q \\ PE &= A^\top \\ (AE)\lambda &= -(b + Az) \\ Px &= -(A^\top \lambda + q). \end{aligned}$$

Observe that $(AP^{-1}A^\top)^{-1}$ is the Schur complement of P in the KKT matrix.

Since the KKT-matrix is symmetric, if it is invertible, we can convert it to $LDL^\top$ form using Proposition 8.6. This method is only practical when the problem is small or when A and P are sparse.

If the KKT-matrix is invertible but P is not, then we can use a trick involving Proposition 50.11. We find a symmetric positive semidefinite matrix Q such that $P + A^\top Q A$ is symmetric positive definite, and since a solution (v, w) of the KKT-system should have $Av = b$, we also have $A^\top QAv = A^\top Qb$, so the KKT-system is equivalent to

$$\begin{pmatrix} P + A^\top QA & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} -q + A^\top Qb \\ b \end{pmatrix},$$

and since $P + A^\top QA$ is symmetric positive definite, we can solve this system by elimination.

Another way to solve Problem (P) is to use variants of Newton's method as described in Section 49.9 dealing with equality constraints. Such methods are discussed extensively in Boyd and Vandenberghe [29] (Chapter 10, Sections 10.2-10.4).

There are two variants of this method:

(1) The first method, called *feasible start Newton method*, assumes that the starting point u_0 is feasible, which means that $Au_0 = b$. The Newton step d_{nt} is a feasible direction, which means that $Ad_{\mathrm{nt}} = 0$.

(2) The second method, called *infeasible start Newton method*, does *not* assume that the starting point u_0 is feasible, which means that $Au_0 = b$ may not hold. This method is a little more complicated than the other method.

We only briefly discuss the feasible start Newton method, leaving it to the reader to consult Boyd and Vandenberghe [29] (Chapter 10, Section 10.3) for a discussion of the infeasible start Newton method.

The Newton step d_{nt} is the solution of the linear system

$$\begin{pmatrix} \nabla^2 J(x) & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} d_{\mathrm{nt}} \\ w \end{pmatrix} = \begin{pmatrix} -\nabla J_x \\ 0 \end{pmatrix}.$$

The Newton decrement $\lambda(x)$ is defined as in Section 49.9 as

$$\lambda(x) = (d_{\mathrm{nt}}^\top \nabla^2 J(x)\, d_{\mathrm{nt}})^{1/2} = \left((\nabla J_x)^\top (\nabla^2 J(x))^{-1} \nabla J_x\right)^{1/2}.$$

Newton's method with equality constraints (with feasible start) consists of the following steps: Given a starting point $u_0 \in \mathrm{dom}(J)$ with $Au_0 = b$, and a tolerance $\epsilon > 0$ do:

repeat

(1) Compute the Newton step and decrement
$d_{\mathrm{nt},k} = -(\nabla^2 J(u_k))^{-1} \nabla J_{u_k}$ and $\lambda(u_k)^2 = (\nabla J_{u_k})^\top (\nabla^2 J(u_k))^{-1} \nabla J_{u_k}$.

(2) Stopping criterion. **quit** if $\lambda(u_k)^2/2 \leq \epsilon$.

(3) Line Search. Perform an exact or backtracking line search to find ρ_k.

(4) Update. $u_{k+1} = u_k + \rho_k d_{\mathrm{nt},k}$.

Newton's method requires that the KKT-matrix be invertible. Under some mild assumptions, Newton's method (with feasible start) converges; see Boyd and Vandenberghe [29] (Chapter 10, Section 10.2.4).

We now give an example illustrating Proposition 50.7, the *Support Vector Machine* (abbreviated as *SVM*).

50.5 Hard Margin Support Vector Machine; Version I

In this section we describe the following *classification problem*, or perhaps more accurately, *separation problem* (into two classes). Suppose we have two nonempty disjoint finite sets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$ in $\mathbb{R}^n$ (for simplicity, you may assume that these points are in the plane, that is, $n = 2$). Our goal is to find a hyperplane H of equation $w^\top x - b = 0$ (where $w \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$), such that all the blue points u_i are in one of the two open half-spaces determined by H, and all the red points v_j are in the other open half-space determined by H; see Figure 50.11.

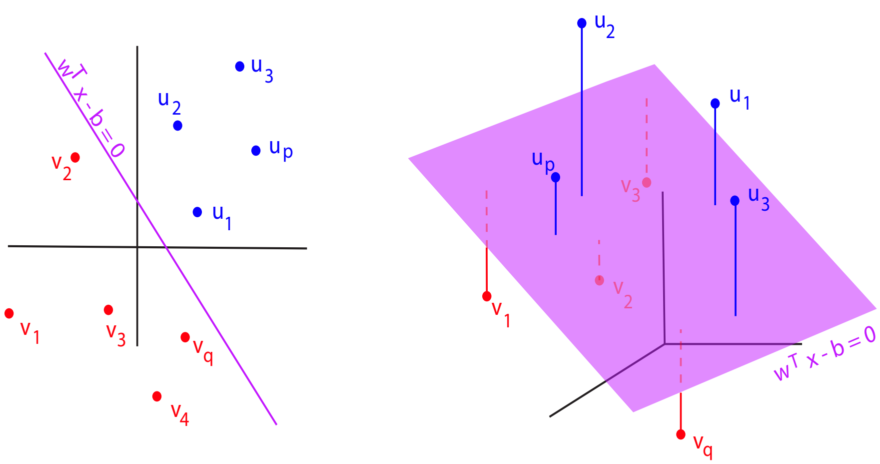


Figure 50.11: Two examples of the SVM separation problem. The left figure is SVM in $\mathbb{R}^2$, while the right figure is SVM in $\mathbb{R}^3$.

Without loss of generality, we may assume that

$$\begin{aligned} w^\top u_i - b &> 0 && \text{for } i = 1, \ldots, p \\ w^\top v_j - b &< 0 && \text{for } j = 1, \ldots, q. \end{aligned}$$

Of course, separating the blue and the red points may be impossible, as we see in Figure 50.12 for four points where the line segments (u_1, u_2) and (v_1, v_2) intersect. If a hyperplane separating the two subsets of blue and red points exists, we say that they are *linearly separable*.

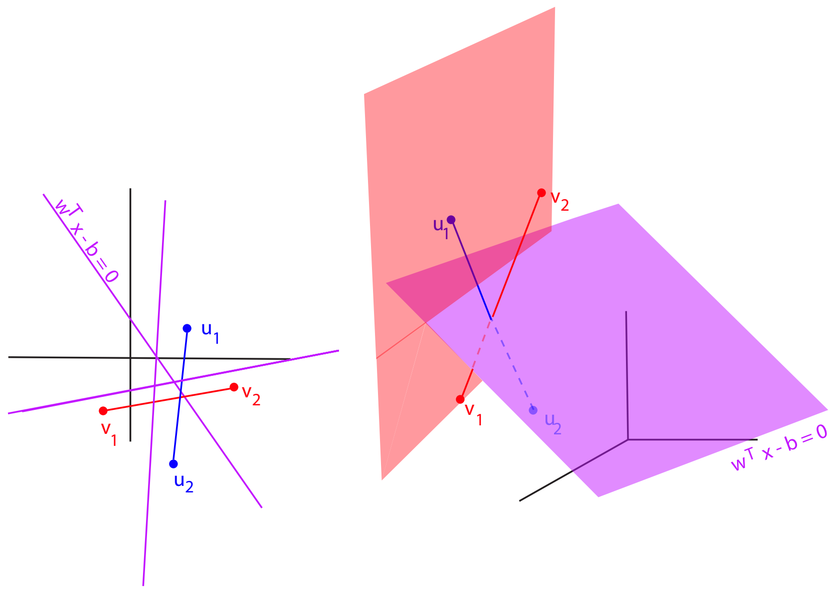


Figure 50.12: Two examples in which it is impossible to find purple hyperplanes which separate the red and blue points.

Remark: Write $m = p + q$. The reader should be aware that in machine learning the classification problem is usually defined as follows. We assign m so-called *class labels* $y_k = \pm 1$ to the data points in such a way that $y_i = +1$ for each blue point u_i, and $y_{p+j} = -1$ for each red point v_j, and we denote the m points by x_k, where $x_k = u_k$ for $k = 1, \ldots, p$ and $x_k = v_{k-p}$ for $k = p + 1, \ldots, p + q$. Then the classification constraints can be written as

$$y_k(w^\top x_k - b) > 0 \quad \text{for } k = 1, \ldots, m.$$

The set of pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ is called a set of *training data* (or *training set*).

In the sequel, we will not use the above method, and we will stick to our two subsets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$.

Since there are infinitely many hyperplanes separating the two subsets (if indeed the two subsets are linearly separable), we would like to come up with a "good" criterion for choosing such a hyperplane.

The idea that was advocated by Vapnik (see Vapnik [182]) is to consider the distances $d(u_i, H)$ and $d(v_j, H)$ from *all* the points to the hyperplane H, and to pick a hyperplane H that maximizes the smallest of these distances. In machine learning this strategy is called finding a *maximal margin hyperplane*, or *hard margin support vector machine*, which definitely sounds more impressive.

Since the distance from a point x to the hyperplane H of equation $w^\top x - b = 0$ is

$$d(x, H) = \frac{|w^\top x - b|}{\|w\|},$$

(where $\|w\| = \sqrt{w^\top w}$ is the Euclidean norm of w), it is convenient to temporarily assume that $\|w\| = 1$, so that

$$d(x, H) = |w^\top x - b|.$$

See Figure 50.13. Then with our sign convention, we have

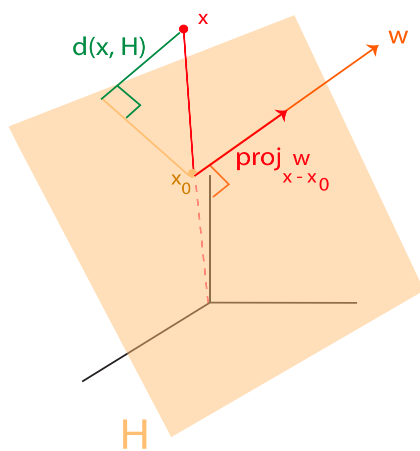


Figure 50.13: In $\mathbb{R}^3$, the distance from a point to the plane $w^\top x - b = 0$ is given by the projection onto the normal w.

$$\begin{aligned} d(u_i, H) &= w^\top u_i - b && i = 1, \ldots, p \\ d(v_j, H) &= -w^\top v_j + b && j = 1, \ldots, q. \end{aligned}$$

If we let

$$\delta = \min\{d(u_i, H),\ d(v_j, H) \mid 1 \leq i \leq p,\ 1 \leq j \leq q\},$$

then the hyperplane H should chosen so that

$$\begin{aligned} w^\top u_i - b &\geq \delta && i = 1, \ldots, p \\ -w^\top v_j + b &\geq \delta && j = 1, \ldots, q, \end{aligned}$$

and such that $\delta > 0$ is *maximal*. The distance δ is called the *margin* associated with the hyperplane H. This is indeed one way of formulating the two-class separation problem as an optimization problem with a linear objective function $J(\delta, w, b) = \delta$, and affine and quadratic constraints (SVM$_{h1}$):

$$\begin{aligned} &\text{maximize} \quad \delta \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p \\ &\qquad - w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q \\ &\qquad \|w\| \leq 1. \end{aligned}$$

Observe that the Problem (SVM$_{h1}$) has an optimal solution $\delta > 0$ iff the two subsets are linearly separable. We used the constraint $\|w\| \leq 1$ rather than $\|w\| = 1$ because the former is qualified, whereas the latter is not. But if (w, b, δ) is an optimal solution, then $\|w\| = 1$, as shown in the following proposition.

Proposition 50.12. *If (w, b, δ) is an optimal solution of Problem* (SVM$_{h1}$)*, so in particular $\delta > 0$, then we must have $\|w\| = 1$.*

Proof. First, if $w = 0$, then we get the two inequalities

$$-b \geq \delta, \quad b \geq \delta,$$

which imply that $b \leq -\delta$ and $b \geq \delta$ for some positive δ, which is impossible. But then, if $w \neq 0$ and $\|w\| < 1$, by dividing both sides of the inequalities by $\|w\| < 1$ we would obtain the better solution $(w/\|w\|, b/\|w\|, \delta/\|w\|)$, since $\|w\| < 1$ implies that $\delta/\|w\| > \delta$. ☐

We now prove that if the two subsets are linearly separable, then Problem (SVM$_{h1}$) has a unique optimal solution.

Theorem 50.13. *If two disjoint subsets of p blue points $\{u_i\}_{i=1}^p$ and q red points $\{v_j\}_{j=1}^q$ are linearly separable, then Problem* (SVM$_{h1}$) *has a unique optimal solution consisting of a hyperplane of equation $w^\top x - b = 0$ separating the two subsets with maximum margin δ. Furthermore, if we define $c_1(w)$ and $c_2(w)$ by*

$$\begin{aligned} c_1(w) &= \min_{1 \leq i \leq p} w^\top u_i \\ c_2(w) &= \max_{1 \leq j \leq q} w^\top v_j, \end{aligned}$$

then w is the unique maximum of the function

$$\rho(w) = \frac{c_1(w) - c_2(w)}{2}$$

over the convex subset U of $\mathbb{R}^n$ given by the inequalities

$$\begin{aligned} w^\top u_i - b &\geq \delta & i &= 1, \ldots, p \\ -w^\top v_j + b &\geq \delta & j &= 1, \ldots, q \\ \|w\| &\leq 1, \end{aligned}$$

and

$$b = \frac{c_1(w) + c_2(w)}{2}.$$

Proof. Our proof is adapted from Vapnik [182] (Chapter 10, Theorem 10.1). For any separating hyperplane H, since

$$\begin{aligned} d(u_i, H) &= w^\top u_i - b & i &= 1, \ldots, p \\ d(v_j, H) &= -w^\top v_j + b & j &= 1, \ldots, q, \end{aligned}$$

and since the smallest distance to H is

$$\begin{aligned} \delta &= \min\{d(u_i, H),\ d(v_j, H) \mid 1 \leq i \leq p,\ 1 \leq j \leq q\} \\ &= \min\{w^\top u_i - b,\ -w^\top v_j + b \mid 1 \leq i \leq p,\ 1 \leq j \leq q\} \\ &= \min\{\min\{w^\top u_i - b \mid 1 \leq i \leq p\}, \min\{-w^\top v_j + b \mid 1 \leq j \leq q\}\} \\ &= \min\{\min\{w^\top u_i \mid 1 \leq i \leq p\} - b\}, \min\{-w^\top v_j \mid 1 \leq j \leq q\} + b\} \\ &= \min\{\min\{w^\top u_i \mid 1 \leq i \leq p\} - b\}, -\max\{w^\top v_j \mid 1 \leq j \leq q\} + b\} \\ &= \min\{c_1(w) - b, -c_2(w) + b\}, \end{aligned}$$

in order for δ to be maximal we must have

$$c_1(w) - b = -c_2(w) + b,$$

which yields

$$b = \frac{c_1(w) + c_2(w)}{2}.$$

In this case,

$$c_1(w) - b = \frac{c_1(w) - c_2(w)}{2} = -c_2(w) + b,$$

so the maximum margin δ is indeed obtained when $\rho(w) = (c_1(w) - c_2(w))/2$ is maximal over U. Conversely, it is easy to see that any hyperplane of equation $w^\top x - b = 0$ associated with a w maximizing ρ over U and $b = (c_1(w) + c_2(w))/2$ is an optimal solution.

It remains to show that an optimal separating hyperplane exists and is unique. Since the unit ball is compact, U (as defined in Theorem 50.13) is compact, and since the function $w \mapsto \rho(w)$ is continuous, it achieves its maximum for some w_0 such that $\|w_0\| \leq 1$. Actually, we must have $\|w_0\| = 1$, since otherwise, by the reasoning used in Proposition 50.12, $w_0/\|w_0\|$ would be an even better solution. Therefore, w_0 is on the boundary of U. But ρ is a concave function (as an infimum of affine functions), so if it had two distinct maxima w_0 and w_0' with $\|w_0\| = \|w_0'\| = 1$, these would be global maxima since U is also convex, so we would have $\rho(w_0) = \rho(w_0')$ and then ρ would also have the same value along the segment (w_0, w_0') and in particular at $(w_0 + w_0')/2$, an interior point of U, a contradiction. $\square$

We can proceed with the above formulation (SVM_{h1}) but there is a way to reformulate the problem so that the constraints are all *affine*, which might be preferable since they will be *automatically qualified*.

50.6 Hard Margin Support Vector Machine; Version II

Since $\delta > 0$ (otherwise the data would not be separable into two disjoint sets), we can divide the affine constraints by δ to obtain

$$\begin{aligned} w'^{\top} u_i - b' &\geq 1 & i &= 1, \ldots, p \\ -w'^{\top} v_j + b' &\geq 1 & j &= 1, \ldots, q, \end{aligned}$$

except that now, w' is not necessarily a unit vector. To obtain the distances to the hyperplane H, we need to divide by $\|w'\|$ and then we have

$$\begin{aligned} \frac{w'^{\top} u_i - b'}{\|w'\|} &\geq \frac{1}{\|w'\|} & i &= 1, \ldots, p \\ \frac{-w'^{\top} v_j + b'}{\|w'\|} &\geq \frac{1}{\|w'\|} & j &= 1, \ldots, q, \end{aligned}$$

which means that the shortest distance from the data points to the hyperplane is $1/\|w'\|$. Therefore, we wish to maximize $1/\|w'\|$, that is, to minimize $\|w'\|$, so we obtain the following optimization Problem (SVM$_{h2}$):

Hard margin SVM (SVM$_{h2}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \\ &\qquad w^{\top} u_i - b \geq 1 \qquad i = 1, \ldots, p \\ &\qquad - w^{\top} v_j + b \geq 1 \qquad j = 1, \ldots, q. \end{aligned}$$

The objective function $J(w) = 1/2\|w\|^2$ is convex, so Proposition 50.7 applies and gives us a necessary and sufficient condition for having a minimum in terms of the KKT conditions. First observe that the trivial solution $w = 0$ is impossible, because the blue constraints would be

$$-b \geq 1,$$

that is $b \leq -1$, and the red constraints would be

$$b \geq 1,$$

but these are contradictory. **Our goal is to find w and b, and optionally, δ.** We proceed in four steps first demonstrated on the following example.

Suppose that $p = q = n = 2$, so that we have two blue points

$$u_1^{\top} = (u_{11}, u_{12}) \qquad u_2^{\top} = (u_{21}, u_{22}),$$

two red points

$$v_1^\top = (v_{11}, v_{12}) \qquad v_2^\top = (v_{21}, v_{22}),$$

and

$$w^\top = (w_1, w_2).$$

Step 1: Write the constraints in matrix form. Let

$$C = \begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix} \qquad d = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}. \tag{M}$$

The constraints become

$$C \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}. \tag{C}$$

Step 2: Write the objective function in matrix form.

$$J(w_1, w_2, b) = \frac{1}{2} \begin{pmatrix} w_1 & w_2 & b \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ b \end{pmatrix}. \tag{O}$$

Step 3: Apply Proposition 50.7 to solve for w in terms of λ and μ. We obtain

$$\begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix} + \begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{12} & v_{22} \\ 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

i.e.

$$\nabla J_{(w,b)} + C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0_3, \qquad \lambda^\top = (\lambda_1, \lambda_2),\ \ \mu^\top = (\mu_1, \mu_2).$$

Then

$$\begin{pmatrix} w_1 \\ w_2 \\ 0 \end{pmatrix} = \begin{pmatrix} u_{11} & u_{21} & -v_{11} & -v_{21} \\ u_{12} & u_{22} & -v_{12} & -v_{22} \\ -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix},$$

which implies

$$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} + \lambda_2 \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} - \mu_1 \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} - \mu_2 \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix} \tag{$*_1$}$$

with respect to

$$\mu_1 + \mu_2 - \lambda_1 - \lambda_2 = 0. \qquad (*_2)$$

Step 4: Rewrite the constraints at (C) using $(*_1)$. In particular $C\begin{pmatrix} w \\ b \end{pmatrix} \leq d$ becomes

$$\begin{pmatrix} -u_{11} & -u_{12} & 1 \\ -u_{21} & -u_{22} & 1 \\ v_{11} & v_{12} & -1 \\ v_{21} & v_{22} & -1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{21} & -v_{11} & -v_{21} & 0 \\ u_{12} & u_{22} & -v_{21} & -v_{22} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}.$$

Rewriting the previous equation in "block" format gives us

$$-\begin{pmatrix} -u_{11} & -u_{12} \\ -u_{21} & -u_{22} \\ v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} + b \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

which with the definition

$$X = \begin{pmatrix} -u_{11} & -u_{21} & v_{11} & v_{21} \\ -u_{12} & -u_{22} & v_{21} & v_{22} \end{pmatrix}$$

yields

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_2 \\ -\mathbf{1}_2 \end{pmatrix} + \mathbf{1}_4 \leq 0_4. \qquad (*_3)$$

Let us now consider the general case.

Step 1: Write the constraints in matrix form. First we rewrite the constraints as

$$\begin{aligned} -u_i^\top w + b &\leq -1 & i &= 1, \ldots, p \\ v_j^\top w - b &\leq -1 & j &= 1, \ldots, q, \end{aligned}$$

and we get the $(p+q) \times (n+1)$ matrix C and the vector $d \in \mathbb{R}^{p+q}$ given by

$$C = \begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix}, \quad d = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix},$$

so the set of inequality constraints is

$$C\begin{pmatrix} w \\ b \end{pmatrix} \leq d.$$

Step 2: The objective function in matrix form is given by

$$J(w,b) = \frac{1}{2}\begin{pmatrix} w^\top & b \end{pmatrix}\begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix}.$$

Note that the corresponding matrix is symmetric positive semidefinite, but it is *not* invertible. Thus, the function J is convex but not strictly convex. This will cause some minor trouble in finding the dual function of the problem.

Step 3: If we introduce the generalized Lagrange multipliers $\lambda \in \mathbb{R}^p$ and $\mu \in \mathbb{R}^q$, according to Proposition 50.7, the first KKT condition is

$$\nabla J_{(w,b)} + C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0_{n+1},$$

with $\lambda \geq 0, \mu \geq 0$. By the result of Example 39.5,

$$\nabla J_{(w,b)} = \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} w \\ 0 \end{pmatrix},$$

so we get

$$\begin{pmatrix} w \\ 0 \end{pmatrix} = -C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

that is,

$$\begin{pmatrix} w \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 & \cdots & u_p & -v_1 & \cdots & -v_q \\ -1 & \cdots & -1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

Consequently,

$$w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j, \tag{$*_1$}$$

and

$$\sum_{j=1}^{q} \mu_j - \sum_{i=1}^{p} \lambda_i = 0. \tag{$*_2$}$$

Step 4: Rewrite the constraint using $(*_1)$. Plugging the above expression for w into the constraints $C\begin{pmatrix} w \\ b \end{pmatrix} \leq d$ we get

$$\begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix}\begin{pmatrix} u_1 & \cdots & u_p & -v_1 & \cdots & -v_q & 0_n \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}\begin{pmatrix} \lambda \\ \mu \\ b \end{pmatrix} \leq \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix},$$

so if let X be the $n \times (p+q)$ matrix given by

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

we obtain

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \tag{$*_1'$}$$

and the above inequalities are written in matrix form as

$$\begin{pmatrix} X^\top & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} \end{pmatrix} \begin{pmatrix} -X & 0_n \\ 0_{p+q}^\top & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ b \end{pmatrix} \leq -\mathbf{1}_{p+q};$$

that is,

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \leq 0_{p+q}. \tag{$*_3$}$$

Equivalently, the ith inequality is

$$-\sum_{j=1}^p u_i^\top u_j \lambda_j + \sum_{k=1}^q u_i^\top v_k \mu_k + b + 1 \leq 0 \qquad i = 1, \ldots, p,$$

and the $(p+j)$th inequality is

$$\sum_{i=1}^p v_j^\top u_i \lambda_i - \sum_{k=1}^q v_j^\top v_k \mu_k - b + 1 \leq 0 \qquad j = 1, \ldots, q.$$

We also have $\lambda \geq 0, \mu \geq 0$. Furthermore, if the ith inequality is inactive, then $\lambda_i = 0$, and if the $(p+j)$th inequality is inactive, then $\mu_j = 0$. Since the constraints are affine and since J is convex, if we can find $\lambda \geq 0$, $\mu \geq 0$, and b such that the inequalities in $(*_3)$ are satisfied, and $\lambda_i = 0$ and $\mu_j = 0$ when the corresponding constraint is inactive, then by Proposition 50.7 we have an optimum solution.

Remark: The second KKT condition can be written as

$$\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \right) = 0;$$

that is,

$$-\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} = 0.$$

Since $(*_2)$ says that $\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j$, the second term is zero, and by $(*_1')$ we get

$$w^\top w = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j.$$

Thus, we obtain a simple expression for $\|w\|^2$ in terms of λ and μ.

The vectors u_i and v_j for which the i-th inequality is active and the $(p+j)$th inequality is active are called *support vectors*. For every vector u_i or v_j that is not a support vector, the corresponding inequality is inactive, so $\lambda_i = 0$ and $\mu_j = 0$. Thus we see that *only the support vectors contribute to a solution.* If we can *guess* which vectors u_i and v_j are support vectors, namely, those for which $\lambda_i \neq 0$ and $\mu_j \neq 0$, then for each support vector u_i we have an equation

$$-\sum_{j=1}^{p} u_i^\top u_j \lambda_j + \sum_{k=1}^{q} u_i^\top v_k \mu_k + b + 1 = 0,$$

and for each support vector v_j we have an equation

$$\sum_{i=1}^{p} v_j^\top u_i \lambda_i - \sum_{k=1}^{q} v_j^\top v_k \mu_k - b + 1 = 0,$$

with $\lambda_i = 0$ and $\mu_j = 0$ for all non-support vectors, so together with the Equation $(*_2)$ we have a linear system with an equal number of equations and variables, which is solvable if our separation problem has a solution. Thus, in principle we can find λ, μ, and b by solving a linear system.

Remark: We can first solve for λ and μ (by eliminating b), and by $(*_1)$ and since $w \neq 0$, there is a least some nonzero λ_{i_0} and thus some nonzero μ_{j_0}, so the corresponding inequalities are equations

$$\begin{aligned} -\sum_{j=1}^{p} u_{i_0}^\top u_j \lambda_j + \sum_{k=1}^{q} u_{i_0}^\top v_k \mu_k + b + 1 &= 0 \\ \sum_{i=1}^{p} v_{j_0}^\top u_i \lambda_i - \sum_{k=1}^{q} v_{j_0}^\top v_k \mu_k - b + 1 &= 0, \end{aligned}$$

so b is given in terms of λ and μ by

$$b = \frac{1}{2}(u_{i_0}^\top + v_{j_0}^\top) \left(\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{p} \mu_j v_j \right).$$

Using the dual of the Lagrangian, we can solve for λ and μ, but typically b is not determined, so we use the above method to find b.

The above nondeterministic procedure in which we guess which vectors are support vectors is not practical. We will see later that a practical method for solving for λ and μ consists in maximizing the dual of the Lagrangian.

If w is an optimal solution, then $\delta = 1/\|w\|$ is the shortest distance from the support vectors to the separating hyperplane $H_{w,b}$ of equation $w^\top x - b = 0$. If we consider the two hyperplanes $H_{w,b+1}$ and $H_{w,b-1}$ of equations

$$w^\top x - b - 1 = 0 \quad \text{and} \quad w^\top x - b + 1 = 0,$$

then $H_{w,b+1}$ and $H_{w,b-1}$ are two hyperplanes parallel to the hyperplane $H_{w,b}$ and the distance between them is 2δ. Furthermore, $H_{w,b+1}$ contains the support vectors u_i, $H_{w,b-1}$ contains the support vectors v_j, and there are no data points u_i or v_j in the open region between these two hyperplanes containing the separating hyperplane $H_{w,b}$ (called a "slab" by Boyd and Vandenberghe; see [29], Section 8.6). This situation is illustrated in Figure 50.14.

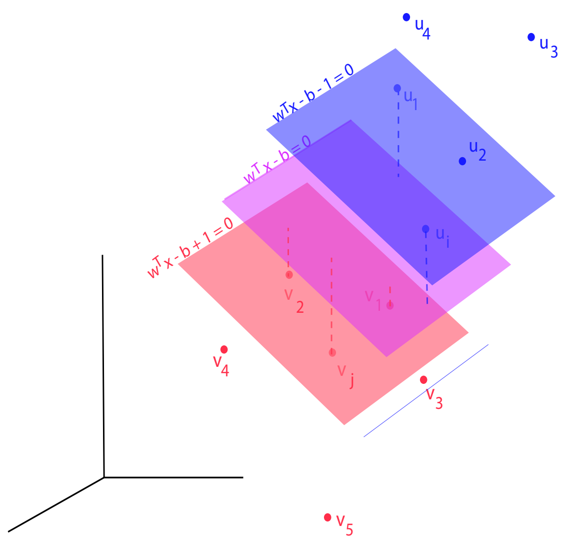


Figure 50.14: In $\mathbb{R}^3$, the solution to Hard Margin SVM$_{h2}$ is the purple plane sandwiched between the red plane $w^\top x - b + 1 = 0$ and the blue plane $w^\top x - b - 1 = 0$, each of which contains the appropriate support vectors u_i and v_j.

Even if $p = 1$ and $q = 2$, a solution is not obvious. In the plane, there are four possibilities:

(1) If u_1 is on the segment (v_1, v_2), there is no solution.

(2) If the projection h of u_1 onto the line determined by v_1 and v_2 is between v_1 and v_2, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $0 \leq \alpha \leq 1$, then it is the line parallel to $v_2 - v_1$ and equidistant to u and both v_1 and v_2, as illustrated in Figure 50.15.

(3) If the projection h of u_1 onto the line determined by v_1 and v_2 is to the right of v_2, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $\alpha > 1$, then it is the bisector of the line segment (u_1, v_2).

(4) If the projection h of u_1 onto the line determined by v_1 and v_2 is to the left of v_1, that is $h = (1-\alpha)v_1 + \alpha_2 v_2$ with $\alpha < 0$, then it is the bisector of the line segment (u_1, v_1).

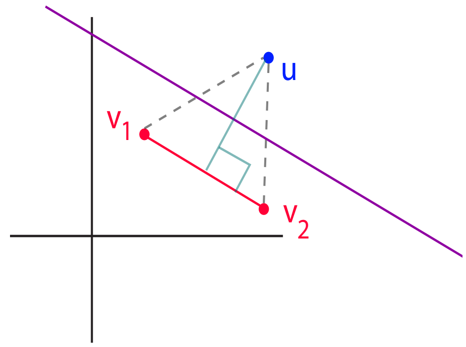


Figure 50.15: The purple line, which is the bisector of the altitude of the isosceles triangle, separates the two red points from the blue point in a manner which satisfies Hard Margin SVM$_{h2}$.

If $p = q = 1$, we can find a solution explicitly. Then $(*_2)$ yields

$$\lambda = \mu,$$

and if we guess that the constraints are active, the corresponding equality constraints are

$$\begin{aligned} -u^\top u\lambda + u^\top v\mu + b + 1 &= 0 \\ u^\top v\lambda - v^\top v\mu - b + 1 &= 0, \end{aligned}$$

so we get

$$\begin{aligned} (-u^\top u + u^\top v)\lambda + b + 1 &= 0 \\ (u^\top v - v^\top v)\lambda - b + 1 &= 0, \end{aligned}$$

Adding up the two equations we find

$$(2u^\top v - u^\top u - v^\top v)\lambda + 2 = 0,$$

that is

$$\lambda = \frac{2}{(u - v)^\top (u - v)}.$$

By subtracting the first equation from the second, we find

$$(u^\top u - v^\top v)\lambda - 2b = 0,$$

which yields

$$b = \lambda \frac{(u^\top u - v^\top v)}{2} = \frac{u^\top u - v^\top v}{(u - v)^\top (u - v)}.$$

Then by $(*_1)$ we obtain

$$w = \frac{2(u-v)}{(u-v)^\top (u-v)}.$$

We verify easily that

$$2(u_1 - v_1)x_1 + \cdots + 2(u_n - v_n)x_n = (u_1^2 + \cdots + u_n^2) - (v_1^2 + \cdots + v_n^2)$$

is the equation of the bisector hyperplane between u and v; see Figure 50.16.

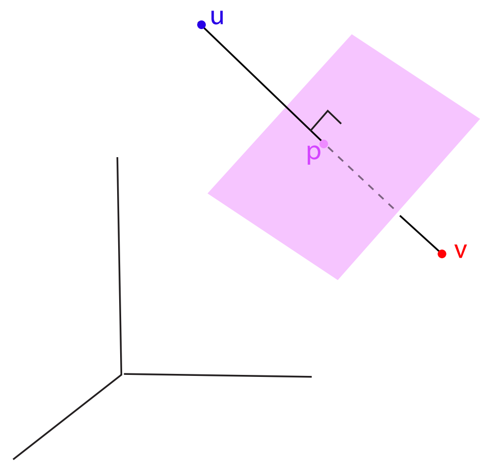


Figure 50.16: In $\mathbb{R}^3$, the solution to Hard Margin SVM$_{h2}$ for the points u and v is the purple perpendicular planar bisector of $u - v$.

In the next section we will derive the dual of the optimization problem discussed in this section. We will also consider a more flexible solution involvlng a *soft margin*.

50.7 Lagrangian Duality and Saddle Points

In this section we investigate methods to solve the *Minimization Problem (P):*

$$\begin{array}{ll} \text{minimize} & J(v) \\ \text{subject to} & \varphi_i(v) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

It turns out that under certain conditions the original Problem (P), called *primal problem*, can be solved in two stages with the help another Problem (D), called the *dual problem*. The Dual Problem (D) is a *maximization problem* involving a function G, called the *Lagrangian* *dual*, and it is obtained by *minimizing* the *Lagrangian* $L(v, \mu)$ of Problem (P) over the variable $v \in \mathbb{R}^n$, holding μ fixed, where $L \colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ is given by

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

with $\mu \in \mathbb{R}^m_+$.

The two steps of the method are:

(1) Find the dual function $\mu \mapsto G(\mu)$ explictly by solving the minimization problem of finding the minimum of $L(v, \mu)$ with respect to $v \in \Omega$, holding μ fixed. This is an unconstrained minimization problem (with $v \in \Omega$). If we are lucky, a unique minimizer u_μ such that $G(\mu) = L(u_\mu, \mu)$ can be found. We will address the issue of uniqueness later on.

(2) Solve the maximization problem of finding the maximum of the function $\mu \mapsto G(\mu)$ over all $\mu \in \mathbb{R}^m_+$. This is basically an unconstrained problem, except for the fact that $\mu \in \mathbb{R}^m_+$.

If Steps (1) and (2) are successful, under some suitable conditions on the function J and the constraints φ_i (for example, if they are convex), for any solution $\lambda \in \mathbb{R}^m_+$ obtained in Step (2), the vector u_λ obtained in Step (1) is an optimal solution of Problem (P). This is proven in Theorem 50.17.

In order to prove Theorem 50.17, which is our main result, we need two intermediate technical results of independent interest involving the notion of saddle point.

The local minima of a function $J \colon \Omega \to \mathbb{R}$ over a domain U defined by inequality constraints are saddle points of the Lagrangian $L(v, \mu)$ associated with J and the constraints φ_i. Then, under some mild hypotheses, the set of solutions of the *Minimization Problem (P)*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \le 0, \quad i = 1, \dots, m \end{aligned}$$

coincides with the set of first arguments of the saddle points of the Lagrangian

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v).$$

This is proved in Theorem 50.15. To prove Theorem 50.17, we also need Proposition 50.14, a basic property of saddle points.

Definition 50.7. Let $L\colon \Omega \times M \to \mathbb{R}$ be a function defined on a set of the form $\Omega \times M$, where Ω and M are open subsets of two normed vector spaces. A point $(u, \lambda) \in \Omega \times M$ is a *saddle point* of L if u is a minimum of the function $L(-, \lambda)\colon \Omega \to \mathbb{R}$ given by $v \mapsto L(v, \lambda)$ for all $v \in \Omega$ and λ fixed, and λ is a maximum of the function $L(u, -)\colon M \to \mathbb{R}$ given by $\mu \mapsto L(u, \mu)$ for all $\mu \in M$ and u fixed; equivalently,

$$\sup_{\mu \in M} L(u, \mu) = L(u, \lambda) = \inf_{v \in \Omega} L(v, \lambda).$$

Note that the order of the arguments u and λ is important. The second set M will be the set of generalized multipliers, and this is why we use the symbol M. Typically, $M = \mathbb{R}^m_+$.

A saddle point is often depicted as a mountain pass, which explains the terminology; see Figure 50.17. However, this is a bit misleading since other situations are possible; see Figure 50.18.

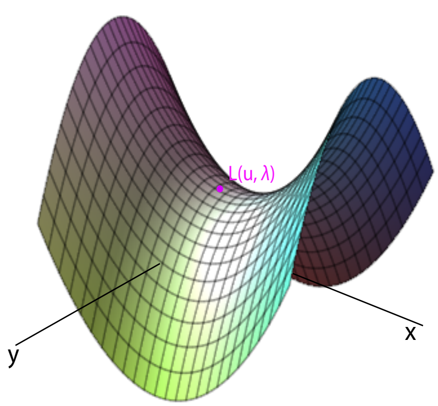


Figure 50.17: A three-dimensional rendition of a saddle point $L(u, \lambda)$ for the function $L(u, \lambda) = u^2 - \lambda^2$. The plane $x = u$ provides a maximum as the apex of a downward opening parabola, while the plane $y = \lambda$ provides a minimum as the apex of an upward opening parabola.

Proposition 50.14. *If (u, λ) is a saddle point of a function $L\colon \Omega \times M \to \mathbb{R}$, then*

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu) = L(u, \lambda) = \inf_{v \in \Omega} \sup_{\mu \in M} L(v, \mu).$$

Proof. First we prove that the following inequality always holds:

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu) \leq \inf_{v \in \Omega} \sup_{\mu \in M} L(v, \mu). \tag{$*_1$}$$

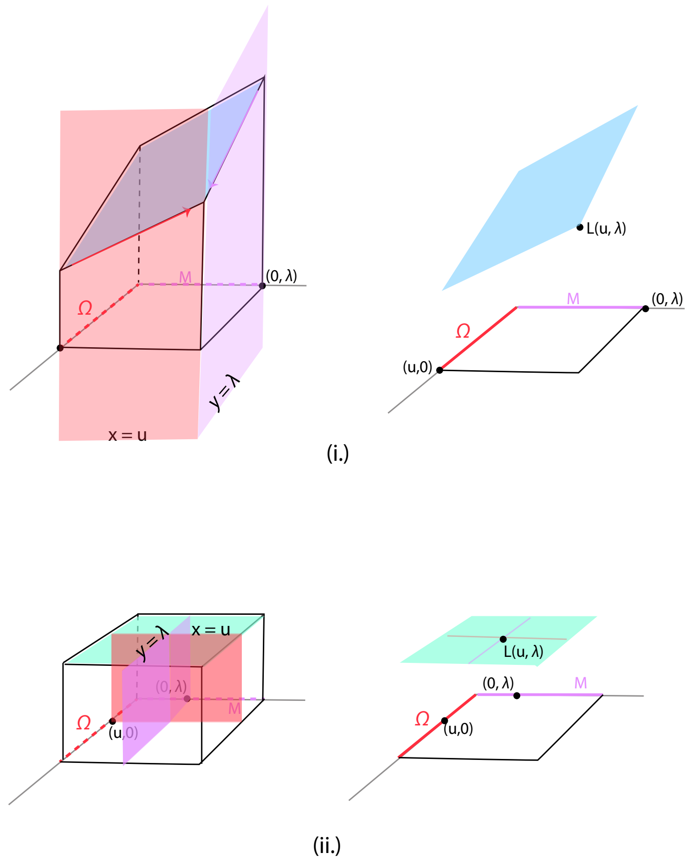


Figure 50.18: Let $\Omega = \{[t, 0, 0] \mid 0 \leq t \leq 1\}$ and $M = \{[0, t, 0] \mid 0 \leq t \leq 1\}$. In Figure (*i.*), $L(u, \lambda)$ is the blue slanted quadrilateral whose forward vertex is a saddle point. In Figure (*ii.*), $L(u, \lambda)$ is the planar green rectangle composed entirely of saddle points.

Pick any $w \in \Omega$ and any $\rho \in M$. By definition of inf (the greatest lower bound) and sup (the least upper bound), we have

$$\inf_{v \in \Omega} L(v, \rho) \leq L(w, \rho) \leq \sup_{\mu \in M} L(w, \mu).$$

The cases where $\inf_{v \in \Omega} L(v, \rho) = -\infty$ or where $\sup_{\mu \in M} L(w, \mu) = +\infty$ may arise, but this is not a problem. Since

$$\inf_{v \in \Omega} L(v, \rho) \leq \sup_{\mu \in M} L(w, \mu)$$

and the right-hand side is independent of ρ, it is an upper bound of the left-hand side for all ρ, so

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu) \leq \sup_{\mu \in M} L(w, \mu).$$

Since the left-hand side is independent of w, it is a lower bound for the right-hand side for all w, so we obtain $(*_1)$:

$$\sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu) \leq \inf_{v \in \Omega} \sup_{\mu \in M} L(v, \mu).$$

To obtain the reverse inequality, we use the fact that (u, λ) is a saddle point, so

$$\inf_{v \in \Omega} \sup_{\mu \in M} L(v, \mu) \leq \sup_{\mu \in M} L(u, \mu) = L(u, \lambda)$$

and

$$L(u, \lambda) = \inf_{v \in \Omega} L(v, \lambda) \leq \sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu),$$

and these imply that

$$\inf_{v \in \Omega} \sup_{\mu \in M} L(v, \mu) \leq \sup_{\mu \in M} \inf_{v \in \Omega} L(v, \mu), \tag{$*_2$}$$

as desired. □

We now return to our main Minimization Problem (P):

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where $J \colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i \colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V).

Definition 50.8. The *Lagrangian* of the Minimization Problem (P) defined above is the function $L \colon \Omega \times \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

with $\mu = (\mu_1, \ldots, \mu_m)$. The numbers μ_i are called *generalized Lagrange multipliers*.

The following theorem shows that under some suitable conditions, every solution u of the Problem (P) is the first argument of a saddle point (u, λ) of the Lagrangian L, and conversely, if (u, λ) is a saddle point of the Lagrangian L, then u is a solution of the Problem (P).

Theorem 50.15. *Consider Problem (P) defined above where $J \colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i \colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V). The following facts hold.*

(1) If $(u, \lambda) \in \Omega \times \mathbb{R}^m_+$ is a saddle point of the Lagrangian L associated with Problem (P), then $u \in U$, u is a solution of Problem (P), and $J(u) = L(u, \lambda)$.

(2) *If Ω is convex (open), if the functions φ_i $(1 \leq i \leq m)$ and J are convex and differentiable at the point $u \in U$, if the constraints are qualified, and if $u \in U$ is a minimum of Problem (P), then there exists some vector $\lambda \in \mathbb{R}^m_+$ such that the pair $(u, \lambda) \in \Omega \times \mathbb{R}^m_+$ is a saddle point of the Lagrangian L.*

Proof. (1) Since (u, λ) is a saddle point of L we have $\sup_{\mu \in \mathbb{R}^m_+} L(u, \mu) = L(u, \lambda)$ which implies that $L(u, \mu) \leq L(u, \lambda)$ for all $\mu \in \mathbb{R}^m_+$, which means that

$$J(u) + \sum_{i=1}^m \mu_i \varphi_i(u) \leq J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u),$$

that is,

$$\sum_{i=1}^m (\mu_i - \lambda_i)\varphi_i(u) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+.$$

If we let each μ_i be large enough, then $\mu_i - \lambda_i > 0$, and if we had $\varphi_i(u) > 0$, then the term $(\mu_i - \lambda_i)\varphi_i(u)$ could be made arbitrarily large and positive, so we conclude that $\varphi_i(u) \leq 0$ for $i = 1, \ldots, m$, and consequently, $u \in U$. For $\mu = 0$, we conclude that $\sum_{i=1}^m \lambda_i \varphi_i(u) \geq 0$. However, since $\lambda_i \geq 0$ and $\varphi_i(u) \leq 0$, (since $u \in U$), we have $\sum_{i=1}^m \lambda_i \varphi_i(u) \leq 0$. Combining these two inequalities shows that

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0. \tag{$*_1$}$$

This shows that $J(u) = L(u, \lambda)$. Since the inequality $L(u, \lambda) \leq L(v, \lambda)$ is

$$J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) \leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v),$$

by $(*_1)$ we obtain

$$\begin{aligned} J(u) &\leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) && \text{for all } v \in \Omega \\ &\leq J(v) && \text{for all } v \in U \text{ (since } \varphi_i(v) \leq 0 \text{ and } \lambda_i \geq 0), \end{aligned}$$

which shows that u is a minimum of J on U.

(2) The hypotheses required to apply Theorem 50.6(1) are satisfied. Consequently if $u \in U$ is a solution of Problem (P), then there exists some vector $\lambda \in \mathbb{R}^m_+$ such that the KKT conditions hold:

$$J'(u) + \sum_{i=1}^m \lambda_i (\varphi'_i)_u = 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i \varphi_i(u) = 0.$$

The second equation yields

$$L(u, \mu) = J(u) + \sum_{i=1}^{m} \mu_i \varphi_i(u) \leq J(u) = J(u) + \sum_{i=1}^{m} \lambda_i \varphi_i(u) = L(u, \lambda),$$

that is,

$$L(u, \mu) \leq L(u, \lambda) \quad \text{for all } \mu \in \mathbb{R}_+^m \tag{$*_2$}$$

(since $\varphi_i(u) \leq 0$ as $u \in U$), and since the function $v \mapsto J(v) + \sum_{i=1} \lambda_i \varphi_i(v) = L(v, \lambda)$ is convex as a sum of convex functions, by Theorem 40.13(4), the first equation is a sufficient condition for the existence of minimum. Consequently,

$$L(u, \lambda) \leq L(v, \lambda) \quad \text{for all } v \in \Omega, \tag{$*_3$}$$

and $(*_2)$ and $(*_3)$ show that (u, λ) is a saddle point of L. $\square$

To recap what we just proved, under some mild hypotheses, the set of solutions of the Minimization Problem (P)

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

coincides with the set of first arguments of the saddle points of the Lagrangian

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

and for any optimum $u \in U$ of Problem (P), we have $J(u) = L(u, \lambda)$.

Therefore, if we knew some particular second argument λ of these saddle points, then the *constrained* Problem (P) would be replaced by the *unconstrained* Problem (P_λ):

$$\begin{aligned} &\text{find } u_\lambda \in \Omega \text{ such that} \\ &L(u_\lambda, \lambda) = \inf_{v \in \Omega} L(v, \lambda). \end{aligned}$$

How do we find such an element $\lambda \in \mathbb{R}_+^m$?

For this, remember that for a saddle point (u_λ, λ), by Proposition 50.14, we have

$$L(u_\lambda, \lambda) = \inf_{v \in \Omega} L(v, \lambda) = \sup_{\mu \in \mathbb{R}_+^m} \inf_{v \in \Omega} L(v, \mu),$$

so we are naturally led to introduce the function $G \colon \mathbb{R}_+^m \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}_+^m,$$

and then λ will be a solution of the problem

$$\begin{aligned} & \text{find } \lambda \in \mathbb{R}^m_+ \text{ such that} \\ & G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu), \end{aligned}$$

which is equivalent to the *Maximization Problem (D):*

$$\begin{aligned} & \text{maximize} \quad G(\mu) \\ & \text{subject to} \quad \mu \in \mathbb{R}^m_+. \end{aligned}$$

Definition 50.9. Given the Minimization Problem (P)

$$\begin{aligned} & \text{minimize} \quad J(v) \\ & \text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where $J\colon \Omega \to \mathbb{R}$ and the constraints $\varphi_i\colon \Omega \to \mathbb{R}$ are some functions defined on some open subset Ω of some finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), the function $G\colon \mathbb{R}^m_+ \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}^m_+,$$

is called the *Lagrange dual function* (or simply *dual function*). The *Problem (D)*

$$\begin{aligned} & \text{maximize} \quad G(\mu) \\ & \text{subject to} \quad \mu \in \mathbb{R}^m_+ \end{aligned}$$

is called the *Lagrange dual problem*. The Problem (P) is often called the *primal problem*, and (D) is the *dual problem*. The variable μ is called the *dual variable*. The variable $\mu \in \mathbb{R}^m_+$ is said to be *dual feasible* if $G(\mu)$ is defined (not $-\infty$). If $\lambda \in \mathbb{R}^m_+$ is a maximum of G, then we call it a *dual optimal* or an *optimal Lagrange multiplier*.

Since

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v),$$

the function $G(\mu) = \inf_{v \in \Omega} L(v, \mu)$ is the pointwise infimum of some affine functions of μ, so it is *concave*, even if the φ_i are not convex. One of the main advantages of the dual problem over the primal problem is that it is a *convex optimization problem*, since we wish to maximize a concave objective function G (thus minimize $-G$, a convex function), and the constraints $\mu \geq 0$ are convex. In a number of practical situations, the dual function G can indeed be computed.

To be perfectly rigorous, we should mention that the dual function G is actually a *partial function*, because it takes the value $-\infty$ when the map $v \mapsto L(v, \mu)$ is unbounded below.

Example 50.5. Consider the Linear Program (P)

$$\begin{aligned} &\text{minimize} \quad c^\top v \\ &\text{subject to} \quad Av \leq b, \;\; v \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix. The constraints $v \geq 0$ are rewritten as $-v_i \leq 0$, so we introduce Lagrange multipliers $\mu \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^n_+$, and we have the Lagrangian

$$\begin{aligned} L(v, \mu, \nu) &= c^\top v + \mu^\top (Av - b) - \nu^\top v \\ &= -b^\top \mu + (c + A^\top \mu - \nu)^\top v. \end{aligned}$$

The linear function $v \mapsto (c + A^\top \mu - \nu)^\top v$ is unbounded below unless $c + A^\top \mu - \nu = 0$, so the dual function $G(\mu, \nu) = \inf_{v \in \mathbb{R}^n} L(v, \mu, \nu)$ is given for all $\mu \geq 0$ and $\nu \geq 0$ by

$$G(\mu, \nu) = \begin{cases} -b^\top \mu & \text{if } A^\top \mu - \nu + c = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The domain of G is a proper subset of $\mathbb{R}^m_+ \times \mathbb{R}^n_+$.

Observe that the value $G(\mu, \nu)$ of the function G, when it is defined, is independent of the second argument ν. Since we are interested in maximizing G, this suggests introducing the function $\widehat{G}$ of the single argument μ given by

$$\widehat{G}(\mu) = -b^\top \mu,$$

which is defined for all $\mu \in \mathbb{R}^m_+$.

Of course, $\sup_{\mu \in \mathbb{R}^m_+} \widehat{G}(\mu)$ and $\sup_{(\mu,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^n_+} G(\mu, \, \nu)$ are generally different, but note that $\widehat{G}(\mu) = G(\mu, \nu)$ iff there is some $\nu \in \mathbb{R}^n_+$ such that $A^\top \mu - \nu + c = 0$ iff $A^\top \mu + c \geq 0$. Therefore, finding $\sup_{(\mu,\nu) \in \mathbb{R}^m_+ \times \mathbb{R}^n_+} G(\mu, \nu)$ is equivalent to the constrained Problem (D_1)

$$\begin{aligned} &\text{maximize} \quad -b^\top \mu \\ &\text{subject to} \quad A^\top \mu \geq -c, \;\; \mu \geq 0. \end{aligned}$$

The above problem is the dual of the Linear Program (P).

In summary, the dual function G of a primary Problem (P) often contains hidden inequality constraints that define its domain, and sometimes it is possible to make these domain constraints $\psi_1(\mu) \leq 0, \ldots, \psi_p(\mu) \leq 0$ explicit, to define a new function $\widehat{G}$ that depends only on $q < m$ of the variables μ_i and is defined for all values $\mu_i \geq 0$ of these variables, and to replace the Maximization Problem (D), find $\sup_{\mu \in \mathbb{R}^m_+} G(\mu)$, by the constrained Problem (D_1)

$$\begin{aligned} &\text{maximize} \quad \widehat{G}(\mu) \\ &\text{subject to} \quad \psi_i(\mu) \leq 0, \quad i = 1, \ldots, p. \end{aligned}$$

Problem (D_1) is different from the Dual Program (D), but it is equivalent to (D) as a maximization problem.

50.8 Weak and Strong Duality

Another important property of the dual function G is that it provides a *lower bound* on the value of the objective function J. Indeed, we have

$$G(\mu) \leq L(u,\mu) \leq J(u) \quad \text{for all } u \in U \text{ and all } \mu \in \mathbb{R}^m_+, \tag{$\dagger$}$$

since $\mu \geq 0$ and $\varphi_i(u) \leq 0$ for $i = 1, \ldots, m$, so

$$G(\mu) = \inf_{v \in \Omega} L(v,\mu) \leq L(u,\mu) = J(u) + \sum_{i=1}^m \mu_i \varphi_i(u) \leq J(u).$$

If the Primal Problem (P) has a minimum denoted p^* and the Dual Problem (D) has a maximum denoted d^*, then the above inequality implies that

$$d^* \leq p^* \tag{$\dagger_w$}$$

known as *weak duality*. Equivalently, for every optimal solution λ^* of the dual problem and every optimal solution u^* of the primal problem, we have

$$G(\lambda^*) \leq J(u^*). \tag{$\dagger_{w'}$}$$

In particular, if $p^* = -\infty$, which means that the primal problem is unbounded below, then the dual problem is unfeasible. Conversely, if $d^* = +\infty$, which means that the dual problem is unbounded above, then the primal problem is unfeasible.

Definition 50.10. The difference $p^* - d^* \geq 0$ is called the *optimal duality gap*. If the duality gap is zero, that is, $p^* = d^*$, then we say that *strong duality* holds.

Even when the duality gap is strictly positive, the inequality $(\dagger_w)$ can be helpful to find a lower bound on the optimal value of a primal problem that is difficult to solve, since the dual problem is *always* convex.

If the primal problem and the dual problem are feasible and if the optimal values p^* and d^* are finite and $p^* = d^*$ (no duality gap), then the complementary slackness conditions hold for the inequality constraints.

Proposition 50.16. *(Complementary Slackness) Given the Minimization Problem (P)*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

and its Dual Problem (D)

$$\begin{aligned} &\text{maximize} && G(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+, \end{aligned}$$

if both (P) and (D) are feasible, $u \in U$ is an optimal solution of (P), $\lambda \in \mathbb{R}^m_+$ is an optimal solution of (D), and $J(u) = G(\lambda)$, then

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0.$$

In other words, if the constraint φ_i is inactive at u, then $\lambda_i = 0$.

Proof. Since $J(u) = G(\lambda)$ we have

$$\begin{aligned}
J(u) &= G(\lambda) \\
&= \inf_{v \in \Omega} \left(J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) \right) && \text{by definition of } G \\
&\leq J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) && \text{the greatest lower bound is a lower bound} \\
&\leq J(u) && \text{since } \lambda_i \geq 0, \varphi_i(u) \leq 0.
\end{aligned}$$

which implies that $\sum_{i=1}^m \lambda_i \varphi_i(u) = 0$. ☐

Going back to Example 50.5, we see that weak duality says that for any feasible solution u of the Primal Problem (P), that is, some $u \in \mathbb{R}^n$ such that

$$Au \leq b, \ \ u \geq 0,$$

and for any feasible solution $\mu \in \mathbb{R}^m$ of the Dual Problem (D_1), that is,

$$A^\top \mu \geq -c, \ \ \mu \geq 0,$$

we have

$$-b^\top \mu \leq c^\top u.$$

Actually, if u and λ are optimal, then we know from Theorem 47.7 that strong duality holds, namely $-b^\top \mu = c^\top u$, but the proof of this fact is nontrivial.

The following theorem establishes a link between the solutions of the Primal Problem (P) and those of the Dual Problem (D). It also gives sufficient conditions for the duality gap to be zero.

Theorem 50.17. *Consider the Minimization Problem (P):*

$$\begin{aligned}
&\text{minimize} && J(v) \\
&\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m,
\end{aligned}$$

where the functions J and φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V).

(1) *Suppose the functions* $\varphi_i \colon \Omega \to \mathbb{R}$ *are continuous, and that for every* $\mu \in \mathbb{R}^m_+$*, the Problem* (P_μ)*:*

$$\begin{aligned} &\text{minimize} && L(v, \mu) \\ &\text{subject to} && v \in \Omega, \end{aligned}$$

has a unique solution u_μ*, so that*

$$L(u_\mu, \mu) = \inf_{v \in \Omega} L(v, \mu) = G(\mu),$$

and the function $\mu \mapsto u_\mu$ *is continuous (on* $\mathbb{R}^m_+$*). Then the function* G *is differentiable for all* $\mu \in \mathbb{R}^m_+$*, and*

$$G'_\mu(\xi) = \sum_{i=1}^m \xi_i \varphi_i(u_\mu) \quad \textit{for all } \xi \in \mathbb{R}^m.$$

If λ *is any solution of Problem* (D)*:*

$$\begin{aligned} &\text{maximize} && G(\mu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+, \end{aligned}$$

then the solution u_λ *of the corresponding problem* (P_λ) *is a solution of Problem* (P)*.*

(2) *Assume Problem* (P) *has some solution* $u \in U$*, and that* Ω *is convex (open), the functions* φ_i $(1 \leq i \leq m)$ *and* J *are convex and differentiable at* u*, and that the constraints are qualified. Then Problem* (D) *has a solution* $\lambda \in \mathbb{R}^m_+$*, and* $J(u) = G(\lambda)$*; that is, the duality gap is zero.*

Proof. (1) Our goal is to prove that for any solution λ of Problem (D), the pair (u_λ, λ) is a saddle point of L. By Theorem 50.15(1), the point $u_\lambda \in U$ is a solution of Problem (P).

Since $\lambda \in \mathbb{R}^m_+$ is a solution of Problem (D), by definition of $G(\lambda)$ and since u_λ satisfies Problem (P_λ), we have

$$G(\lambda) = \inf_{v \in \Omega} L(v, \lambda) = L(u_\lambda, \lambda),$$

which is one of the two equations characterizing a saddle point. In order to prove the second equation characterizing a saddle point,

$$\sup_{\mu \in \mathbb{R}^m_+} L(u_\mu, \mu) = L(u_\lambda, \lambda),$$

we will begin by proving that the function G is differentiable for all $\mu \in \mathbb{R}^m_+$, in order to be able to apply Theorem 40.9 to conclude that since G has a maximum at λ, that is, $-G$ has minimum at λ, then $-G'_\lambda(\mu - \lambda) \geq 0$ for all $\mu \in \mathbb{R}^m_+$. In fact, we prove that

$$G'_\mu(\xi) = \sum_{i=1}^m \xi_i \varphi_i(u_\mu) \quad \text{for all } \xi \in \mathbb{R}^m. \tag{$*_{\text{deriv}}$}$$

Consider any two points μ and $\mu + \xi$ in $\mathbb{R}^m_+$. By definition of u_μ we have

$$L(u_\mu, \mu) \leq L(u_{\mu+\xi}, \mu),$$

which means that

$$J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_\mu) \leq J(u_{\mu+\xi}) + \sum_{i=1}^m \mu_i \varphi_i(u_{\mu+\xi}), \tag{$*_1$}$$

and since $G(\mu) = L(u_\mu, \mu) = J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_\mu)$ and $G(\mu + \xi) = L(u_{\mu+\xi}, \mu + \xi) = J(u_{\mu+\xi}) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi})$, we have

$$G(\mu + \xi) - G(\mu) = J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m \mu_i \varphi_i(u_\mu). \tag{$*_2$}$$

Since $(*_1)$ can be written as

$$0 \leq J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m \mu_i \varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m \mu_i \varphi_i(u_\mu),$$

by adding $\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi})$ to both sides of the above inequality and using $(*_2)$ we get

$$\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi}) \leq G(\mu + \xi) - G(\mu). \tag{$*_3$}$$

By definition of $u_{\mu+\xi}$ we have

$$L(u_{\mu+\xi}, \mu + \xi) \leq L(u_\mu, \mu + \xi),$$

which means that

$$J(u_{\mu+\xi}) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) \leq J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_\mu). \tag{$*_4$}$$

This can be written as

$$J(u_{\mu+\xi}) - J(u_\mu) + \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_{\mu+\xi}) - \sum_{i=1}^m (\mu_i + \xi_i)\varphi_i(u_\mu) \leq 0,$$

and by adding $\sum_{i=1}^m \xi_i \varphi_i(u_\mu)$ to both sides of the above inequality and using $(*_2)$ we get

$$G(\mu + \xi) - G(\mu) \leq \sum_{i=1}^m \xi_i \varphi_i(u_\mu). \tag{$*_5$}$$

By putting $(*_3)$ and $(*_5)$ together we obtain

$$\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi}) \leq G(\mu + \xi) - G(\mu) \leq \sum_{i=1}^m \xi_i \varphi_i(u_\mu). \tag{$*_6$}$$

Consequently there is some $\theta \in [0, 1]$ such that

$$\begin{aligned} G(\mu + \xi) - G(\mu) &= (1 - \theta)\bigg(\sum_{i=1}^m \xi_i \varphi_i(u_\mu)\bigg) + \theta\bigg(\sum_{i=1}^m \xi_i \varphi_i(u_{\mu+\xi})\bigg) \\ &= \sum_{i=1}^m \xi_i \varphi_i(u_\mu) + \theta\bigg(\sum_{i=1}^m \xi_i(\varphi_i(u_{\mu+\xi}) - \varphi_i(u_\mu))\bigg). \end{aligned}$$

Since by hypothesis the functions $\mu \mapsto u_\mu$ (from $\mathbb{R}^m_+$ to Ω) and $\varphi_i \colon \Omega \to \mathbb{R}$ are continuous, for any $\mu \in \mathbb{R}^m_+$ we can write

$$G(\mu + \xi) - G(\mu) = \sum_{i=1}^m \xi_i \varphi_i(u_\mu) + \|\xi\| \, \epsilon(\xi), \quad \text{with } \lim_{\xi \mapsto 0} \epsilon(\xi) = 0, \tag{$*_7$}$$

for any $\| \, \|$ norm on $\mathbb{R}^m$. Equation $(*_7)$ show that G is differentiable for any $\mu \in \mathbb{R}^m_+$, and that

$$G'_\mu(\xi) = \sum_{i=1}^m \xi_i \varphi_i(u_\mu) \quad \text{for all } \xi \in \mathbb{R}^m. \tag{$*_8$}$$

Actually there is a small problem, namely that the notion of derivative was defined for a function defined on an *open* set, but $\mathbb{R}^m_+$ is not open. The difficulty only arises to ensure that the derivative is unique, but in our case we have a unique expression for the derivative so there is no problem as far as defining the derivative. There is still a potential problem, which is that we would like to apply Theorem 40.9 to conclude that since G has a maximum at λ, that is, $-G$ has a minimum at λ, then

$$-G'_\lambda(\mu - \lambda) \geq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_9$}$$

but the hypotheses of Theorem 40.9 require the domain of the function to be open. Fortunately, close examination of the proof of Theorem 40.9 shows that the proof still holds with $U = \mathbb{R}^m_+$. Therefore, $(*_8)$ holds, Theorem 40.9 is valid, which in turn implies

$$G'_\lambda(\mu - \lambda) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_{10}$}$$

which, using the expression for G'_λ given in $(*_8)$ gives

$$\sum_{i=1}^m \mu_i \varphi_i(u_\lambda) \leq \sum_{i=1}^m \lambda_i \varphi_i(u_\lambda), \quad \text{for all } \mu \in \mathbb{R}^m_+. \tag{$*_{11}$}$$

As a consequence of $(*_{11})$, we obtain

$$\begin{aligned}L(u_\lambda, \mu) &= J(u_\lambda) + \sum_{i=1}^m \mu_i \varphi_i(u_\lambda) \\ &\leq J(u_\lambda) + \sum_{i=1}^m \lambda_i \varphi_i(u_\lambda) = L(u_\lambda, \lambda),\end{aligned}$$

for all $\mu \in \mathbb{R}^m_+$, that is,

$$L(u_\lambda, \mu) \leq L(u_\lambda, \lambda), \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$*_{12}$}$$

which implies the second inequality

$$\sup_{\mu \in \mathbb{R}^m_+} L(u_\mu, \mu) = L(u_\lambda, \lambda)$$

stating that (u_λ, λ) is a saddle point. Therefore, (u_λ, λ) is a saddle point of L, as claimed.

(2) The hypotheses are exactly those required by Theorem 50.15(2), thus there is some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point of the Lagrangian L, and by Theorem 50.15(1) we have $J(u) = L(u, \lambda)$. By Proposition 50.14, we have

$$J(u) = L(u, \lambda) = \inf_{v \in \Omega} L(v, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{v \in \Omega} L(v, \mu),$$

which can be rewritten as

$$J(u) = G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu).$$

In other words, Problem (D) has a solution, and $J(u) = G(\lambda)$. □

Remark: Note that Theorem 50.17(2) could have already be obtained as a consequence of Theorem 50.15(2), but the dual function G was not yet defined. If (u, λ) is a saddle point of the Lagrangian L (defined on $\Omega \times \mathbb{R}^m_+$), then by Proposition 50.14, the vector λ is a solution of Problem (D). Conversely, under the hypotheses of Part (1) of Theorem 50.17, if λ is a solution of Problem (D), then (u_λ, λ) is a saddle point of L. *Consequently, under the above hypotheses, the set of solutions of the Dual Problem (D) coincide with the set of second arguments λ of the saddle points (u, λ) of L.* In some sense, this result is the "dual" of the result stated in Theorem 50.15, namely that the set of solutions of Problem (P) coincides with set of first arguments u of the saddle points (u, λ) of L.

Informally, in Theorem 50.17(1), the hypotheses say that if $G(\mu)$ can be "computed nicely," in the sense that there is a unique minimizer u_μ of $L(v, \mu)$ (with $v \in \Omega$) such that $G(\mu) = L(u_\mu, \mu)$, and if a maximizer λ of $G(\mu)$ (with $\mu \in \mathbb{R}^m_+$) can be determined, then u_λ yields the minimum value of J, that is, $p^* = J(u_\lambda)$. If the constraints are qualified and if the functions J and φ_i are convex and differentiable, then since the KKT conditions hold, the duality gap is zero; that is,

$$G(\lambda) = L(u_\lambda, \lambda) = J(u_\lambda).$$

Example 50.6. Going back to Example 50.5 where we considered the linear program (P)

$$\begin{aligned} &\text{minimize} \quad c^\top v \\ &\text{subject to} \quad Av \leq b, \;\; v \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, the Lagrangian $L(v, \mu, \nu)$ is given by

$$L(v, \mu, \nu) = -b^\top \mu + (c + A^\top \mu - \nu)^\top v,$$

and we found that the dual function $G(\mu, \nu) = \inf_{v \in \mathbb{R}^n} L(v, \mu, \nu)$ is given for all $\mu \geq 0$ and $\nu \geq 0$ by

$$G(\mu, \nu) = \begin{cases} -b^\top \mu & \text{if } A^\top \mu - \nu + c = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The hypotheses of Theorem 50.17(1) certainly fail since there are infinitely $u_{\mu,\nu} \in \mathbb{R}^n$ such that $G(\mu, \nu) = \inf_{v \in \mathbb{R}^n} L(v, \mu, \nu) = L(u_{\mu,\nu}, \mu, \nu)$. Therefore, the dual function G is no help in finding a solution of the Primal Problem (P). As we saw earlier, if we consider the modified dual Problem (D_1) then strong duality holds, but this *does not* follow from Theorem 50.17, and a different proof is required.

Thus, we have the somewhat counter-intuitive situation that the *general* theory of Lagrange duality does not apply, at least directly, to linear programming, a fact that is not sufficiently emphasized in many expositions. A separate treatment of duality is required.

Unlike the case of linear programming, which needs a separate treatment, Theorem 50.17 applies to the optimization problem involving a convex quadratic objective function and a set of affine inequality constraints. So in some sense, convex quadratic programming is simpler than linear programming!

Example 50.7. Consider the quadratic objective function

$$J(v) = \frac{1}{2} v^\top A v - v^\top b,$$

where A is an $n \times n$ matrix which is symmetric positive definite, $b \in \mathbb{R}^n$, and the constraints are affine inequality constraints of the form

$$Cv \leq d,$$

where C is an $m \times n$ matrix and $d \in \mathbb{R}^m$. For the time being, we do not assume that C has rank m. Since A is symmetric positive definite, J is strictly convex, as implied by Proposition 40.11 (see Example 40.6). The Lagrangian of this quadratic optimization problem is given by

$$\begin{aligned} L(v, \mu) &= \frac{1}{2} v^\top A v - v^\top b + (Cv - d)^\top \mu \\ &= \frac{1}{2} v^\top A v - v^\top (b - C^\top \mu) - \mu^\top d. \end{aligned}$$

Since A is symmetric positive definite, by Proposition 42.2, the function $v \mapsto L(v, \mu)$ has a unique minimum obtained for the solution u_μ of the linear system

$$Av = b - C^\top \mu;$$

that is,

$$u_\mu = A^{-1}(b - C^\top \mu).$$

This shows that the Problem (P_μ) has a unique solution which depends continuously on μ. Then for any solution λ of the dual problem, $u_\lambda = A^{-1}(b - C^\top \lambda)$ is an optimal solution of the primal problem.

We compute $G(\mu)$ as follows:

$$\begin{aligned}
G(\mu) = L(u_\mu, \mu) &= \frac{1}{2} u_\mu^\top A u_\mu - u_\mu^\top (b - C^\top \mu) - \mu^\top d \\
&= \frac{1}{2} u_\mu^\top (b - C^\top \mu) - u_\mu^\top (b - C^\top \mu) - \mu^\top d \\
&= -\frac{1}{2} u_\mu^\top (b - C^\top \mu) - \mu^\top d \\
&= -\frac{1}{2} (b - C^\top \mu)^\top A^{-1} (b - C^\top \mu) - \mu^\top d \\
&= -\frac{1}{2} \mu^\top C A^{-1} C^\top \mu + \mu^\top (C A^{-1} b - d) - \frac{1}{2} b^\top A^{-1} b.
\end{aligned}$$

Since A is symmetric positive definite, the matrix $CA^{-1}C^\top$ is symmetric positive semidefinite. Since A^{-1} is also symmetric positive definite, $\mu^\top C A^{-1} C^\top \mu = 0$ iff $(C^\top \mu)^\top A^{-1} (C^\top \mu) = 0$ iff $C^\top \mu = 0$ implies $\mu = 0$, that is, $\operatorname{Ker} C^\top = (0)$, which is equivalent to $\operatorname{Im}(C) = \mathbb{R}^m$, namely if C has rank m (in which case, $m \leq n$). Thus $CA^{-1}C^\top$ is symmetric positive definite iff C has rank m.

We showed just after Theorem 49.8 that the functional $v \mapsto (1/2) v^\top A v$ is elliptic iff A is symmetric positive definite, and Theorem 49.8 shows that an elliptic functional is coercive, which is the hypothesis used in Theorem 49.4. Therefore, by Theorem 49.4, if the inequalities $Cx \leq d$ have a solution, the primal problem has a unique solution. In this case, as a consequence, by Theorem 50.17(2), the function $-G(\mu)$ always has a minimum, which is unique if C has rank m. The fact that $-G(\mu)$ has a minimum is not obvious when C has rank $< m$, since in this case $CA^{-1}C^\top$ is not invertible.

We also verify easily that the gradient of G is given by

$$\nabla G_\mu = C u_\mu - d = -C A^{-1} C^\top \mu + C A^{-1} b - d.$$

Observe that since $CA^{-1}C^\top$ is symmetric positive semidefinite, $-G(\mu)$ is convex.

Therefore, if C has rank m, a solution of Problem (P) is obtained by finding the unique solution λ of the equation

$$-C A^{-1} C^\top \mu + C A^{-1} b - d = 0,$$

and then the minimum u_λ of Problem (P) is given by

$$u_\lambda = A^{-1}(b - C^\top \lambda).$$

If C has rank $< m$, then we can find $\lambda \geq 0$ by finding a feasible solution of the linear program whose set of constraints is given by

$$-CA^{-1}C^\top \mu + CA^{-1}b - d = 0,$$

using the standard method of adding nonnegative slack variables $\xi_1, \ldots, \xi_m$ and maximizing $-(\xi_1 + \cdots + \xi_m)$.

50.9 Handling Equality Constraints Explicitly

Sometimes it is desirable to handle equality constraints explicitly (for instance, this is what Boyd and Vandenberghe do, see [29]). The only difference is that the Lagrange multipliers associated with *equality constraints* are *not required* to be nonnegative, as we now show.

Consider the *Optimization Problem* (P')

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \\ & && \psi_j(v) = 0, \quad j = 1, \ldots, p. \end{aligned}$$

We treat each equality constraint $\psi_j(u) = 0$ as the conjunction of the inequalities $\psi_j(u) \leq 0$ and $-\psi_j(u) \leq 0$, and we associate Lagrange multipliers $\lambda \in \mathbb{R}^m_+$, and $\nu^+, \nu^- \in \mathbb{R}^p_+$. Assuming that the constraints are qualified, by Theorem 50.5, the KKT conditions are

$$J'_u + \sum_{i=1}^m \lambda_i (\varphi'_i)_u + \sum_{j=1}^p \nu_j^+ (\psi'_j)_u - \sum_{j=1}^p \nu_j^- (\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) + \sum_{j=1}^p \nu_j^+ \psi_j(u) - \sum_{j=1}^p \nu_j^- \psi_j(u) = 0,$$

with $\lambda \geq 0, \nu^+ \geq 0, \nu^- \geq 0$. Since $\psi_j(u) = 0$ for $j = 1, \ldots, p$, these equations can be rewritten as

$$J'_u + \sum_{i=1}^m \lambda_i (\varphi'_i)_u + \sum_{j=1}^p (\nu_j^+ - \nu_j^-)(\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^m \lambda_i \varphi_i(u) = 0$$

with $\lambda \geq 0, \nu^+ \geq 0, \nu^- \geq 0$, and if we introduce $\nu_j = \nu_j^+ - \nu_j^-$ we obtain the following KKT conditions for programs with explicit equality constraints:

$$J'_u + \sum_{i=1}^{m} \lambda_i (\varphi'_i)_u + \sum_{j=1}^{p} \nu_j (\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^{m} \lambda_i \varphi_i(u) = 0$$

with $\lambda \geq 0$ and $\nu \in \mathbb{R}^p$ arbitrary.

Let us now assume that the functions φ_i and ψ_j are *convex*. As we explained just after Definition 50.6, nonaffine equality constraints are never qualified. Thus, in order to generalize Theorem 50.6 to explicit equality constraints, we assume that the *equality constraints* ψ_j are *affine*.

Theorem 50.18. *Let $\varphi_i\colon \Omega \to \mathbb{R}$ be m convex inequality constraints and $\psi_j\colon \Omega \to \mathbb{R}$ be p affine equality constraints defined on some open convex subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), let $J\colon \Omega \to \mathbb{R}$ be some function, let U be given by*

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ \psi_j(x) = 0,\ \ 1 \leq i \leq m,\ 1 \leq j \leq p\},$$

and let $u \in U$ be any point such that the functions φ_i and J are differentiable at u, and the functions ψ_j are affine.

(1) If J has a local minimum at u with respect to U, and if the constraints are qualified, then there exist some vectors $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$, such that the KKT condition hold:

$$J'_u + \sum_{i=1}^{m} \lambda_i(u) (\varphi'_i)_u + \sum_{j=1}^{p} \nu_j (\psi'_j)_u = 0,$$

and

$$\sum_{i=1}^{m} \lambda_i(u) \varphi_i(u) = 0, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m.$$

Equivalently, in terms of gradients, the above conditions are expressed as

$$\nabla J_u + \sum_{i=1}^{m} \lambda_i \nabla(\varphi_i)_u + \sum_{j=1}^{p} \nu_j \nabla(\psi_j)_u = 0$$

and

$$\sum_{i=1}^{m} \lambda_i(u) \varphi_i(u) = 0, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m.$$

(2) *Conversely, if the restriction of J to U is convex and if there exist vectors $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$ such that the KKT conditions hold, then the function J has a (global) minimum at u with respect to U.*

The Lagrangian $L(v, \lambda, \nu)$ of Problem (P') is defined as

$$L(v, \mu, \nu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v) + \sum_{j=1}^{p} \nu_i \psi_j(v),$$

where $v \in \Omega$, $\mu \in \mathbb{R}^m_+$, and $\nu \in \mathbb{R}^p$.

The function $G\colon \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ given by

$$G(\mu, \nu) = \inf_{v \in \Omega} L(v, \mu, \nu) \quad \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p$$

is called the *Lagrange dual function* (or *dual function*), and the *Dual Problem (D') is*

$$\begin{aligned} &\text{maximize} && G(\mu, \nu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p. \end{aligned}$$

Observe that the Lagrange multipliers ν *are not restricted* to be nonnegative.

Theorem 50.15 and Theorem 50.17 are immediately generalized to Problem (P'). We only state the new version of 50.17, leaving the new version of Theorem 50.15 as an exercise.

Theorem 50.19. *Consider the minimization problem (P'):*

$$\begin{aligned} &\text{minimize} && J(v) \\ &\text{subject to} && \varphi_i(v) \leq 0, \quad i = 1, \ldots, m \\ & && \psi_j(v) = 0, \quad j = 1, \ldots, p. \end{aligned}$$

where the functions J, φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V), and the functions ψ_j are affine.

(1) *Suppose the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous, and that for every $\mu \in \mathbb{R}^m_+$ and every $\nu \in \mathbb{R}^p$, the Problem $(P_{\mu,\nu})$:*

$$\begin{aligned} &\text{minimize} && L(v, \mu, \nu) \\ &\text{subject to} && v \in \Omega, \end{aligned}$$

has a unique solution $u_{\mu,\nu}$, so that

$$L(u_{\mu,\nu}, \mu, \nu) = \inf_{v \in \Omega} L(v, \mu, \nu) = G(\mu, \nu),$$

and the function $(\mu, \nu) \mapsto u_{\mu,\nu}$ *is continuous (on* $\mathbb{R}^m_+ \times \mathbb{R}^p$*). Then the function* G *is differentiable for all* $\mu \in \mathbb{R}^m_+$ *and all* $\nu \in \mathbb{R}^p$*, and*

$$G'_{\mu,\nu}(\xi, \zeta) = \sum_{i=1}^m \xi_i \varphi_i(u_{\mu,\nu}) + \sum_{j=1}^p \zeta_j \psi_j(u_{\mu,\nu}) \quad \text{for all } \xi \in \mathbb{R}^m \text{ and all } \zeta \in \mathbb{R}^p\ .$$

If (λ, η) *is any solution of Problem* (D)*:*

$$\begin{aligned} &\text{maximize} && G(\mu, \nu) \\ &\text{subject to} && \mu \in \mathbb{R}^m_+,\ \nu \in \mathbb{R}^p, \end{aligned}$$

then the solution $u_{\lambda,\eta}$ *of the corresponding Problem* $(P_{\lambda,\eta})$ *is a solution of Problem* (P')*.*

(2) *Assume Problem* (P') *has some solution* $u \in U$*, and that* Ω *is convex (open), the functions* φ_i $(1 \leq i \leq m)$ *and* J *are convex, differentiable at* u*, and that the constraints are qualified. Then Problem* (D') *has a solution* $(\lambda, \eta) \in \mathbb{R}^m_+ \times \mathbb{R}^p$*, and* $J(u) = G(\lambda, \eta)$*; that is, the duality gap is zero.*

In the next section we derive the dual function and the dual program of the optimization problem of Section 50.6 (Hard margin SVM), which involves both inequality and equality constraints. We also derive the KKT conditions associated with the dual program.

50.10 Dual of the Hard Margin Support Vector Machine

Recall the **Hard margin SVM** problem (SVM_{h2}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2, \qquad w \in \mathbb{R}^n \\ &\text{subject to} \\ &\qquad\quad w^\top u_i - b \geq 1 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 \qquad j = 1, \ldots, q. \end{aligned}$$

We proceed in six steps.

Step 1: Write the constraints in matrix form.

The inequality constraints are written as

$$C \begin{pmatrix} w \\ b \end{pmatrix} \leq d,$$

where C is a $(p+q)\times(n+1)$ matrix C and $d \in \mathbb{R}^{p+q}$ is the vector given by

$$C = \begin{pmatrix} -u_1^\top & 1 \\ \vdots & \vdots \\ -u_p^\top & 1 \\ v_1^\top & -1 \\ \vdots & \vdots \\ v_q^\top & -1 \end{pmatrix}, \quad d = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix} = -\mathbf{1}_{p+q}.$$

If we let X be the $n \times (p+q)$ matrix given by

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

then

$$C = \begin{pmatrix} X^\top & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} \end{pmatrix}$$

and so

$$C^\top = \begin{pmatrix} X \\ \mathbf{1}_p^\top \quad -\mathbf{1}_q^\top \end{pmatrix}.$$

Step 2: Write the objective function in matrix form.

The objective function is given by

$$J(w, b) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix}.$$

Note that the corresponding matrix is symmetric positive semidefinite, but it is *not* invertible. Thus the function J is convex but not strictly convex.

Step 3: Write the Lagrangian in matrix form.

As in Example 50.7, we obtain the Lagrangian

$$L(w, b, \lambda, \mu) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \left(0_{n+1} - C^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right) + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q},$$

that is,

$$L(w, b, \lambda, \mu) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} + \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

Step 4: Find the dual function $G(\lambda, \mu)$.

In order to find the dual function $G(\lambda, \mu)$, we need to minimize $L(w, b, \lambda, \mu)$ with respect to w and b and for this, since the objective function J is convex and since $\mathbb{R}^{n+1}$ is convex and open, we can apply Theorem 40.13, which gives a necessary and sufficient condition for a minimum. The gradient of $L(w, b, \lambda, \mu)$ with respect to w and b is

$$\begin{aligned}
\nabla L_{w,b} &= \begin{pmatrix} I_n & 0_n \\ 0_n^\top & 0 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} + \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ \mathbf{1}_p^\top \lambda \quad -\mathbf{1}_q^\top \mu \end{pmatrix} \\
&= \begin{pmatrix} w \\ 0 \end{pmatrix} + \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ \mathbf{1}_p^\top \lambda \quad -\mathbf{1}_q^\top \mu \end{pmatrix}.
\end{aligned}$$

The necessary and sufficient condition for a minimum is

$$\nabla L_{w,b} = 0,$$

which yields

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \tag{$*_1$}$$

and

$$\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu = 0. \tag{$*_2$}$$

The second equation can be written as

$$\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j. \tag{$*_3$}$$

Plugging back w from $(*_1)$ into the Lagrangian and using $(*_2)$ we get

$$G(\lambda, \mu) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q}; \tag{$*_4$}$$

of course, $\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} = \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j$. Actually, to be perfectly rigorous $G(\lambda, \mu)$ is only defined on the intersection of the hyperplane of equation $\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j$ with the convex octant in $\mathbb{R}^{p+q}$ given by $\lambda \geq 0, \mu \geq 0$, so for all $\lambda \in \mathbb{R}_+^p$ and all $\mu \in \mathbb{R}_+^q$, we have

$$G(\lambda, \mu) = \begin{cases} -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} & \text{if } \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ -\infty & \text{otherwise.} \end{cases}$$

Note that the condition

$$\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j$$

is Condition $(*_2)$ of Example 50.6, which is not surprising.

Step 5: Write the dual program in matrix form.

Maximizing the dual function $G(\lambda, \mu)$ over its domain of definition is equivalent to maximizing

$$\widehat{G}(\lambda, \mu) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q}$$

subject to the constraint

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j,$$

so we formulate the dual program as,

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j \\ &\qquad \lambda \geq 0, \ \mu \geq 0, \end{aligned}$$

or equivalently,

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j \\ &\qquad \lambda \geq 0, \ \mu \geq 0. \end{aligned}$$

The constraints of the dual program are a lot simpler than the constraints

$$\begin{pmatrix} X^\top & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \leq -\mathbf{1}_{p+q}$$

of the primal program because these constraints have been "absorbed" by the objective function $\widehat{G}(\lambda, \nu)$ of the dual program which involves the matrix $X^\top X$. The matrix $X^\top X$ is symmetric positive semidefinite, but not invertible in general.

Step 6: Solve the dual program.

This step involves using numerical procedures typically based on gradient descent to find λ and μ, for example, ADMM from Section 52.6. Once λ and μ are determined, w is determined by $(*_1)$ and b is determined as in Section 50.6 using the fact that there is at least some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$.

Remarks:

(1) Since the constraints are affine and the objective function is convex, by Theorem 50.19(2) the duality gap is zero, so for any minimum w of $J(w,b) = (1/2)w^\top w$ and any maximum (λ, μ) of G, we have

$$J(w,b) = \frac{1}{2} w^\top w = G(\lambda, \mu).$$

But by $(*_1)$

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

so

$$\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = w^\top w,$$

and we get

$$\frac{1}{2} w^\top w = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} = -\frac{1}{2} w^\top w + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q},$$

so

$$w^\top w = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j,$$

which yields

$$G(\lambda, \mu) = \frac{1}{2} \left(\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \right).$$

The above formulae are stated in Vapnik [182] (Chapter 10, Section 1).

(2) It is instructive to compute the Lagrangian of the dual program and to derive the KKT conditions for this Lagrangian.

The conditions $\lambda \geq 0$ being equivalent to $-\lambda \leq 0$, and the conditions $\mu \geq 0$ being equivalent to $-\mu \leq 0$, we introduce Lagrange multipliers $\alpha \in \mathbb{R}^p_+$ and $\beta \in \mathbb{R}^q_+$ as well as a multiplier $\rho \in \mathbb{R}$ for the equational constraint, and we form the Lagrangian

$$\begin{aligned} L(\lambda, \mu, \alpha, \beta, \rho) = &\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &- \sum_{i=1}^{p} \alpha_i \lambda_i - \sum_{j=1}^{q} \beta_j \mu_j + \rho \Big(\sum_{j=1}^{q} \mu_j - \sum_{i=1}^{p} \lambda_i \Big). \end{aligned}$$

It follows that the KKT conditions are

$$X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \mathbf{1}_{p+q} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \rho \begin{pmatrix} -\mathbf{1}_p \\ \mathbf{1}_q \end{pmatrix} = 0_{p+q}, \qquad (*_4)$$

and $\alpha_i \lambda_i = 0$ for $i = 1, \ldots, p$ and $\beta_j \mu_j = 0$ for $j = 1, \ldots, q$.

But $(*_4)$ is equivalent to

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \rho \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0_{p+q},$$

which is precisely the result of adding $\alpha \geq 0$ and $\beta \geq 0$ as slack variables to the inequalities $(*_3)$ of Example 50.6, namely

$$-X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + b \begin{pmatrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{pmatrix} + \mathbf{1}_{p+q} \leq 0_{p+q},$$

to make them equalities, where ρ plays the role of b.

When the constraints are *affine*, the dual function $G(\lambda, \nu)$ can be expressed in terms of the conjugate of the objective function J.

50.11 Conjugate Function and Legendre Dual Function

The notion of conjugate function goes back to Legendre and plays an important role in classical mechanics for converting a Lagrangian to a Hamiltonian; see Arnold [5] (Chapter 3, Sections 14 and 15).

Definition 50.11. Let $f\colon A \to \mathbb{R}$ be a function defined on some subset A of $\mathbb{R}^n$. The *conjugate* f^* of the function f is the partial function $f^*\colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f^*(y) = \sup_{x \in A}(\langle y, x\rangle - f(x)) = \sup_{x \in A}(y^\top x - f(x)), \quad y \in \mathbb{R}^n.$$

The conjugate of a function is also called the *Fenchel conjugate*, or *Legendre transform* when f is differentiable.

As the pointwise supremum of a family of affine functions in y, the conjugate function f^* is *convex*, even if the original function f is not convex.

By definition of f^* we have

$$f(x) + f^*(y) \geq \langle x, y\rangle = x^\top y,$$

whenever the left-hand side is defined. The above is known as *Fenchel's inequality* (or *Young's inequality* if f is differentiable).

If $f\colon A \to \mathbb{R}$ is convex (so A is convex) and if $\mathbf{epi}(f)$ is closed, then it can be shown that $f^{**} = f$. In particular, this is true if $A = \mathbb{R}^n$.

The domain of f^* can be very small, even if the domain of f is big. For example, if $f\colon \mathbb{R} \to \mathbb{R}$ is the affine function given by $f(x) = ax + b$ (with $a, b \in \mathbb{R}$), then the function $x \mapsto yx - ax - b$ is unbounded above unless $y = a$, so

$$f^*(y) = \begin{cases} -b & \text{if } y = a \\ +\infty & \text{otherwise.} \end{cases}$$

The domain of f^* can also be bigger than the domain of f; see Example 50.8(3).

The conjugates of many functions that come up in optimization are derived in Boyd and Vandenberghe; see [29], Section 3.3. We mention a few that will be used in this chapter.

Example 50.8.

(1) *Negative logarithm*: $f(x) = -\log x$, with $\mathrm{dom}(f) = \{x \in \mathbb{R} \mid x > 0\}$. The function $x \mapsto yx + \log x$ is unbounded above if $y \geq 0$, and when $y < 0$, its maximum is obtained iff its derivative is zero, namely
$$y + \frac{1}{x} = 0.$$
Substituting for $x = -1/y$ in $yx + \log x$, we obtain $-1 + \log(-1/y) = -1 - \log(-y)$, so we have
$$f^*(y) = -\log(-y) - 1,$$
with $\mathrm{dom}(f^*) = \{y \in \mathbb{R} \mid y < 0\}$.

(2) *Exponential*: $f(x) = e^x$, with $\mathrm{dom}(f) = \mathbb{R}$. The function $x \mapsto yx - e^x$ is unbounded if $y < 0$. When $y > 0$, it reaches a maximum iff its derivative is zero, namely
$$y - e^x = 0.$$
Substituting for $x = \log y$ in $yx - e^x$, we obtain $y \log y - y$, so we have
$$f^*(y) = y \log y - y,$$
with $\mathrm{dom}(f^*) = \{y \in \mathbb{R} \mid y \geq 0\}$, with the convention that $0 \log 0 = 0$.

(3) *Negative Entropy*: $f(x) = x \log x$, with $\mathrm{dom}(f) = \{x \in \mathbb{R} \mid x \geq 0\}$, with the convention that $0 \log 0 = 0$. The function $x \mapsto yx - x \log x$ is bounded above for all $y > 0$, and it attains its maximum when its derivative is zero, namely
$$y - \log x - 1 = 0.$$
Substituting for $x = e^{y-1}$ in $yx - x \log x$, we obtain $ye^{y-1} - e^{y-1}(y - 1) = e^{y-1}$, which yields
$$f^*(y) = e^{y-1},$$
with $\mathrm{dom}(f^*) = \mathbb{R}$.

(4) *Strictly convex quadratic function*: $f(x) = \frac{1}{2}x^\top Ax$, where A is an $n \times n$ symmetric positive definite matrix, with $\mathrm{dom}(f) = \mathbb{R}^n$. The function $x \mapsto y^\top x - \frac{1}{2}x^\top Ax$ has a unique maximum when is gradient is zero, namely

$$y = Ax.$$

Substituting for $x = A^{-1}y$ in $y^\top x - \frac{1}{2}x^\top Ax$, we obtain

$$y^\top A^{-1}y - \frac{1}{2}y^\top A^{-1}y = -\frac{1}{2}y^\top A^{-1}y,$$

so

$$f^*(y) = -\frac{1}{2}y^\top A^{-1}y$$

with $\mathrm{dom}(f^*) = \mathbb{R}^n$.

(5) *Log-determinant*: $f(X) = \log\det(X^{-1})$, where X is an $n \times n$ symmetric positive definite matrix. Then

$$f(Y) = \log\det((-Y)^{-1}) - n,$$

where Y is an $n \times n$ symmetric negative definite matrix; see Boyd and Vandenberghe; see [29], Section 3.3.1, Example 3.23.

(6) *Norm on* $\mathbb{R}^n$: $f(x) = \|x\|$ for any norm $\|\ \|$ on $\mathbb{R}^n$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Recall from Section 14.7 that the dual norm $\|\ \|^D$ of the norm $\|\ \|$ (with respect to the canonical inner product $x \cdot y = y^\top x$ on $\mathbb{R}^n$ is given by

$$\|y\|^D = \sup_{\|x\|=1} |y^\top x|,$$

and that

$$|y^\top x| \leq \|x\| \, \|y\|^D .$$

We have

$$\begin{aligned} f^*(y) &= \sup_{x\in\mathbb{R}^n} (y^\top x - \|x\|) \\ &= \sup_{x\in\mathbb{R}^n, x\neq 0} \left(y^\top \frac{x}{\|x\|} - 1 \right) \|x\| \\ &\leq \sup_{x\in\mathbb{R}^n, x\neq 0} (\|y\|^D - 1) \, \|x\| , \end{aligned}$$

so if $\|y\|^D > 1$ this last term goes to $+\infty$, but if $\|y\|^D \leq 1$, then its maximum is 0. Therefore,

$$f^*(y) = \|y\|^* = \begin{cases} 0 & \text{if } \|y\|^D \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

(7) *Norm squared*: $f(x) = \frac{1}{2}\|x\|^2$ for any norm $\| \ \|$ on $\mathbb{R}^n$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Since $|y^\top x| \leq \|x\| \, \|y\|^D$, we have

$$y^\top x - (1/2)\|x\|^2 \leq \|y\|^D \|x\| - (1/2)\|x\|^2 .$$

The right-hand side is a quadratic function of $\|x\|$ which achieves its maximum at $\|x\| = \|y\|^D$, with maximum value $(1/2)(\|y\|^D)^2$. Therefore

$$y^\top x - (1/2)\|x\|^2 \leq (1/2)\Big(\|y\|^D\Big)^2$$

for all x, which shows that

$$f^*(y) \leq (1/2)\Big(\|y\|^D\Big)^2.$$

By definition of the dual norm and because the unit sphere is compact, for any $y \in \mathbb{R}^n$, there is some $x \in \mathbb{R}^n$ such that $\|x\| = 1$ and $y^\top x = \|y\|^D$, so multiplying both sides by $\|y\|^D$ we obtain

$$y^\top \|y\|^D x = \Big(\|y\|^D\Big)^2$$

and for $z = \|y\|^D x$, since $\|x\| = 1$ we have $\|z\| = \|y\|^D \|x\| = \|y\|^D$, so we get

$$y^\top z - (1/2)(\|z\|)^2 = \Big(\|y\|^D\Big)^2 - (1/2)\Big(\|y\|^D\Big)^2 = (1/2)\Big(\|y\|^D\Big)^2,$$

which shows that the upper bound $(1/2)\Big(\|y\|^D\Big)^2$ is achieved. Therefore,

$$f^*(y) = \frac{1}{2}\Big(\|y\|^D\Big)^2,$$

and $\mathrm{dom}(f^*) = \mathbb{R}^n$.

(8) *Log-sum-exp function*: $f(x) = \log\Big(\sum_{i=1}^n e^{x_i}\Big)$, where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. To determine the values of $y \in \mathbb{R}^n$ for which the maximum of $g(x) = y^\top x - f(x)$ over $x \in \mathbb{R}^n$ is attained, we compute its gradient and we find

$$\nabla f_x = \begin{pmatrix} y_1 - \dfrac{e^{x_1}}{\sum_{i=1}^n e^{x_i}} \\ \vdots \\ y_n - \dfrac{e^{x_n}}{\sum_{i=1}^n e^{x_i}} \end{pmatrix}.$$

Therefore, $(y_1, \ldots, y_n)$ must satisfy the system of equations

$$y_j = \frac{e^{x_j}}{\sum_{i=1}^n e^{x_i}}, \quad j = 1, \ldots, n. \qquad (*)$$

The condition $\sum_{i=1}^n y_i = 1$ is obviously necessary, as well as the conditions $y_i > 0$, for $i = 1, \ldots, n$. Conversely, if $\mathbf{1}^\top y = 1$ and $y > 0$, then $x_j = \log y_i$ for $i = 1, \ldots, n$ is a solution. Since $(*)$ implies that

$$x_i = \log y_i + \log\left(\sum_{i=1}^n e^{x_i}\right), \tag{$**$}$$

we get

$$\begin{aligned} y^\top x - f(x) &= \sum_{i=1}^n y_i x_i - \log\left(\sum_{i=1}^n e^{x_i}\right) \\ &= \sum_{i=1}^n y_i \log y_i + \sum_{i=1}^n y_i \log\left(\sum_{i=1}^n e^{x_i}\right) - \log\left(\sum_{i=1}^n e^{x_i}\right) \quad \text{by } (**) \\ &= \sum_{i=1}^n y_i \log y_i + \left(\sum_{i=1}^n y_i - 1\right) \log\left(\sum_{i=1}^n e^{x_i}\right) \\ &= \sum_{i=1}^n y_i \log y_i \qquad \text{since } \textstyle\sum_{i=1}^n y_i = 1. \end{aligned}$$

Consequently, if $f^*(y)$ is defined, then $f^*(y) = \sum_{i=1}^n y_i \log y_i$. If we agree that $0 \log 0 = 0$, then it is an easy exercise (or, see Boyd and Vandenberghe [29], Section 3.3, Example 3.25) to show that

$$f^*(y) = \begin{cases} \sum_{i=1}^n y_i \log y_i & \text{if } \mathbf{1}^\top y = 1 \text{ and } y \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

Thus we obtain the negative entropy restricted to the domain $\mathbf{1}^\top y = 1$ and $y \geq 0$.

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, then x^* maximizes $x^\top y - f(x)$ iff x^* minimizes $-x^\top y + f(x)$ iff

$$\nabla f_{x^*} = y,$$

and so

$$f^*(y) = (x^*)^\top \nabla f_{x^*} - f(x^*).$$

Consequently, if we can solve the equation

$$\nabla f_z = y$$

for z given y, then we obtain $f^*(y)$.

It can be shown that if f is twice differentiable, strictly convex, and surlinear, which means that

$$\lim_{\|y\| \mapsto +\infty} \frac{f(y)}{\|y\|} = +\infty,$$

then there is a unique x_y such that $\nabla f_{x_y} = y$, so that

$$f^*(y) = x_y^\top \nabla f_{x_y} - f(x_y),$$

and f^* is differentiable with

$$\nabla f^*_y = x_y.$$

We now return to our optimization problem.

Proposition 50.20. *Consider Problem* (P),

$$\begin{aligned} \text{minimize} \quad & J(v) \\ \text{subject to} \quad & Av \leq b \\ & Cv = d, \end{aligned}$$

with affine inequality and equality constraints (with A *an* $m \times n$ *matrix,* C *an* $p \times n$ *matrix,* $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$*). The dual function* $G(\lambda, \nu)$ *is given by*

$$G(\lambda, \nu) = \begin{cases} -b^\top \lambda - d^\top \nu - J^*(-A^\top \lambda - C^\top \nu) & \textit{if } -A^\top \lambda - C^\top \nu \in \mathrm{dom}(J^*), \\ -\infty & \textit{otherwise}, \end{cases}$$

for all $\lambda \in \mathbb{R}^m_+$ *and all* $\nu \in \mathbb{R}^p$*, where* J^* *is the conjugate of* J*.*

Proof. The Lagrangian associated with the above program is

$$\begin{aligned} L(v, \lambda, \nu) &= J(v) + (Av - b)^\top \lambda + (Cv - d)^\top \nu \\ &= -b^\top \lambda - d^\top \nu + J(v) + (A^\top \lambda + C^\top \nu)^\top v, \end{aligned}$$

with $\lambda \in \mathbb{R}^m_+$ and $\nu \in \mathbb{R}^p$. By definition

$$\begin{aligned} G(\lambda, \nu) &= -b^\top \lambda - d^\top \nu + \inf_{v \in \mathbb{R}^n} (J(v) + (A^\top \lambda + C^\top \nu)^\top v) \\ &= -b^\top \lambda - d^\top \nu - \sup_{v \in \mathbb{R}^n} (-(A^\top \lambda + C^\top \nu)^\top v - J(v)) \\ &= -b^\top \lambda - d^\top \nu - J^*(-A^\top \lambda - C^\top \nu). \end{aligned}$$

Therefore, for all $\lambda \in \mathbb{R}^m_+$ and all $\nu \in \mathbb{R}^p$, we have

$$G(\lambda, \nu) = \begin{cases} -b^\top \lambda - d^\top \nu - J^*(-A^\top \lambda - C^\top \nu) & \text{if } -A^\top \lambda - C^\top \nu \in \mathrm{dom}(J^*), \\ -\infty & \text{otherwise}, \end{cases}$$

as claimed. ∎

As application of Proposition 50.20, consider the following example.

Example 50.9. Consider the following problem:

$$\begin{aligned} &\text{minimize} && \|v\| \\ &\text{subject to} && Av = b, \end{aligned}$$

where $\| \ \|$ is any norm on $\mathbb{R}^n$. Using the result of Example 50.8(6), we obtain

$$G(\nu) = -b^\top \nu - \left\| -A^\top \nu \right\|^*,$$

that is,

$$G(\nu) = \begin{cases} -b^\top \nu & \text{if } \left\| A^\top \nu \right\|^D \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

In the special case where $\| \ \| = \| \ \|_2$, we also have $\| \ \|^D = \| \ \|_2$.

Another interesting application is to the entropy minimization problem.

Example 50.10. Consider the following problem known as *entropy minimization*:

$$\begin{aligned} &\text{minimize} && f(x) = \sum_{i=1}^n x_i \log x_i \\ &\text{subject to} && Ax \leq b \\ & && \mathbf{1}^\top x = 1, \end{aligned}$$

where $\mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid x \geq 0\}$. By Example 50.8(3), the conjugate of the negative entropy function $u \log u$ is e^{v-1}, so we easily see that

$$f^*(y) = \sum_{i=1}^n e^{y_i - 1},$$

which is defined on $\mathbb{R}^n$. Proposition 50.20 implies that the dual function $G(\lambda, \mu)$ of the entropy minimization problem is given by

$$G(\lambda, \mu) = -b^\top \lambda - \mu - e^{-\mu-1} \sum_{i=1}^n e^{-(A^i)^\top \lambda},$$

for all $\lambda \in \mathbb{R}^n_+$ and all $\mu \in \mathbb{R}$, where A^i is the ith column of A. It follows that the dual program is:

$$\begin{aligned} &\text{maximize} && -b^\top \lambda - \mu - e^{-\mu-1} \sum_{i=1}^n e^{-(A^i)^\top \lambda} \\ &\text{subject to} && \lambda \geq 0. \end{aligned}$$

We can simplify this problem by maximizing over the variable $\mu \in \mathbb{R}$. For fixed λ, the objective function is maximized when the derivative is zero, that is,

$$-1 + e^{-\mu-1} \sum_{i=1}^{n} e^{-(A^i)^\top \lambda} = 0,$$

which yields

$$\mu = \log\left(\sum_{i=1}^{n} e^{-(A^i)^\top \lambda}\right) - 1.$$

By plugging the above value back into the objective function of the dual, we obtain the following program:

$$\begin{aligned} &\text{maximize} \quad -b^\top \lambda - \log\left(\sum_{i=1}^{n} e^{-(A^i)^\top \lambda}\right) \\ &\text{subject to} \quad \lambda \geq 0. \end{aligned}$$

The entropy minimization problem is another problem for which Theorem 50.18 applies, and thus can be solved using the dual program. Indeed, the Lagrangian of the primal program is given by

$$L(x, \lambda, \mu) = \sum_{i-1}^{n} x_i \log x_i + \lambda^\top (Ax - b) + \mu(\mathbf{1}^\top x - 1).$$

Using the second derivative criterion for convexity, we see that $L(x, \lambda, \mu)$ is strictly convex for $x \in \mathbb{R}^n_+$ and is bounded below, so it has a unique minimum which is obtain by setting the gradient ∇L_x to zero. We have

$$\nabla L_x = \begin{pmatrix} \log x_1 + 1 + (A^1)^\top \lambda + \mu \\ \vdots \\ \log x_n + 1 + (A^n)^\top \lambda + \mu. \end{pmatrix}$$

so by setting ∇L_x to 0 we obtain

$$x_i = e^{-((A^n)^\top \lambda + \mu + 1)}, \quad i = 1, \ldots, n. \tag{$*$}$$

By Theorem 50.18, since the objective function is convex and the constraints are affine, if the primal has a solution then so does the dual, and λ and μ constitute an optimal solution of the dual, then $x = (x_1, \ldots, x_n)$ given by the equations in $(*)$ is an optimal solution of the primal.

Other examples are given in Boyd and Vandenberghe; see [29], Section 5.1.6.

The derivation of the dual function of Problem (SVM_{h1}) from Section 50.5 involves a similar type of reasoning.

Example 50.11. Consider the Hard Margin Problem (SVM_{h1}):

$$\begin{aligned}
&\text{maximize} \quad \delta\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q\\
&\qquad \|w\|_2 \leq 1,
\end{aligned}$$

which is converted to the following minimization problem:

$$\begin{aligned}
&\text{minimize} \quad -2\delta\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \delta \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \delta \qquad j = 1, \ldots, q\\
&\qquad \|w\|_2 \leq 1.
\end{aligned}$$

We replaced δ by 2δ because this will make it easier to find a nice geometric interpretation. Recall from Section 50.5 that Problem (SVM_{h1}) has a an optimal solution iff $\delta > 0$, in which case $\|w\| = 1$.

The corresponding Lagrangian with $\lambda \in \mathbb{R}^p_+, \mu \in \mathbb{R}^q_+, \gamma \in \mathbb{R}^+$, is

$$\begin{aligned}
L(w, b, \delta, \lambda, \mu, \gamma) &= -2\delta + \sum_{i=1}^p \lambda_i(\delta + b - w^\top u_i) + \sum_{j=1}^q \mu_j(\delta - b + w^\top v_j) + \gamma(\|w\|_2 - 1)\\
&= w^\top\left(-\sum_{i=1}^p \lambda_i u_i + \sum_{j=1}^q \mu_j v_j\right) + \gamma \|w\|_2 + \left(\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j\right) b\\
&\quad + \left(-2 + \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j\right)\delta - \gamma.
\end{aligned}$$

Next to find the dual function $G(\lambda, \mu, \gamma)$ we need to minimize $L(w, b, \delta, \lambda, \mu, \gamma)$ with respect to w, b and δ, so its gradient with respect to w, b and δ must be zero. This implies that

$$\begin{aligned}
\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0\\
-2 + \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j &= 0,
\end{aligned}$$

which yields

$$\sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = 1.$$

Observe that we did not compute the partial derivative with respect to w because it does not yield any useful information due to the presence of the term $\|w\|_2$ (as opposed to $\|w\|_2^2$). Our minimization problem is reduced to: find

$$\begin{aligned}
&\inf_{w, \|w\| \leq 1} \left(w^\top \left(\sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right) + \gamma \|w\|_2 - \gamma \right) \\
&= -\gamma - \gamma \inf_{w, \|w\| \leq 1} \left(-w^\top \frac{1}{\gamma} \left(\sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right) + \|-w\|_2 \right) \\
&= \begin{cases} -\gamma & \text{if } \left\| \frac{1}{\gamma} \left(\sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right) \right\|_2^D \leq 1 \\ -\infty & \text{otherwise} \end{cases} && \text{by Example 50.8(6)} \\
&= \begin{cases} -\gamma & \text{if } \left\| \sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right\|_2 \leq \gamma \\ -\infty & \text{otherwise.} \end{cases} && \text{since } \| \, \|_2^D = \| \, \|_2 \text{ and } \gamma > 0
\end{aligned}$$

It is immediately verified that the above formula is still correct if $\gamma = 0$. Therefore

$$G(\lambda, \mu, \gamma) = \begin{cases} -\gamma & \text{if } \left\| \sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right\|_2 \leq \gamma \\ -\infty & \text{otherwise.} \end{cases}$$

Since $\left\| \sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right\|_2 \leq \gamma$ iff $-\gamma \leq - \left\| \sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right\|_2$, the dual program, maximizing $G(\lambda, \mu, \gamma)$, is equivalent to

$$\begin{aligned}
&\text{maximize} \quad - \left\| \sum_{j=1}^q \mu_j v_j - \sum_{i=1}^p \lambda_i u_i \right\|_2 \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i = 1, \ \ \lambda \geq 0 \\
&\qquad \sum_{j=1}^q \mu_j = 1, \ \ \mu \geq 0,
\end{aligned}$$

equivalently

$$\begin{aligned}
&\text{minimize} \quad \left\| \sum_{j=1}^{q} \mu_j v_j - \sum_{i=1}^{p} \lambda_i u_i \right\|_2 \\
&\text{subject to} \\
&\qquad \sum_{i=1}^{p} \lambda_i = 1, \;\; \lambda \geq 0 \\
&\qquad \sum_{j=1}^{q} \mu_j = 1, \;\; \mu \geq 0.
\end{aligned}$$

Geometrically, $\sum_{i=1}^{p} \lambda_i u_i$ with $\sum_{i=1}^{p} \lambda_i = 1$ and $\lambda \geq 0$ is a convex combination of the u_is, and $\sum_{j=1}^{q} \mu_j v_j$ with $\sum_{j=1}^{q} \mu_j = 1$ and $\mu \geq 0$ is a convex combination of the v_js, so the dual program is to minimize the distance between the polyhedron $\mathrm{conv}(u_1, \ldots, u_p)$ (the convex hull of the u_is) and the polyhedron $\mathrm{conv}(v_1, \ldots, v_q)$ (the convex hull of the v_js). Since both polyhedra are compact, the shortest distance between then is achieved. In fact, there is some vertex u_i such that if $P(u_i)$ is its projection onto $\mathrm{conv}(v_1, \ldots, v_q)$ (which exists by Hilbert space theory), then the length of the line segment $(u_i, P(u_i))$ is the shortest distance between the two polyhedra (and similarly there is some vertex v_j such that if $P(v_j)$ is its projection onto $\mathrm{conv}(u_1, \ldots, u_p)$ then the length of the line segment $(v_j, P(v_j))$ is the shortest distance between the two polyhedra).

If the two subsets are separable, in which case Problem (SVM$_{h1}$) has an optimal solution $\delta > 0$, because the objective function is convex and the convex constraint $\|w\|_2 \leq 1$ is qualified since δ may be negative, by Theorem 50.17(2) the duality gap is zero, so δ is half of the minimum distance between the two convex polyhedra $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$; see Figure 50.19.

It should be noted that the constraint $\|w\| \leq 1$ yields a formulation of the dual problem which has the advantage of having a nice geometric interpretation: finding the minimal distance between the convex polyhedra $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$. Unfortunately this formulation is not useful for actually solving the problem. However, if the equivalent constraint $\|w\|^2$ $(= w^\top w) \leq 1$ is used, then the dual problem is much more useful as a solving tool.

In Chapter 54 we consider the case where the sets of points $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_q\}$ are not linearly separable.

50.12 Some Techniques to Obtain a More Useful Dual Program

In some cases, it is advantageous to reformulate a primal optimization problem to obtain a more useful dual problem. Three different reformulations are proposed in Boyd and Van-

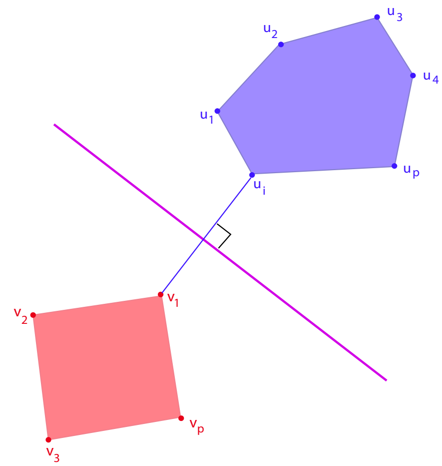


Figure 50.19: In $\mathbb{R}^2$ the convex hull of the u_is, namely the blue hexagon, is separated from the convex hull of the v_js, i.e. the red square, by the purple hyperplane (line) which is the perpendicular bisector to the blue line segment between u_i and v_1, where this blue line segment is the shortest distance between the two convex polygons.

denberghe; see [29], Section 5.7:

(1) Introducing new variables and associated equality constraints.

(2) Replacing the objective function with an increasing function of the the original function.

(3) Making explicit constraints implicit, that is, incorporating them into the domain of the objective function.

We only give illustrations of (1) and (2) and refer the reader to Boyd and Vandenberghe [29] (Section 5.7) for more examples of these techniques.

Consider the unconstrained program:

$$\text{minimize} \quad f(Ax + b),$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. While the conditions for a zero duality gap are satisfied, the Lagrangian is

$$L(x) = f(Ax + b),$$

so the dual function G is the constant function whose value is

$$G = \inf_{x \in \mathbb{R}^n} f(Ax + b),$$

which is not useful at all.

Let us reformulate the problem as

$$\begin{aligned}
&\text{minimize} \quad f(y)\\
&\text{subject to}\\
&\qquad\qquad Ax + b = y,
\end{aligned}$$

where we introduced the new variable $y \in \mathbb{R}^m$ and the equality constraint $Ax + b = y$. The two problems are obviously equivalent. The Lagrangian of the reformulated problem is

$$L(x, y, \mu) = f(y) + \mu^\top (Ax + b - y)$$

where $\mu \in \mathbb{R}^m$. To find the dual function $G(\mu)$ we minimize $L(x, y, \mu)$ over x and y. Minimizing over x we see that $G(\mu) = -\infty$ unless $A^\top \mu = 0$, in which case we are left with

$$G(\mu) = b^\top \mu + \inf_y (f(y) - \mu^\top y) = b^\top \mu - \inf_y (\mu^\top y - f(y)) = b^\top \mu - f^*(\mu),$$

where f^* is the conjugate of f. It follows that the dual program can be expressed as

$$\begin{aligned}
&\text{maximize} \quad b^\top \mu - f^*(\mu)\\
&\text{subject to}\\
&\qquad\qquad A^\top \mu = 0.
\end{aligned}$$

This formulation of the dual is much more useful than the dual of the original program.

Example 50.12. As a concrete example, consider the following unconstrained program:

$$\text{minimize} \quad f(x) = \log\left(\sum_{i=1}^n e^{(A^i)^\top x + b_i}\right)$$

where A^i is a column vector in $\mathbb{R}^n$. We reformulate the problem by introducing new variables and equality constraints as follows:

$$\begin{aligned}
&\text{minimize} \quad f(y) = \log\left(\sum_{i=1}^n e^{y_i}\right)\\
&\text{subject to}\\
&\qquad\qquad Ax + b = y,
\end{aligned}$$

where A is the $n \times n$ matrix whose columns are the vectors A^i and $b = (b_1, \ldots, b_n)$. Since by Example 50.8(8), the conjugate of the log-sum-exp function $f(y) = \log\left(\sum_{i=1}^n e^{y_i}\right)$ is

$$f^*(\mu) = \begin{cases} \sum_{i=1}^n \mu_i \log \mu_i & \text{if } \mathbf{1}^\top \mu = 1 \text{ and } \mu \geq 0\\ \infty & \text{otherwise,} \end{cases}$$

the dual of the reformulated problem can be expressed as

$$\begin{aligned}
&\text{maximize} \quad b^\top \mu - \log\left(\sum_{i=1}^n \mu_i \log \mu_i\right) \\
&\text{subject to} \\
&\qquad\qquad \mathbf{1}^\top \mu = 1 \\
&\qquad\qquad A^\top \mu = 0 \\
&\qquad\qquad \mu \geq 0,
\end{aligned}$$

an entropy maximization problem.

Example 50.13. Similarly the unconstrained norm minimization problem

$$\text{minimize} \quad \|Ax - b\|,$$

where $\| \|$ is any norm on $\mathbb{R}^m$, has a dual function which is a constant, and is not useful. This problem can be reformulated as

$$\begin{aligned}
&\text{minimize} \quad \|y\| \\
&\text{subject to} \\
&\qquad\qquad Ax - b = y.
\end{aligned}$$

By Example 50.8(6), the conjugate of the norm is given by

$$\|y\|^* = \begin{cases} 0 & \text{if } \|y\|^D \leq 1 \\ +\infty & \text{otherwise,} \end{cases}$$

so the dual of the reformulated program is:

$$\begin{aligned}
&\text{maximize} \quad b^\top \mu \\
&\text{subject to} \\
&\qquad\qquad \|\mu\|^D \leq 1 \\
&\qquad\qquad A^\top \mu = 0.
\end{aligned}$$

Here is now an example of (2), replacing the objective function with an increasing function of the the original function.

Example 50.14. The norm minimization of Example 50.13 can be reformulated as

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\|y\|^2 \\
&\text{subject to} \\
&\qquad\qquad Ax - b = y.
\end{aligned}$$

This program is obviously equivalent to the original one. By Example 50.8(7), the conjugate of the square norm is given by

$$\frac{1}{2}\Big(\|y\|^D\Big)^2,$$

so the dual of the reformulated program is

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{2}\Big(\|\mu\|^D\Big)^2 + b^\top \mu \\ &\text{subject to} \\ &\qquad\qquad A^\top \mu = 0. \end{aligned}$$

Note that this dual is different from the dual obtained in Example 50.13.

The objective function of the dual program in Example 50.13 is linear, but we have the nonlinear constraint $\|\mu\|^D \leq 1$. On the other hand, the objective function of the dual program of Example 50.14 is quadratic, whereas its constraints are affine. We have other examples of this trade-off with the Programs (SVM$_{h2}$) (quadratic objective function, affine constraints), and (SVM$_{h1}$) (linear objective function, one nonlinear constraint).

Sometimes, it is also helpful to replace a constraint by an increasing function of this constraint; for example, to use the constraint $\|w\|_2^2\,(= w^\top w) \leq 1$ instead of $\|w\|_2 \leq 1$.

In Chapter 55 we revisit the problem of solving an overdetermined or underdetermined linear system $Ax = b$ considered in Section 23.1, from a different point of view.

50.13 Uzawa's Method

Let us go back to our Minimization Problem

$$\begin{aligned} &\text{minimize} \quad J(v) \\ &\text{subject to} \quad \varphi_i(v) \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where the functions J and φ_i are defined on some open subset Ω of a finite-dimensional Euclidean vector space V (more generally, a real Hilbert space V). As usual, let

$$U = \{v \in V \mid \varphi_i(v) \leq 0,\ 1 \leq i \leq m\}.$$

If the functional J satisfies the inequalities of Proposition 49.18 and if the functions φ_i are convex, in theory, the projected-gradient method converges to the unique minimizer of J over U. Unfortunately, it is usually impossible to compute the projection map $p_U\colon V \to U$.

On the other hand, the domain of the Lagrange dual function $G\colon \mathbb{R}_+^m \to \mathbb{R}$ given by

$$G(\mu) = \inf_{v \in \Omega} L(v, \mu) \quad \mu \in \mathbb{R}_+^m,$$

is $\mathbb{R}^m_+$, where

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v)$$

is the Lagrangian of our problem. Now the projection p_+ from $\mathbb{R}^m$ to $\mathbb{R}^m_+$ is very simple, namely

$$(p_+(\lambda))_i = \max\{\lambda_i, 0\}, \quad 1 \leq i \leq m.$$

It follows that the projection-gradient method should be applicable to the *Dual Problem (D):*

$$\begin{aligned} &\text{maximize} \quad G(\mu) \\ &\text{subject to} \quad \mu \in \mathbb{R}^m_+. \end{aligned}$$

If the hypotheses of Theorem 50.17 hold, then a solution λ of the Dual Program (D) yields a solution u_λ of the primal problem.

Uzawa's method is essentially the gradient method with fixed stepsize applied to the Dual Problem (D). However, it is designed to yield a solution of the primal problem.

Uzawa's method:

Given an arbitrary initial vector $\lambda^0 \in \mathbb{R}^m_+$, two sequences $(\lambda^k)_{k\geq 0}$ and $(u^k)_{k\geq 0}$ are constructed, with $\lambda^k \in \mathbb{R}^m_+$ and $u^k \in V$.

Assuming that $\lambda^0, \lambda^1, \ldots, \lambda^k$ are known, u^k and λ^{k+1} are determined as follows:

u^k is the unique solution of the minimization problem, find $u^k \in V$ such that

$$\text{(UZ)} \quad \begin{cases} J(u^k) + \displaystyle\sum_{i=1}^m \lambda_i^k \varphi_i(u^k) = \inf_{v \in V}\left(J(v) + \sum_{i=1}^m \lambda_i^k \varphi_i(v)\right); \text{ and} \\ \qquad\qquad\qquad\qquad \lambda_i^{k+1} = \max\{\lambda_i^k + \rho \varphi_i(u^k), 0\}, \quad 1 \leq i \leq m, \end{cases}$$

where $\rho > 0$ is a suitably chosen parameter.

Recall that in the proof of Theorem 50.17 we showed $(*_{\text{deriv}})$, namely

$$G'_{\lambda^k}(\xi) = \langle \nabla G_{\lambda^k}, \xi \rangle = \sum_{i=1}^m \xi_i \varphi_i(u^k),$$

which means that $(\nabla G_{\lambda^k})_i = \varphi_i(u^k)$. Then the second equation in (UZ) corresponds to the gradient-projection step

$$\lambda^{k+1} = p_+(\lambda^k + \rho \nabla G_{\lambda^k}).$$

Note that because the problem is a maximization problem we use a positive sign instead of a negative sign. Uzawa's method is indeed a gradient method.

Basically, Uzawa's method replaces a constrained optimization problem by a sequence of unconstrained optimization problems involving the Lagrangian of the (primal) problem.

Interestingly, under certain hypotheses, it is possible to prove that the sequence of approximate solutions $(u^k)_{k\geq 0}$ converges to the minimizer u of J over U, even if the sequence $(\lambda^k)_{k\geq 0}$ does not converge. We prove such a result when the constraints φ_i are *affine*.

Theorem 50.21. *Suppose $J\colon \mathbb{R}^n \to \mathbb{R}$ is an elliptic functional, which means that J is continuously differentiable on $\mathbb{R}^n$, and there is some constant $\alpha > 0$ such that*

$$\langle \nabla J_v - \nabla J_u, v - u\rangle \geq \alpha \left\|v - u\right\|^2 \quad \textit{for all } u, v \in \mathbb{R}^n,$$

and that U is a nonempty closed convex subset given by

$$U = \{v \in \mathbb{R}^n \mid Cv \leq d\},$$

where C is a real $m \times n$ matrix and $d \in \mathbb{R}^m$. If the scalar ρ satisfies the condition

$$0 < \rho < \frac{2\alpha}{\left\|C\right\|_2^2},$$

where $\left\|C\right\|_2$ is the spectral norm of C, then the sequence $(u^k)_{k\geq 0}$ computed by Uzawa's method converges to the unique minimizer $u \in U$ of J.

Furthermore, if C has rank m, then the sequence $(\lambda^k)_{k\geq 0}$ converges to the unique maximizer of the Dual Problem (D).

Proof.

Step 1. We establish algebraic conditions relating the unique minimizer $u \in U$ of J over U and some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point.

Since J is elliptic and U is nonempty closed and convex, by Theorem 49.8, the functional J is strictly convex, so it has a unique minimizer $u \in U$. Since J is convex and the constraints are affine, by Theorem 50.17(2) the Dual Problem (D) has at least one solution. By Theorem 50.15(2), there is some $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point of the Lagrangian L.

If we define the affine function φ by

$$\varphi(v) = (\varphi_1(v), \ldots, \varphi_m(v)) = Cv - d,$$

then the Lagrangian $L(v, \mu)$ can be written as

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v) = J(v) + \langle C^\top \mu, v\rangle - \langle \mu, d\rangle.$$

Since

$$L(u, \lambda) = \inf_{v\in\mathbb{R}^n} L(v, \lambda),$$

by Theorem 40.13(4) we must have

$$\nabla J_u + C^\top \lambda = 0, \tag{$*_1$}$$

and since

$$G(\lambda) = L(u, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} L(u, \mu),$$

by Theorem 40.13(3) (and since maximing a function g is equivalent to minimizing $-g$), we must have

$$G'_\lambda(\mu - \lambda) \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+,$$

and since as noted earlier $\nabla G_\lambda = \varphi(u)$, we get

$$\langle \varphi(u), \mu - \lambda \rangle \leq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+. \tag{$*_2$}$$

As in the proof of Proposition 49.18, $(*_2)$ can be expressed as follows for every $\rho > 0$:

$$\langle \lambda - (\lambda + \rho\varphi(u)), \mu - \lambda \rangle \geq 0 \quad \text{for all } \mu \in \mathbb{R}^m_+, \tag{$**_2$}$$

which shows that λ can be viewed as the projection onto $\mathbb{R}^m_+$ of the vector $\lambda + \rho\varphi(u)$. In summary we obtain the equations

$$(\dagger_1) \quad \begin{cases} \nabla J_u + C^\top \lambda = 0 \\ \lambda = p_+(\lambda + \rho\varphi(u)). \end{cases}$$

Step 2. We establish algebraic conditions relating the unique solution u_k of the minimization problem arising during an iteration of Uzawa's method in (UZ) and λ^k.

Observe that the Lagrangian $L(v, \mu)$ is strictly convex as a function of v (as the sum of a strictly convex function and an affine function). As in the proof of Theorem 49.8(1) and using Cauchy–Schwarz, we have

$$\begin{aligned} J(v) + \langle C^\top \mu, v \rangle &\geq J(0) + \langle \nabla J_0, v \rangle + \frac{\alpha}{2} \|v\|^2 + \langle C^\top \mu, v \rangle \\ &\geq J(0) - \|\nabla J_0\| \|v\| - \left\|C^\top \mu\right\| \|v\| + \frac{\alpha}{2} \|v\|^2, \end{aligned}$$

and the term $(- \|\nabla J_0\| - \left\|C^\top \mu\right\| \|v\| + \frac{\alpha}{2} \|v\|) \|v\|$ goes to $+\infty$ when $\|v\|$ tends to $+\infty$, so $L(v, \mu)$ is coercive as a function of v. Therefore, the minimization problem find u^k such that

$$J(u^k) + \sum_{i=1}^m \lambda_i^k \varphi_i(u^k) = \inf_{v \in \mathbb{R}^n} \left(J(v) + \sum_{i=1}^m \lambda_i^k \varphi_i(v) \right)$$

has a unique solution $u^k \in \mathbb{R}^n$. It follows from Theorem 40.13(4) that the vector u^k must satisfy the equation

$$\nabla J_{u^k} + C^\top \lambda^k = 0, \tag{$*_3$}$$

and since by definition of Uzawa's method

$$\lambda^{k+1} = p_+(\lambda^k + \rho\varphi(u^k)), \tag{$*_4$}$$

we obtain the equations

$$(\dagger_2) \quad \begin{cases} \nabla J_{u^k} + C^\top \lambda^k = 0 \\ \lambda^{k+1} = p_+(\lambda^k + \rho\varphi(u^k)). \end{cases}$$

Step 3. By subtracting the first of the two equations of $(\dagger_1)$ and $(\dagger_2)$ we obtain

$$\nabla J_{u^k} - \nabla J_u + C^\top(\lambda^k - \lambda) = 0,$$

and by subtracting the second of the two equations of $(\dagger_1)$ and $(\dagger_2)$ and using Proposition 48.6, we obtain

$$\left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda + \rho C(u^k - u)\right\|.$$

In summary, we proved

$$(\dagger) \quad \begin{cases} \nabla J_{u^k} - \nabla J_u + C^\top(\lambda^k - \lambda) = 0 \\ \left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda + \rho C(u^k - u)\right\|. \end{cases}$$

Step 4. Convergence of the sequence $(u^k)_{k\geq 0}$ to u.

Squaring both sides of the inequality in $(\dagger)$ we obtain

$$\left\|\lambda^{k+1} - \lambda\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 + 2\rho\langle C^\top(\lambda^k - \lambda), u_k - u\rangle + \rho^2 \left\|C(u^k - u)\right\|^2.$$

Using the equation in $(\dagger)$ and the inequality

$$\langle \nabla J_{u^k} - \nabla J_u, u^k - u\rangle \geq \alpha \left\|u^k - u\right\|^2,$$

we get

$$\begin{aligned} \left\|\lambda^{k+1} - \lambda\right\|^2 &\leq \left\|\lambda^k - \lambda\right\|^2 - 2\rho\langle \nabla J_{u^k} - \nabla J_u, u^k - u\rangle + \rho^2 \left\|C(u^k - u)\right\|^2 \\ &\leq \left\|\lambda^k - \lambda\right\|^2 - \rho(2\alpha - \rho \left\|C\right\|_2^2) \left\|u^k - u\right\|^2. \end{aligned}$$

Consequently, if

$$0 \leq \rho \leq \frac{2\alpha}{\left\|C\right\|_2^2},$$

we have

$$\left\|\lambda^{k+1} - \lambda\right\| \leq \left\|\lambda^k - \lambda\right\|, \quad \text{for all } k \geq 0. \tag{$*_5$}$$

By $(*_5)$, the sequence $(\left\|\lambda^k - \lambda\right\|)_{k\geq 0}$ is nonincreasing and bounded below by 0, so it converges, which implies that

$$\lim_{k\mapsto\infty} \left(\left\|\lambda^{k+1} - \lambda\right\| - \left\|\lambda^k - \lambda\right\|\right) = 0,$$

and since

$$\left\|\lambda^{k+1} - \lambda\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 - \rho(2\alpha - \rho \left\|C\right\|_2^2) \left\|u^k - u\right\|^2,$$

we also have

$$\rho(2\alpha - \rho \|C\|_2^2) \left\|u^k - u\right\|^2 \leq \left\|\lambda^k - \lambda\right\|^2 - \left\|\lambda^{k+1} - \lambda\right\|^2.$$

So if

$$0 < \rho < \frac{2\alpha}{\|C\|_2^2},$$

then $\rho(2\alpha - \rho \|C\|_2^2) > 0$, and we conclude that

$$\lim_{k \mapsto \infty} \left\|u^k - u\right\| = 0,$$

that is, the sequence $(u^k)_{k\geq 0}$ converges to u.

Step 5. Convergence of the sequence $(\lambda^k)_{k\geq 0}$ to λ if C has rank m.

Since the sequence $(\left\|\lambda^k - \lambda\right\|)_{k\geq 0}$ is nonincreasing, the sequence $(\lambda^k)_{k\geq 0}$ is bounded, and thus it has a convergent subsequence $(\lambda^{i(k)})_{i\geq 0}$ whose limit is some $\lambda' \in \mathbb{R}^m_+$. Since J' is continuous, by $(\dagger_2)$ we have

$$\nabla J_u + C^\top \lambda' = \lim_{i \mapsto \infty} (\nabla J_{u^{i(k)}} + C^\top \lambda^{i(k)}) = 0. \tag{$*_6$}$$

If C has rank m, then $\mathrm{Im}(C) = \mathbb{R}^m$, which is equivalent to $\mathrm{Ker}\,(C^\top) = (0)$, so $C^\top$ is injective and since by $(\dagger_1)$ we also have $\nabla J_u + C^\top \lambda = 0$, we conclude that $\lambda' = \lambda$. The above reasoning applies to any subsequence of $(\lambda^k)_{k\geq 0}$, so $(\lambda^k)_{k\geq 0}$ converges to λ. $\square$

In the special case where J is an elliptic quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle,$$

where A is symmetric positive definite, by $(\dagger_2)$ an iteration of Uzawa's method gives

$$\begin{aligned} &Au^k - b + C^\top \lambda^k = 0 \\ &\lambda_i^{k+1} = \max\{(\lambda^k + \rho(Cu^k - d))_i, 0\}, \quad 1 \leq i \leq m. \end{aligned}$$

Theorem 50.21 implies that Uzawa's method converges if

$$0 < \rho < \frac{2\lambda_1}{\|C\|_2^2},$$

where λ_1 is the smallest eigenvalue of A.

If we solve for u^k using the first equation, we get

$$\lambda^{k+1} = p_+(\lambda^k + \rho(-CA^{-1}C^\top \lambda^k + CA^{-1}b - d)). \tag{$*_7$}$$

In Example 50.7 we showed that the gradient of the dual function G is given by

$$\nabla G_\mu = Cu_\mu - d = -CA^{-1}C^\top \mu + CA^{-1}b - d,$$

so $(*_7)$ can be written as

$$\lambda^{k+1} = p_+(\lambda^k + \rho \nabla G_{\lambda^k});$$

this shows that Uzawa's method is indeed the gradient method with fixed stepsize applied to the dual program.

50.14 Summary

The main concepts and results of this chapter are listed below:

- The cone of feasible directions.
- Cone with apex.
- Active and inactive constraints.
- Qualified constraint at u.
- Farkas lemma.
- Farkas–Minkowski lemma.
- Karush–Kuhn–Tucker optimality conditions (or KKT-conditions).
- Complementary slackness conditions.
- Generalized Lagrange multipliers.
- Qualified convex constraint.
- Lagrangian of a minimization problem.
- Equality constrained minimization.
- KKT matrix.
- Newton's method with equality constraints (feasible start and infeasible start).
- Hard margin support vector machine
- Training data
- Linearly separable sets of points.
- Maximal margin hyperplane.
- Support vectors
- Saddle points.
- Lagrange dual function.
- Lagrange dual program.
- Duality gap.

- Weak duality.
- Strong duality.
- Handling equality constraints in the Lagrangian.
- Dual of the Hard Margin SVM (SVM_{h2}).
- Conjugate functions and Legendre dual functions.
- Dual of the Hard Margin SVM (SVM_{h1}).
- Uzawa's Method.

50.15 Problems

Problem 50.1. Prove (3) and (4) of Proposition 50.11.

Problem 50.2. Assume that in Theorem 50.17, $V = \mathbb{R}^n$, J is elliptic and the constraints φ_i are of the form

$$\sum_{j=1}^{n} c_{ij} v_j \leq d_i,$$

that is, affine. Prove that the Problem (P_μ) has a unique solution which is continuous in μ.

Problem 50.3. (1) Prove that the set of saddle points of a function $L\colon \Omega \times M \to \mathbb{R}$ is of the form $V_0 \times M_0$, for some $V_0 \subseteq \Omega$ and some $M_0 \subseteq M$.

(2) Assume that Ω and M are convex subsets of some normed vector spaces, assume that for any fixed $v \in \Omega$ the map

$$\mu \mapsto L(v, \mu) \quad \text{is concave,}$$

and for any fixed $\mu \in M$ the map

$$v \mapsto L(v, \mu) \quad \text{is convex.}$$

Prove that $V_0 \times M_0$ is convex.

(3) Prove that if for every fixed $\mu \in M$ the map

$$v \mapsto L(v, \mu) \quad \text{is strictly convex,}$$

then V_0 has a most one element.

Problem 50.4. Prove that the conjugate of the function f given by $f(X) = \log\det(X^{-1})$, where X is an $n \times n$ symmetric positive definite matrix, is

$$f^*(Y) = \log\det((-Y)^{-1}) - n,$$

where Y is an $n \times n$ symmetric negative definite matrix.

Problem 50.5. (From Boyd and Vandenberghe [29], Problem 5.12) Given an $m \times n$ matrix A and any vector $b \in \mathbb{R}^n$, consider the problem

$$\begin{aligned} &\text{minimize} \quad -\sum_{i=1}^{m} \log(b_i - a_i x) \\ &\text{subject to} \;\; Ax < b, \end{aligned}$$

where a_i is the ith row of A. This is called the *analytic centering problem*. It can be shown that the problem has a unique solution iff the open polyhedron $\{x \in \mathbb{R}^n \mid Ax < b\}$ is nonempty and bounded.

(1) Prove that necessary and sufficient conditions for the problem to have an optimal solution are

$$Ax < b, \quad \sum_{i=1}^{m} \frac{a_i^\top}{b_i - a_x x} = 0.$$

(2) Derive a dual program for the above program.
Hint. First introduce new variables y_i and equations $y_i = b_i - a_i x$.

Problem 50.6. (From Boyd and Vandenberghe [29], Problem 5.13) A *Boolean linear program* is the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad c^\top x \\ &\text{subject to} \;\; Ax \leq b \\ &\qquad\qquad\quad x_i \in \{0, 1\}, \; i = 1, \ldots, n, \end{aligned}$$

where A is an $m \times n$ matrix, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. The fact that the solutions $x \in \mathbb{R}^n$ are constrained to have coordinates x_i taking the values 0 or 1 makes it a hard problem. The above problem can be stated as a program with quadratic constraints:

$$\begin{aligned} &\text{minimize} \quad c^\top x \\ &\text{subject to} \;\; Ax \leq b \\ &\qquad\qquad\quad x_i(1 - x_i) = 0, \; i = 1, \ldots, n. \end{aligned}$$

(1) Derive the Lagrangian dual of the above program.

(2) A way to approximate a solution of the Boolean linear program is to consider its linear relaxation where the constraints $x_i \in \{0, 1\}$ are replaced by the *linear constraints* $0 \leq x_i \leq 1$:

$$\begin{aligned} &\text{minimize} \quad c^\top x \\ &\text{subject to} \;\; Ax \leq b \\ &\qquad\qquad\quad 0 \leq x_i \leq 1, \; i = 1, \ldots, n. \end{aligned}$$

Find the dual linear program of the above linear program. Show that the maxima of the dual programs in (1) and (2) are the same.

Chapter 51

Subgradients and Subdifferentials of Convex Functions $\circledast$

In this chapter we consider some deeper aspects of the theory of convex functions that are not necessarily differentiable at every point of their domain. Some substitute for the gradient is needed. Fortunately, for convex functions, there is such a notion, namely subgradients. Geometrically, given a (proper) convex function f, the subgradients at x are vectors normal to supporting hyperplanes to the epigraph of the function at $(x, f(x))$. The subdifferential $\partial f(x)$ to f at x is the set of all subgradients at x. A crucial property is that f is differentiable at x iff $\partial f(x) = \{\nabla f_x\}$, where ∇f_x is the gradient of f at x. Another important property is that a (proper) convex function f attains its minimum at x iff $0 \in \partial f(x)$. A major motivation for developing this more sophisticated theory of "differentiation" of convex functions is to extend the Lagrangian framework to convex functions that are not necessarily differentiable.

Experience shows that the applicability of convex optimization is significantly increased by considering extended real-valued functions, namely functions $f\colon S \to \mathbb{R} \cup \{-\infty, +\infty\}$, where S is some subset of $\mathbb{R}^n$ (usually convex). This is reminiscent of what happens in measure theory, where it is natural to consider functions that take the value $+\infty$. We already encountered functions that take the value $-\infty$ as a result of a minimization that does not converge. For example, if $J(u, v) = u$, and we have the affine constraint $v = 0$, for any fixed λ, the minimization problem

$$\begin{aligned} &\text{minimize} && u + \lambda v \\ &\text{subject to} && v = 0, \end{aligned}$$

yields the solution $u = -\infty$ and $v = 0$.

Until now, we chose not to consider functions taking the value $-\infty$, and instead we considered partial functions, but it turns out to be convenient to admit functions taking the value $-\infty$.

Allowing functions to take the value $+\infty$ is also a convenient alternative to dealing with partial functions. This situation is well illustrated by the indicator function of a convex set.

Definition 51.1. Let $C \subseteq \mathbb{R}^n$ be any subset of $\mathbb{R}^n$. The *indicator function* I_C of C is the function given by

$$I_C(u) = \begin{cases} 0 & \text{if } u \in C \\ +\infty & \text{if } u \notin C. \end{cases}$$

The indicator function I_C is a variant of the characteristic function χ_C of the set C (defined such that $\chi_C(u) = 1$ if $u \in C$ and $\chi_C(u) = 0$ if $u \notin C$). Rockafellar denotes the indicator function I_C by $\delta(-|C)$; that is, $\delta(u|C) = I_C(u)$; see Rockafellar [138], Page 28.

Given a partial function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$, by setting $f(u) = +\infty$ if $u \notin \mathrm{dom}(f)$, we convert the partial function f into a total function with values in $\mathbb{R} \cup \{-\infty, +\infty\}$. Still, one has to remember that such functions are really partial functions, but $-\infty$ and $+\infty$ play different roles. The value $f(x) = -\infty$ indicates that computing $f(x)$ using a *minimization procedure did not terminate*, but $f(x) = +\infty$ means that the *function f is really undefined at x*.

The definition of a convex function $f\colon S \to \mathbb{R} \cup \{-\infty, +\infty\}$ needs to be slightly modified to accommodate the infinite values $\pm\infty$. The cleanest definition uses the notion of epigraph.

A remarkable and very useful fact is that the optimization problem

$$\begin{aligned} &\text{minimize} && J(u) \\ &\text{subject to} && u \in C, \end{aligned}$$

where C is a closed convex set in $\mathbb{R}^n$ and J is a convex function can be rewritten in term of the indicator function I_C of C, as

$$\begin{aligned} &\text{minimize} && J(u) + I_C(z) \\ &\text{subject to} && u - z = 0. \end{aligned}$$

But $J(u) + I_C(z)$ is not differentiable, even if J is, which forces us to deal with convex functions which are not differentiable

Convex functions are not necessarily differentiable, but if a convex function f has a finite value $f(u)$ at u (which means that $f(u) \in \mathbb{R}$), then it has a one-sided directional derivative at u. Another crucial notion is the notion of subgradient, which is a substitute for the notion of gradient when the function f is not differentiable at u.

In Section 51.1, we introduce extended real-valued functions, which are functions that may also take the values $\pm\infty$. In particular, we define proper convex functions, and the closure of a convex function. Subgradients and subdifferentials are defined in Section 51.2. We discuss some properties of subgradients in Section 51.3 and Section 51.4. In particular, we relate subgradients to one-sided directional derivatives. In Section 51.5, we discuss the problem of finding the minimum of a proper convex function and give some criteria in terms of subdifferentials. In Section 51.6, we sketch the generalization of the results presented in Chapter 50 about the Lagrangian framework to programs allowing an objective function and inequality constraints which are convex but not necessarily differentiable. In fact, it is fair to say that the theory of extended real-valued convex functions and the notions of subgradient and subdifferential developed in this chapter constitute the machinery needed to extend the Lagrangian framework to convex functions that are not necessarily differentiable.

This chapter relies heavily on Rockafellar [138]. Some of the results in this chapter are also discussed in Bertsekas [16, 19, 17]. It should be noted that Bertsekas has developed a framework to discuss duality that he refers to as the *min common/max crossing* framework, for short MC/MC. Although this framework is elegant and interesting in its own right, the fact that Bertsekas relies on it to prove properties of subdifferentials makes it little harder for a reader to "jump in."

51.1 Extended Real-Valued Convex Functions

We extend the ordering on $\mathbb{R}$ by setting

$$-\infty < x < +\infty, \quad \text{for all } x \in \mathbb{R}.$$

Definition 51.2. A (total) function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is called an *extended real-valued function*. For any $x \in \mathbb{R}^n$, we say that $f(x)$ is *finite* if $f(x) \in \mathbb{R}$ (equivalently, $f(x) \neq \pm\infty$). The function f is *finite* if $f(x)$ is finite for all $x \in \mathbb{R}^n$.

Adapting slightly Definition 40.8, given a function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$, the *epigraph* of f is the subset of $\mathbb{R}^{n+1}$ given by

$$\mathbf{epi}(f) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y\}.$$

See Figure 51.1.

If S is a nonempty subset of $\mathbb{R}^n$, the epigraph of the restriction of f to S is defined as

$$\mathbf{epi}(f|S) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y,\ x \in S\}.$$

Observe the following facts:

1. For any $x \in S$, if $f(x) = -\infty$, then $\mathbf{epi}(f)$ contains the "vertical line" $\{(x, y) \mid y \in \mathbb{R}\}$ in $\mathbb{R}^{n+1}$.

2. For any $x \in S$, if $f(x) \in \mathbb{R}$, then $\mathbf{epi}(f)$ contains the ray $\{(x, y)\} \mid f(x) \leq y\}$ in $\mathbb{R}^{n+1}$.

3. For any $x \in S$, if $f(x) = +\infty$, then $\mathbf{epi}(f)$ does not contain any point (x, y), with $y \in \mathbb{R}$.

4. We have $\mathbf{epi}(f) = \emptyset$ iff f corresponds to the partial function undefined everywhere; that is, $f(x) = +\infty$ for all $x \in \mathbb{R}^n$.

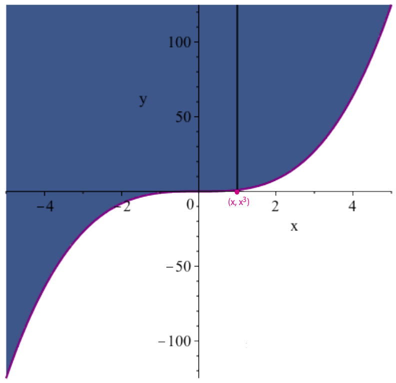


Figure 51.1: Let $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ be given by $f(x) = x^3$ for $x \in \mathbb{R}$. Its graph in $\mathbb{R}^2$ is the magenta curve, and its epigraph is the union of the magenta curve and blue region "above" this curve. For any point $x \in \mathbb{R}$, $\mathbf{epi}(f)$ contains the ray which starts at (x, x^3) and extends upward.

Definition 51.3. Given a nonempty subset S of $\mathbb{R}^n$, a (total) function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is *convex on* S if its epigraph $\mathbf{epi}(f|S)$ is convex as a subset of $\mathbb{R}^{n+1}$. See Figure 51.2. The function f is *concave* on S if $-f$ is convex on S. The function f is *affine* on S if it is finite, convex, and concave. If $S = \mathbb{R}^n$, we simply that f is *convex* (resp. *concave*, resp. *affine*).

Definition 51.4. Given any function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$, the *effective domain* $\mathrm{dom}(f)$ of f is given by

$$\mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid (\exists y \in \mathbb{R})((x, y) \in \mathbf{epi}(f))\} = \{x \in \mathbb{R}^n \mid f(x) < +\infty\}.$$

Observe that the effective domain of f contains the vectors $x \in \mathbb{R}^n$ such that $f(x) = -\infty$, but excludes the vectors $x \in \mathbb{R}^n$ such that $f(x) = +\infty$.

Example 51.1. The above fact is illustrated by the function $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ where

$$f(x) = \begin{cases} -x^2 & \text{if } x \geq 0 \\ +\infty & \text{if } x < 0. \end{cases}$$

The epigraph of this function is illustrated Figure 51.3. By definition $\mathrm{dom}(f) = [0, \infty)$.

If f is a convex function, since $\mathrm{dom}(f)$ is the image of $\mathbf{epi}(f)$ by a linear map (a projection), *it is convex*.

By definition, $\mathbf{epi}(f|S)$ is convex iff for any (x_1, y_1) and (x_2, y_2) with $x_1, x_2 \in S$ and $y_1, y_2 \in \mathbb{R}$ such that $f(x_1) \leq y_1$ and $f(x_2) \leq y_2$, for every λ such that $0 \leq \lambda \leq 1$, we have

$$(1 - \lambda)(x_1, y_1) + \lambda(x_2, y_2) = ((1 - \lambda)x_1 + \lambda x_2, (1 - \lambda)y_1 + \lambda y_2) \in \mathbf{epi}(f|S),$$

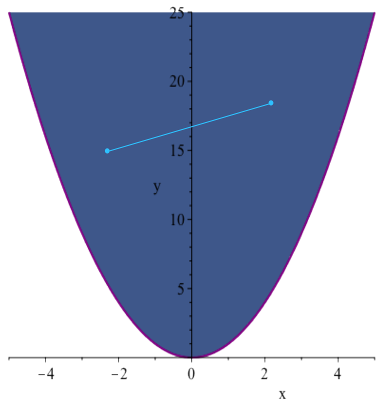


Figure 51.2: Let $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ be given by $f(x) = x^2$ for $x \in \mathbb{R}$. Its graph in $\mathbb{R}^2$ is the magenta curve, and its epigraph is the union of the magenta curve and blue region "above" this curve. Observe that $\mathbf{epi}(f)$ is a convex set of $\mathbb{R}^2$ since the aqua line segment connecting any two points is contained within the epigraph.

which means that $(1-\lambda)x_1 + \lambda x_2 \in S$ and

$$f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda)y_1 + \lambda y_2. \tag{$*$}$$

Thus S *must be convex* and $f((1-\lambda)x_1 + \lambda x_2) < +\infty$. Condition $(*)$ is a little awkward, since it does not refer explicitly to $f(x_1)$ and $f(x_2)$, as these values may be $-\infty$, in which case it is not clear what the expression $(1-\lambda)f(x_1) + \lambda f(x_2)$ means.

In order to perform arithmetic operations involving $-\infty$ and $+\infty$, we adopt the following conventions:

$$\begin{aligned}
\alpha + (+\infty) = +\infty + \alpha &= +\infty & \quad & -\infty < \alpha \leq +\infty \\
\alpha + -\infty = -\infty + \alpha &= -\infty & & -\infty \leq \alpha < +\infty \\
\alpha(+\infty) = (+\infty)\alpha &= +\infty & & 0 < \alpha \leq +\infty \\
\alpha(-\infty) = (-\infty)\alpha &= -\infty & & 0 < \alpha \leq +\infty \\
\alpha(+\infty) = (+\infty)\alpha &= -\infty & & -\infty \leq \alpha < 0 \\
\alpha(-\infty) = (-\infty)\alpha &= +\infty & & -\infty \leq \alpha < 0 \\
0(+\infty) = (+\infty)0 &= 0 & & 0(-\infty) = (-\infty)0 = 0 \\
-(-\infty) &= +\infty & & \\
\inf \emptyset &= +\infty & & \sup \emptyset = -\infty.
\end{aligned}$$

The expressions $+\infty + (-\infty)$ and $-\infty + (+\infty)$ are *meaningless*.

The following characterizations of convex functions are easy to show.

Proposition 51.1. *Let C be a nonempty convex subset of $\mathbb{R}^n$.*

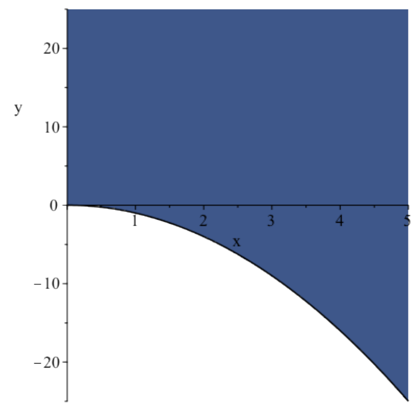


Figure 51.3: The epigraph of the concave function $f(x) = -x^2$ if $x \geq 0$ and $+\infty$ otherwise.

(1) A function $f\colon C \to \mathbb{R}^n \cup \{+\infty\}$ *is convex on* C *iff*

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

for all $x, y \in C$ *and all* λ *such that* $0 < \lambda < 1$.

(2) A function $f\colon \mathbb{R}^n \to \mathbb{R}^n \cup \{-\infty, +\infty\}$ *is convex iff*

$$f((1-\lambda)x + \lambda y) < (1-\lambda)\alpha + \lambda\beta$$

for all $\alpha, \beta \in \mathbb{R}$, *all* $x, y \in \mathbb{R}^n$ *such that* $f(x) < \alpha$ *and* $f(y) < \beta$, *and all* λ *such that* $0 < \lambda < 1$.

The "good" convex functions that we would like to deal with are defined below.

Definition 51.5. A convex function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is *proper*[1] if its epigraph is nonempty and does not contain any vertical line. Equivalently, a convex function f is proper if $f(x) > -\infty$ for all $x \in \mathbb{R}^n$ and $f(x) < +\infty$ for some $x \in \mathbb{R}^n$. A convex function which is not proper is called an *improper function*.

Observe that a convex function f is proper iff $\mathrm{dom}(f) \neq \emptyset$ and if the restriction of f to $\mathrm{dom}(f)$ is a finite function.

It is immediately verified that a set C is convex iff its indicator function I_C is convex, and clearly, the indicator function of a convex set is proper.

The important object of study is the set of proper functions, but improper functions can't be avoided.

[1]This terminology is unfortunate because it clashes with the notion of a proper function from topology, which has to do with the preservation of compact subsets under inverse images.

Example 51.2. Here is an example of an improper convex function $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$:

$$f(x) = \begin{cases} -\infty & \text{if } |x| < 1 \\ 0 & \text{if } |x| = 1 \\ +\infty & \text{if } |x| > 1 \end{cases}$$

Observe that $\mathrm{dom}(f) = [-1, 1]$, and that $\mathbf{epi}(f)$ is not closed. See Figure 51.4.

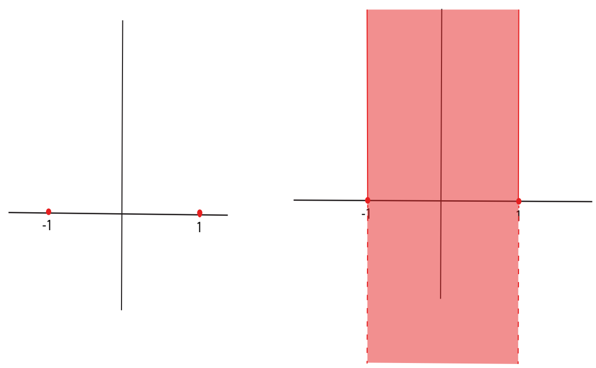


Figure 51.4: The improper convex function of Example 51.2 and its epigraph depicted as a rose colored region of $\mathbb{R}^2$.

Functions whose epigraph are closed tend to have better properties. To characterize such functions we introduce sublevel sets.

Definition 51.6. Given a function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$, for any $\alpha \in \mathbb{R} \cup \{-\infty, +\infty\}$, the *sublevel sets* $\mathrm{sublev}_\alpha(f)$ and $\mathrm{sublev}_{<\alpha}(f)$ are the sets

$$\mathrm{sublev}_\alpha(f) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\} \quad \text{and} \quad \mathrm{sublev}_{<\alpha}(f) = \{x \in \mathbb{R}^n \mid f(x) < \alpha\}.$$

For the improper convex function of Example 51.2, we have

$\mathrm{sublev}_{-\infty}(f) = (-1, 1)$ while $\mathrm{sublev}_{<-\infty}(f) = \emptyset$.

$\mathrm{sublev}_\alpha(f) = (-1, 1) = \mathrm{sublev}_{<\alpha}(f)$ whenever $-\infty < \alpha < 0$.

$\mathrm{sublev}_0(f) = [-1, 1]$ while $\mathrm{sublev}_{<0}(f) = (-1, 1)$.

$\mathrm{sublev}_\alpha(f) = [-1, 1] = \mathrm{sublev}_{<\alpha}(f)$ whenever $0 < \alpha < +\infty$.

$\mathrm{sublev}_{+\infty}(f) = \mathbb{R}$ while $\mathrm{sublev}_{<+\infty}(f) = [-1, 1]$.

A useful corollary of Proposition 51.1 is the following result whose (easy) proof can be found in Rockafellar [138] (Theorem 4.6).

Proposition 51.2. *If f is any convex function on $\mathbb{R}^n$, then for every $\alpha \in \mathbb{R} \cup \{-\infty, +\infty\}$, the sublevel sets* $\mathrm{sublev}_\alpha(f)$ *and* $\mathrm{sublev}_{<\alpha}(f)$ *are convex.*

Definition 51.7. A function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is *lower semi-continuous* if the sublevel sets $\mathrm{sublev}_\alpha(f) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ are closed for all $\alpha \in \mathbb{R}$.

Observe that the improper convex function of Example 51.2 is not lower semi-continuous since $\mathrm{sublev}_\alpha(f) = (-1, 1)$ whenever $-\infty < \alpha < 0$. This result reflects the fact that the epigraph is not closed as shown in the following proposition; see Rockafellar [138] (Theorem 7.1).

Proposition 51.3. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be any function. The following properties are equivalent:*

(1) The function f is lower semi-continuous.

(2) The epigraph of f is a closed set in $\mathbb{R}^{n+1}$.

The notion of the closure of convex function plays an important role. It is a bit subtle because a convex function may be improper.

Definition 51.8. Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be any function. The function whose epigraph is the closure of the epigraph $\mathbf{epi}(f)$ of f (in $\mathbb{R}^{n+1}$) is called the *lower semi-continuous hull* of f. If f is a convex function and if $f(x) > -\infty$ for all $x \in \mathbb{R}^n$, then the *closure* $\mathrm{cl}(f)$ *of* f is equal to its lower semi-continuous hull, else if $f(x) = -\infty$ for some $x \in \mathbb{R}^n$, then the *closure* $\mathrm{cl}(f)$ *of* f is the constant function with value $-\infty$. A convex function f is *closed* if $f = \mathrm{cl}(f)$.

Definition 51.8 implies that there are *only two closed improper convex functions*: the constant function with value $-\infty$ and the constant function with value $+\infty$. Also, by Proposition 51.3, *a proper convex function is closed iff it is equal to its lower semi-continuous hull iff its epigraph is nonempty and closed*.

Given a convex set C in $\mathbb{R}^n$, the interior $\mathrm{int}(C)$ of C (the largest open subset of $\mathbb{R}^n$ contained in C) is often not interesting because C may have dimension smaller than n. For example, a (closed) triangle in $\mathbb{R}^3$ has empty interior.

The remedy is to consider the affine hull $\mathrm{aff}(C)$ of C, which is the smallest affine set containing C; see Section 44.2. The dimension of C is the dimension of $\mathrm{aff}(C)$. Then the relative interior of C is the interior of C in $\mathrm{aff}(C)$ endowed with the subspace topology induced on $\mathrm{aff}(C)$. More explicitly, we can make the following definition.

Definition 51.9. Let C be a subset of $\mathbb{R}^n$. The *relative interior* of C is the set

$$\mathbf{relint}(C) = \{x \in C \mid B_\epsilon(x) \cap \mathrm{aff}(C) \subseteq C \quad \text{for some } \epsilon > 0\},$$

where $B_\epsilon(x) = \{y \in \mathbb{R}^n \mid \|x - y\|_2 < \epsilon\}$, the open ball of center x and radius ϵ. The *relative boundary* of C is defined as $\overline{C} - \mathbf{relint}(C)$, where $\overline{C}$ is the closure of C in $\mathbb{R}^n$ (the smallest closed subset of $\mathbb{R}^n$ containing C).

Remark. Observe that $\text{int}(C) \subseteq \textbf{relint}(C)$. Rockafellar denotes the relative interior of a set C by $\textbf{ri}(C)$.

The following result from Rockafellar [138] (Theorem 7.2) tells us that an improper convex function mostly takes infinite values, except perhaps at relative boundary points of its effective domain.

Proposition 51.4. *If f is an improper convex function, then $f(x) = -\infty$ for every $x \in \textbf{relint}(\text{dom}(f))$. Thus an improper convex function takes infinite values, except at relative boundary points of its effective domain.*

Example 51.2 illustrates Proposition 51.4.

The following result also holds; see Rockafellar [138] (Corollary 7.2.3).

Proposition 51.5. *If f is a convex function whose effective domain is relatively open, which means that $\textbf{relint}(\text{dom}(f)) = \text{dom}(f)$, then either $f(x) > -\infty$ for all $x \in \mathbb{R}^n$, or $f(x) = \pm\infty$ for all $x \in \mathbb{R}^n$.*

We also have the following result showing that the closure of a proper convex function does not differ much from the original function; see Rockafellar [138] (Theorem 7.4).

Proposition 51.6. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then $\text{cl}(f)$ is a closed proper convex function, and $\text{cl}(f)$ agrees with f on $\text{dom}(f)$ except possibly at relative boundary points.*

Example 51.3. For an example of Propositions 51.6 and 51.5, let $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be the proper convex function

$$f(x) = \begin{cases} x^2 & \text{if } x < 1 \\ +\infty & \text{if } |x| \geq 1. \end{cases}$$

Then $\text{cl}(f)$ is

$$\text{cl} f(x) = \begin{cases} x^2 & \text{if } x \leq 1 \\ +\infty & \text{if } |x| > 1, \end{cases}$$

and $\text{cl} f(x) = f(x)$ whenever $x \in (-\infty, 1) = \textbf{relint}(\text{dom}(f)) = \text{dom}(f)$. Furthermore, since $\textbf{relint}(\text{dom}(f)) = \text{dom}(f)$, $f(x) > -\infty$ for all $x \in \mathbb{R}$. See Figure 51.5.

Small miracle: *the indicator function I_C of any closed convex set is proper and closed.* Indeed, for any $\alpha \in \mathbb{R}$ the sublevel set $\{x \in \mathbb{R}^n \mid I_C(x) \leq \alpha\}$ is either empty if $\alpha < 0$, or equal to C if $\alpha \geq 0$, and C is closed.

We now discuss briefly continuity properties of convex functions. The fact that a convex function f can take the values $\pm\infty$ causes a difficulty, so we consider the restriction of f to its effective domain. There is still a problem because an improper function may take the value $-\infty$. However, if we consider any subset C of $\text{dom}(f)$ which is relatively open, which means that $\textbf{relint}(C) = C$, then $C \subseteq \textbf{relint}(\text{dom}(f))$, so by Proposition 51.4, the function

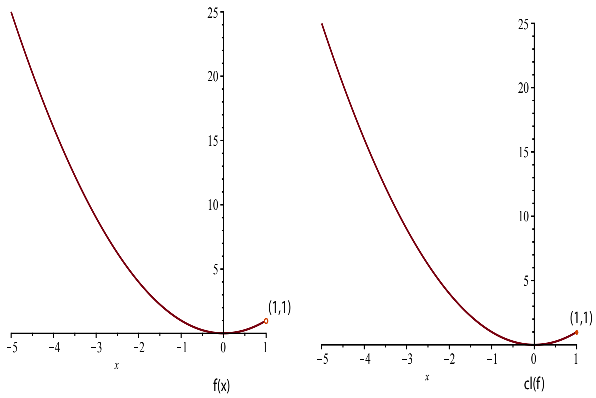


Figure 51.5: The proper convex function of Example 51.3 and its closure. These two functions only differ at the relative boundary point of $\mathrm{dom}(f)$, namely $x = 1$.

f has the constant value $-\infty$ on C, and so it can be considered to be continuous on C. Thus we are led to consider proper functions.

Definition 51.10. Given a proper convex function f, for any subset $S \subseteq \mathrm{dom}(f)$, we say that f is *continuous relative to* S if the restriction of f to S is continuous, with S endowed with the subspace topology.

The following result is proven in Rockafellar [138] (Theorem 10.1).

Proposition 51.7. *If f is a proper convex function, then f is continuous on any convex relatively open subset C (*$\mathbf{relint}(C) = C$*) contained in its effective domain* $\mathrm{dom}(f)$*, in particular relative to* $\mathbf{relint}(\mathrm{dom}(f))$*.*

As a corollary, any convex function f which is finite on $\mathbb{R}^n$ is continuous.

The behavior of a convex function at relative boundary points of the effective domain can be tricky. Here is an example due to Rockafellar [138] illustrating the problems.

Example 51.4. Consider the proper convex function (on $\mathbb{R}^2$) given by

$$f(x,y) = \begin{cases} y^2/(2x) & \text{if } x > 0 \\ 0 & \text{if } x = 0, y = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

We have

$$\mathrm{dom}(f) = \{(x,y) \in \mathbb{R}^2 \mid x > 0\} \cup \{(0,0)\}.$$

See Figure 51.6.

The function f is continuous on the open right half-plane $\{(x,y) \in \mathbb{R}^2 \mid x > 0\}$, but not at $(0,0)$. The limit of $f(x,y)$ when (x,y) approaches $(0,0)$ on the parabola of equation

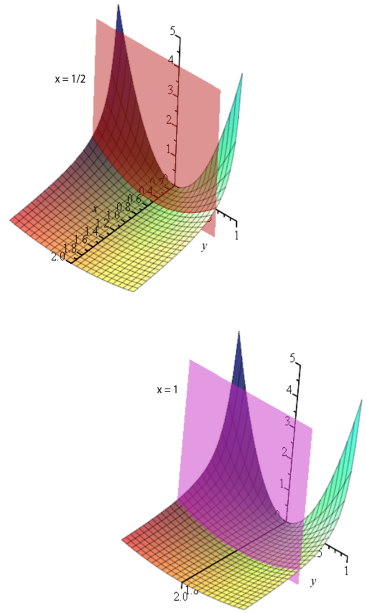


Figure 51.6: The proper convex function of Example 51.4. When intersected by vertical planes of the form $x = \alpha$, for $\alpha > 0$, the trace is an upward parabola. When α is close to zero, this parabola approximates the positive z axis.

$x = y^2/(2\alpha)$ is α for any $\alpha > 0$. See Figure 51.7 However, it is easy to see that the limit along any line segment from $(0, 0)$ to a point in the open right half-plane is 0.

We conclude this quick tour of the basic properties of convex functions with a result involving the Lipschitz condition.

Definition 51.11. Let $f\colon E \to F$ be a function between normed vector spaces E and F, and let U be a nonempty subset of E. We say that f *Lipschitzian on* U (or *has the Lipschitz condition on* U) if there is some $c \geq 0$ such that

$$\|f(x) - f(y)\|_F \leq c \|x - y\|_E \quad \text{for all } x, y \in U.$$

Obviously, if f is Lipschitzian on U it is uniformly continuous on U. The following result is proven in Rockafellar [138] (Theorem 10.4).

Proposition 51.8. *Let f be a proper convex function, and let S be any (nonempty) closed bounded subset of* $\textbf{relint}(\text{dom}(f))$. *Then f is Lipschitzian on S.*

In particular, a finite convex function on $\mathbb{R}^n$ is Lipschitzian on every compact subset of $\mathbb{R}^n$. However, such a function may not be Lipschitzian on $\mathbb{R}^n$ as a whole.

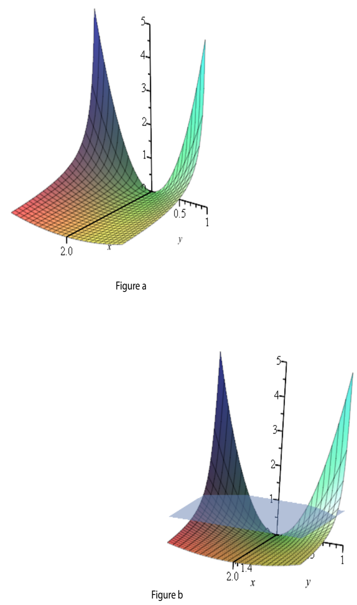


Figure 51.7: Figure (a) illustrates the proper convex function of Example 51.4. Figure (b) illustrates the approach to $(0, 0)$ along the planar parabolic curve $(y^2/2, y)$. Then $f(y^2/2, y) = 1$ and Figure b shows the intersection of the surface with the plane $z = 1$.

51.2 Subgradients and Subdifferentials

We saw in the previous section that proper convex functions have "good" continuity properties. Remarkably, if f is a convex function, for any $x \in \mathbb{R}^n$ such that $f(x)$ is finite, the one-sided derivative $f'(x; u)$ exists for all $u \in \mathbb{R}^n$; This result has been shown at least since 1893, as noted by Stoltz (see Rockafellar [138], page 428). Directional derivatives will be discussed in Section 51.3. If f is differentitable at x, then of course

$$df_x(u) = \langle \nabla f_x, u \rangle \quad \text{for all } u \in \mathbb{R}^n,$$

where ∇f_x is the gradient of f at x.

But even if f is not differentiable at x, it turns out that for "most" $x \in \text{dom}(f)$, in particular if $x \in \textbf{relint}(\text{dom}(f))$, there is a nonempty closed convex set $\partial f(x)$ which may be viewed as a generalization of the gradient ∇f_x. This convex set of $\mathbb{R}^n$, $\partial f(x)$, called the *subdifferential of f at x*, has some of the properties of the gradient ∇f_x. The vectors in $\partial f(x)$ are called *subgradients* at x. For example, if f is a proper convex function, then f achieves its minimum at $x \in \mathbb{R}^n$ iff $0 \in \partial f(x)$. Some of the theorems of Chapter 50 can be generalized to convex functions that are not necessarily differentiable by replacing conditions involving gradients by conditions involving subdifferentials. These generalizations are crucial for the justification that various iterative methods for solving optimization programs converge. For example, they are used to prove the convergence of the ADMM method discussed in Chapter 52.

One should note that the notion of subdifferential is not just a gratuitous mathematical generalization. The remarkable fact that the optimization problem

$$\begin{aligned} &\text{minimize} \quad J(u) \\ &\text{subject to} \quad u \in C, \end{aligned}$$

where C is a closed convex set in $\mathbb{R}^n$ can be rewritten as

$$\begin{aligned} &\text{minimize} \quad J(u) + I_C(z) \\ &\text{subject to} \quad u - z = 0, \end{aligned}$$

where I_C is the indicator function of C, forces us to deal with functions such as $J(u) + I_C(z)$ which are not differentiable, even if J is. ADMM can cope with this situation (under certain conditions), and subdifferentials cannot be avoided in justifying its convergence. However, it should be said that the subdifferential $\partial f(x)$ is a theoretical tool that is never computed in practice (except in very special simple cases).

To define subgradients we need to review (affine) hyperplanes.

Recall that an *affine form* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is a function of the form

$$\varphi(x) = h(x) + c, \quad x \in \mathbb{R}^n,$$

where $h\colon \mathbb{R}^n \to \mathbb{R}$ is a linear form and $c \in \mathbb{R}$ is some constant. An *affine hyperplane* $H \subseteq \mathbb{R}^n$ is the kernel of any nonconstant affine form $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ (which means that the linear form h defining φ is not the zero linear form),

$$H = \varphi^{-1}(0) = \{x \in \mathbb{R}^n \mid \varphi(x) = 0\}.$$

Any two nonconstant affine forms φ and ψ defining the same (affine) hyperplane H, in the sense that $H = \varphi^{-1}(0) = \psi^{-1}(0)$, must be proportional, which means that there is some nonzero $\alpha \in \mathbb{R}$ such that $\psi = \alpha\varphi$.

A nonconstant affine form φ also defines the two *half spaces* H_+ and H_- given by

$$H_+ = \{x \in \mathbb{R}^n \mid \varphi(x) \geq 0\}, \quad H_- = \{x \in \mathbb{R}^n \mid \varphi(x) \leq 0\}.$$

Clearly, $H_+ \cap H_- = H$, their common boundary. See Figure 51.8. The choice of sign is somewhat arbitrary, since the affine form $\alpha\varphi$ with $\alpha < 0$ defines the half spaces with H_- and H_+ (the half spaces are swapped).

By the duality induced by the Euclidean inner product on $\mathbb{R}^n$, a linear form $h\colon \mathbb{R}^n \to \mathbb{R}$ corresponds to a *unique* vector $u \in \mathbb{R}^n$ such that

$$h(x) = \langle x, u \rangle \quad \text{for all } x \in \mathbb{R}^n.$$

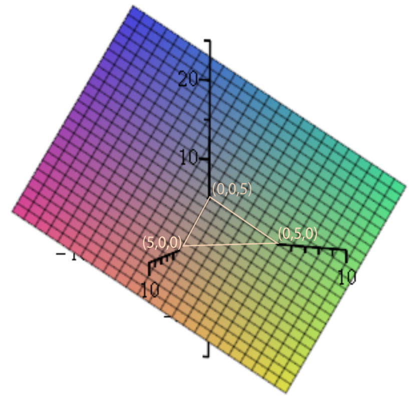


Figure 51.8: The affine hyperplane $H = \{x \in \mathbb{R}^3 \mid x + y + z - 2 = 0\}$. The half space H_+ faces the viewer and contains the point $(0, 0, 10)$, while the half space H_- is behind H and contains the point $(0, 0, 0)$.

Then if φ is the affine form given by $\varphi(x) = \langle x, u\rangle + c$, this affine form is nonconstant iff $u \neq 0$, and u is normal to the hyperplane H, in the sense that if $x_0 \in H$ is any fixed vector in H, and x is any vector in H, then $\langle x - x_0, u\rangle = 0$.

Indeed, $x_0 \in H$ means that $\langle x_0, u\rangle + c = 0$, and $x \in H$ means that $\langle x, u\rangle + c = 0$, so we get $\langle x_0, u\rangle = \langle x, u\rangle$, which implies $\langle x - x_0, u\rangle = 0$.

Here is an observation which plays a key role in defining the notion of subgradient. An illustration of the following proposition is provided by Figure 51.9.

Proposition 51.9. *Let $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ be a nonconstant affine form. Then the map $\omega\colon \mathbb{R}^{n+1} \to \mathbb{R}$ given by*

$$\omega(x, \alpha) = \varphi(x) - \alpha, \quad x \in \mathbb{R}^n,\ \alpha \in \mathbb{R},$$

is a nonconstant affine form defining a hyperplane $\mathcal{H} = \omega^{-1}(0)$ which is the graph of the affine form φ. Furthermore, this hyperplane is nonvertical in $\mathbb{R}^{n+1}$, in the sense that $\mathcal{H}$ cannot be defined by a nonconstant affine form $(x, \alpha) \mapsto \psi(x)$ which does not depend on α.

Proof. Indeed, φ is of the form $\varphi(x) = h(x) + c$ for some nonzero linear form h, so

$$\omega(x, \alpha) = h(x) - \alpha + c.$$

Since h is linear, the map $(x, \alpha) = h(x) - \alpha$ is obviously linear and nonzero, so ω is a nonconstant affine form defining a hyperplane $\mathcal{H}$ in $\mathbb{R}^{n+1}$. By definition,

$$\mathcal{H} = \{(x, \alpha) \in \mathbb{R}^{n+1} \mid \omega(x, \alpha) = 0\} = \{(x, \alpha) \in \mathbb{R}^{n+1} \mid \varphi(x) - \alpha = 0\},$$

which is the graph of φ. If $\mathcal{H}$ was a vertical hyperplane, then $\mathcal{H}$ would be defined by a nonconstant affine form ψ independent of α, but the affine form ω given by $\omega(x, \alpha) = \varphi(x) - \alpha$ and the affine form $\psi(x)$ can't be proportional, a contradiction. $\square$

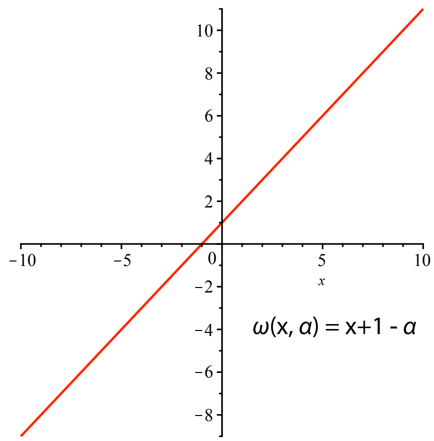


Figure 51.9: Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be the affine form $\varphi(x) = x + 1$. Let $\omega\colon \mathbb{R}^2 \to \mathbb{R}$ be the affine form $\omega(x, \alpha) = x + 1 - \alpha$. The hyperplane $\mathcal{H} = \omega^{-1}(0)$ is the red line with equation $x - \alpha + 1 = 0$.

We say that $\mathcal{H}$ is the *hyperplane (in $\mathbb{R}^{n+1}$) induced by the affine form $\varphi\colon \mathbb{R}^n \to \mathbb{R}$*. Also recall the notion of supporting hyperplane to a convex set.

Definition 51.12. If C is a nonempty convex set in $\mathbb{R}^n$ and x is a vector in C, an affine hyperplane H is a *supporting hyperplane to C at x* if

(1) $x \in H$.

(2) Either $C \subseteq H_+$ or $C \subseteq H_-$.

See Figure 51.10. Equivalently, there is some nonconstant affine form φ such that $\varphi(z) = \langle z, u\rangle - c$ for all $z \in \mathbb{R}^n$, for some nonzero $u \in \mathbb{R}^n$ and some $c \in \mathbb{R}$, such that

(1) $\langle x, u\rangle = c$.

(2) $\langle z, u\rangle \leq c$ for all $z \in C$

The notion of vector normal to a convex set is defined as follows.

Definition 51.13. Given a nonempty convex set C in $\mathbb{R}^n$, for any $a \in C$, a vector $u \in \mathbb{R}^n$ is *normal to C at a* if

$$\langle z - a, u\rangle \leq 0 \quad \text{for all } z \in C.$$

In other words, u does not make an acute angle with any line segment in C with a as endpoint. The set of all vectors u normal to C is called the *normal cone* to C at a and is denoted by $N_C(a)$. See Figure 51.11.

It is easy to check that the normal cone to C at a is a convex cone. Also, if the hyperplane H defined by an affine form $\varphi(z) = \langle z, u\rangle - c$ with $u \neq 0$ is a supporting hyperplane to C at

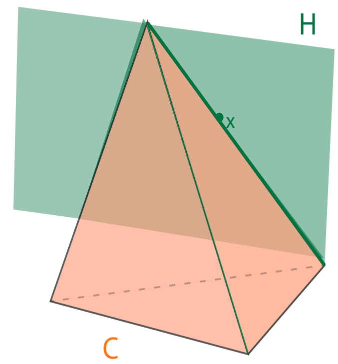


Figure 51.10: Let C be the solid peach tetrahedron in $\mathbb{R}^3$. The green plane H is a supporting hyperplane to the point x since $x \in H$ and $C \subseteq H_+$, i.e. H only intersects C on the edge containing x and so the tetrahedron lies in "front" of H.

x, since $\langle z, u\rangle \leq c$ for all $z \in C$ and $\langle x, u\rangle = c$, we have $\langle z - x, u\rangle \leq 0$ for all $z \in C$, *which means that u is normal to C at x*. This concept is illustrated by Figure 51.12.

The notion of subgradient can be motived as follows. A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$ if

$$f(x + y) = f(x) + df_x(y) + \epsilon(y) \left\|y\right\|_2,$$

for all $y \in \mathbb{R}^n$ in some nonempty subset containing x, where $df_x\colon \mathbb{R}^n \to \mathbb{R}$ is a linear form, and ϵ is some function such that $\lim_{\|y\|\mapsto 0} \epsilon(y) = 0$. Furthermore,

$$df_x(y) = \langle y, \nabla f_x\rangle \quad \text{for all } y \in \mathbb{R}^n,$$

where ∇f_x is the *gradient* of f at x, so

$$f(x + y) = f(x) + \langle y, \nabla f_x\rangle + \epsilon(y) \left\|y\right\|_2.$$

If we assume that f is convex, it makes sense to replace the equality sign by the inequality sign $\geq$ in the above equation and to drop the "error term" $\epsilon(y) \left\|y\right\|_2$, so a vector u is a subgradient of f at x if

$$f(x + y) \geq f(x) + \langle y, u\rangle \quad \text{for all } y \in \mathbb{R}^n.$$

Thus we are led to the following definition.

Definition 51.14. Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be a convex function. For any $x \in \mathbb{R}^n$, a *subgradient of f at x* is any vector $u \in \mathbb{R}^n$ such that

$$f(z) \geq f(x) + \langle z - x, u\rangle, \quad \text{for all } z \in \mathbb{R}^n. \qquad (*_{\text{subgrad}})$$

The above inequality is called the *subgradient inequality*. The set of all subgradients of f at x is denoted $\partial f(x)$ and is called the *subdifferential* of f at x. If $\partial f(x) \neq \emptyset$, then we say that *$f$ is subdifferentiable at x*.

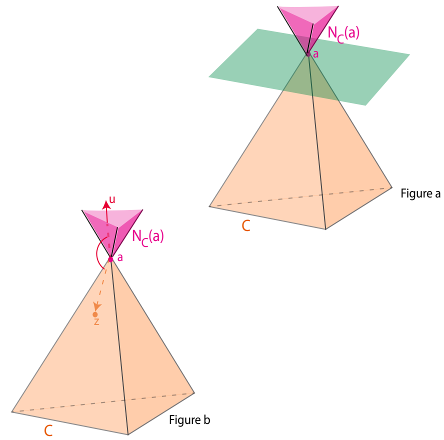


Figure 51.11: Let C be the solid peach tetrahedron in $\mathbb{R}^3$. The small upside-down magenta tetrahedron is the translate of $N_C(a)$. Figure (a) shows that the normal cone is separated from C by the horizontal green supporting hyperplane. Figure (b) shows that any vector $u \in N_C(a)$ does not make an acute angle with a line segment in C emanating from a.

Assume that $f(x)$ is finite. Observe that the subgradient inequality says that 0 is a subgradient at x iff f has a global minimum at x. In this case, the hyperplane $\mathcal{H}$ (in $\mathbb{R}^{n+1}$) defined by the affine form $\omega(y, \alpha) = f(x) - \alpha$ is a horizontal supporting hyperplane to the epigraph $\mathbf{epi}(f)$ at $(x, f(x))$. If $u \in \partial f(x)$ and $u \neq 0$, then $(*_{\text{subgrad}})$ says that the hyperplane induced by the affine form $z \mapsto \langle z - x, u \rangle + f(x)$ as in Proposition 51.9 is a nonvertical supporting hyperplane $\mathcal{H}$ (in $\mathbb{R}^{n+1}$) to the epigraph $\mathbf{epi}(f)$ at $(x, f(x))$. The vector $(u, -1) \in \mathbb{R}^{n+1}$ is normal to the hyperplane $\mathcal{H}$. See Figure 51.13.

Indeed, if $u \neq 0$, the hyperplane $\mathcal{H}$ is given by

$$\mathcal{H} = \{(y, \alpha) \in \mathbb{R}^{n+1} \mid \omega(y, \alpha) = 0\}$$

with $\omega(y, \alpha) = \langle y - x, u \rangle + f(x) - \alpha$, so $\omega(x, f(x)) = 0$, which means that $(x, f(x)) \in \mathcal{H}$. Also, for any $(z, \beta) \in \mathbf{epi}(f)$, by $(*_{\text{subgrad}})$, we have

$$\omega(z, \beta) = \langle z - x, u \rangle + f(x) - \beta \leq f(z) - \beta \leq 0,$$

since $(z, \beta) \in \mathbf{epi}(f)$, so $\mathbf{epi}(f) \subseteq \mathcal{H}_-$, and $\mathcal{H}$ is a nonvertical supporting hyperplane (in $\mathbb{R}^{n+1}$) to the epigraph $\mathbf{epi}(f)$ at $(x, f(x))$. Since

$$\omega(y, \alpha) = \langle y - x, u \rangle + f(x) - \alpha = \langle (y - x, \alpha), (u, -1) \rangle + f(x),$$

the vector $(u, -1)$ is indeed normal to the hyperplane $\mathcal{H}$.

The above facts are important and recorded as the following proposition.

Proposition 51.10. *For every $x \in \mathbb{R}^n$, if $f(x)$ is finite, then f is subdifferentiable at x if and only if there is a nonvertical supporting hyperplane (in $\mathbb{R}^{n+1}$) to the epigraph* $\mathbf{epi}(f)$ *at*

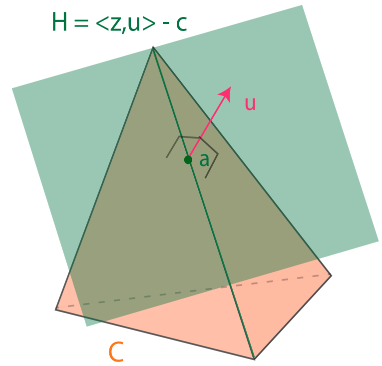


Figure 51.12: Let C be the solid peach tetrahedron in $\mathbb{R}^3$. The green plane H defined by $\varphi(z) = \langle z, u\rangle - c$ is a supporting hyperplane to C at a. The pink normal to H, namely the vector u, is also normal to C at a.

$(x, f(x))$. In this case there is an affine form φ (over $\mathbb{R}^n$) such that $f(y) \geq \varphi(y)$ for all $y \in \mathbb{R}^n$. We can pick φ given by $\varphi(y) = \langle y - x, u\rangle + f(x)$ for all $y \in \mathbb{R}^n$.

It is easy to see that $\partial f(x)$ is closed and convex. The set $\partial f(x)$ may be empty, or reduced to a single element. In $\partial f(x)$ consists of a single element it can be shown that f is finite near x, differentiable at x, and that $\partial f(x) = \{\nabla f_x\}$, the gradient of f at x.

Example 51.5. The ℓ^2 norm $f(x) = \|x\|_2$ is subdifferentiable for all $x \in \mathbb{R}^n$, in fact differentiable for all $x \neq 0$. For $x = 0$, the set $\partial f(0)$ consists of all $u \in \mathbb{R}^n$ such that

$$\|z\|_2 \geq \langle z, u\rangle \quad \text{for all } z \in \mathbb{R}^n,$$

namely (by Cauchy–Schwarz), the Euclidean unit ball $\{u \in \mathbb{R}^n \mid \|u\|_2 \leq 1\}$. See Figure 51.14.

Example 51.6. For the ℓ^∞ norm if $f(x) = \|x\|_\infty$, we leave it as an exercise to show that $\partial f(0)$ is the polyhedron

$$\partial f(0) = \text{conv}\{\pm e_1, \ldots, \pm e_n\}.$$

See Figure 51.15. One can also work out what is $\partial f(x)$ if $x \neq 0$, but this is more complicated; see Rockafellar [138], page 215.

Example 51.7. The following function is an example of a proper convex function which is not subdifferentiable everywhere:

$$f(x) = \begin{cases} -(1 - |x|^2)^{1/2} & \text{if } |x| \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

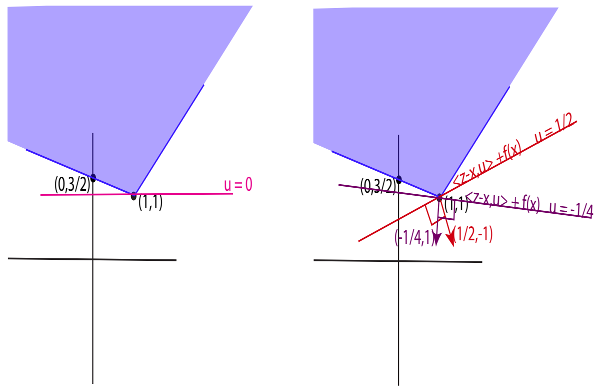


Figure 51.13: Let $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ be the piecewise function defined by $f(x) = x+1$ for $x \geq 1$ and $f(x) = -\frac{1}{2}x + \frac{3}{2}$ for $x < 1$. Its epigraph is the shaded blue region in $\mathbb{R}^2$. Since f has minimum at $x = 1$, $0 \in \partial f(1)$, and the graph of $f(x)$ has a horizontal supporting hyperplane at $(1, 1)$. Since $\{\frac{1}{2}, -\frac{1}{4}\} \subset \partial f(1)$, the maroon line $\frac{1}{2}(x-1)+1$ (with normal $(\frac{1}{2}, -1)$) and the violet line $-\frac{1}{4}(x-1)+1$ (with normal $(-\frac{1}{4}, -1)$) are supporting hyperplanes to the graph of $f(x)$ at $(1, 1)$.

See Figure 51.16. We leave it as an exercise to show that f is subdifferentiable (in fact differentiable) at x when $|x| < 1$, but $\partial f(x) = \emptyset$ when $|x| \geq 1$, even though $x \in \mathrm{dom}(f)$ for $|x| = 1$.

Example 51.8. The subdifferential of an indicator function is interesting. Let C be a nonempty convex set. By definition, $u \in \partial I_C(x)$ iff

$$I_C(z) \geq I_C(x) + \langle z - x, u\rangle \quad \text{for all } z \in \mathbb{R}^n.$$

Since C is nonempty, there is some $z \in C$ such that $I_C(z) = 0$, so the above condition implies that $x \in C$ (otherwise $I_C(x) = +\infty$ but $0 \geq +\infty + \langle z - u, u\rangle$ is impossible), so $0 \geq \langle z - x, u\rangle$ for all $z \in C$, which means that z is normal to C at x. Therefore, *$\partial I_C(x)$ is the normal cone $N_C(x)$ to C at x.*

Example 51.9. The subdifferentials of the indicator function f of the nonnegative orthant of $\mathbb{R}^n$ reveal a connection to complementary slackness conditions. Recall that this indicator function is given by

$$f(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if } x_i \geq 0,\ 1 \leq i \leq n, \\ +\infty & \text{otherwise.} \end{cases}$$

By Example 51.8, the subgradients y of f at $x \geq 0$ form the normal cone to the nonnegative orthant at x. This means that $y \in N_C(x)$ iff

$$\langle z - x, y\rangle \leq 0 \quad \text{for all } z \geq 0$$

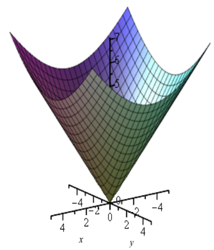


Figure 1

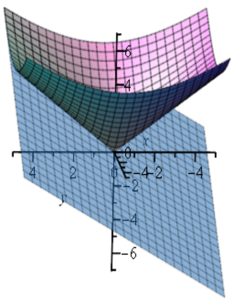


Figure 2

Figure 51.14: Figure (1) shows the graph in $\mathbb{R}^3$ of $f(x, y) = \|(x, y)\|_2 = \sqrt{x^2 + y^2}$. Figure (2) shows the supporting hyperplane with normal $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, -1)$, where $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \in \partial f(0)$.

iff

$$\langle z, y\rangle \leq \langle x, y\rangle \quad \text{for all } z \geq 0.$$

In particular, for $z = 0$ we get $\langle x, y\rangle \geq 0$, and for $z = 2x \geq 0$, we have $\langle x, y\rangle \leq 0$, so $\langle x, y\rangle = 0$. As a consequence, $y \in N_C(x)$ iff $\langle x, y\rangle = 0$ and

$$\langle z, y\rangle \leq 0 \quad \text{for all } z \geq 0.$$

For $z = e_j \geq 0$, we get $y_j \leq 0$. Conversely, if $y \leq 0$ and $\langle x, y\rangle = 0$, since $x \geq 0$, we get $\langle z, y\rangle \leq 0$ for all $z \geq 0$, and so

$$\partial f(x) = \{y = (y_1, \ldots, y_n) \in \mathbb{R}^n \mid y \leq 0,\ \langle x, y\rangle = 0\}.$$

But for $x \geq 0$ and $y \leq 0$ we have $\langle x, y\rangle = \sum_{j=1}^n x_j y_j = 0$ iff $x_j y_j = 0$ for $j = 1, \ldots, n$, thus we see that $y \in \partial f(x)$ iff we have

$$x_j \geq 0,\ y_j \leq 0,\ x_j y_j = 0, \quad 1 \leq j \leq n,$$

which are complementary slackness conditions.

Supporting hyperplanes to the epigraph of a proper convex function f can be used to prove a property which plays a key role in optimization theory. The proof uses a classical result of convex geometry, namely the Minkowski supporting hyperplane theorem.

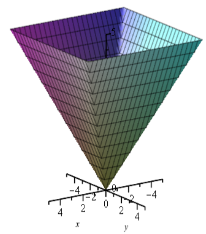


Figure 1

Figure 2

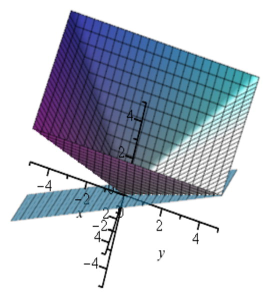


Figure 51.15: Figure (1) shows the graph in $\mathbb{R}^3$ of $f(x, y) = \|(x, y)\|_\infty = \sup\{|x|, |y|\}$. Figure (2) shows the supporting hyperplane with normal $(\frac{1}{2}, \frac{1}{2}, -1)$, where $(\frac{1}{2}, \frac{1}{2}) \in \partial f(0)$.

Theorem 51.11. *(Minkowski) Let C be a nonempty convex set in $\mathbb{R}^n$. For any point $a \in C - \mathbf{relint}(C)$, there is a supporting hyperplane H to C at a.*

Theorem 51.11 is proven in Rockafellar [138] (Theorem 11.6). See also Berger [11] (Proposition 11.5.2). The proof is not as simple as one might expect, and is based on a geometric version of the Hahn–Banach theorem.

In order to prove Theorem 51.14 below we need two technical propositions.

Proposition 51.12. *Let C be any nonempty convex set in $\mathbb{R}^n$. For any $x \in \mathbf{relint}(C)$ and any $y \in \overline{C}$, we have $(1-\lambda)x + \lambda y \in \mathbf{relint}(C)$ for all λ such that $0 \leq \lambda < 1$. In other words, the line segment from x to y including x and excluding y lies entirely within $\mathbf{relint}(C)$.*

Proposition 51.12 is proven in Rockafellar [138] (Theorem 6.1). The proof is not difficult but quite technical.

Proposition 51.13. *For any proper convex function f on $\mathbb{R}^n$, we have*

$$\mathbf{relint}(\mathbf{epi}(f)) = \{(x, \mu) \in \mathbb{R}^{n+1} \mid x \in \mathbf{relint}(\mathrm{dom}(f)),\ f(x) < \mu\}.$$

Proof. Proposition 51.13 is proven in Rockafellar [138] (Lemma 7.3). By working in the affine hull of $\mathbf{epi}(f)$, the statement of Proposition 51.13 is equivalent to

$$\mathrm{int}(\mathbf{epi}(f)) = \{(x, \mu) \in \mathbb{R}^{m+1} \mid x \in \mathrm{int}(\mathrm{dom}(f)),\ f(x) < \mu\},$$

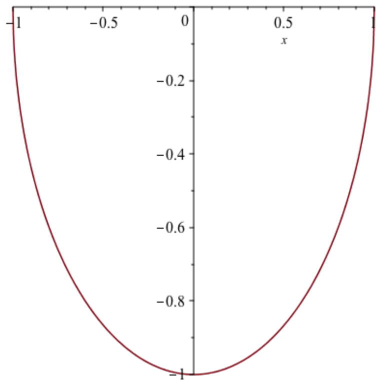


Figure 51.16: The graph of the function in Example 51.7.

assuming that the affine hull of $\mathbf{epi}(f)$ has dimension $m+1$. See Figure (1) of Figure 51.17. The inclusion $\subseteq$ is obvious, so we only need to prove the reverse inclusion. Then for any $z \in \mathrm{int}(\mathrm{dom}(f))$, we can find a convex polyhedral subset $P = \mathrm{conv}(a_1, \ldots, a_{m+1})$ with $a_1, \ldots, a_{m+1} \in \mathrm{dom}(f)$ such that $z \in \mathrm{int}(P)$. Let

$$\alpha = \max\{f(a_1), \ldots, f(a_{m+1})\}.$$

Since any $x \in P$ can be expressed as

$$x = \lambda_1 a_1 + \cdots + \lambda_{m+1} a_{m+1}, \quad \lambda_1 + \cdots + \lambda_{m+1} = 1,\ \lambda_i \geq 0,$$

and since f is convex we have

$$f(x) \leq \lambda_1 f(a_1) + \cdots + \lambda_{m+1} f(a_{m+1}) \leq (\lambda_1 + \cdots + \lambda_{m+1})\alpha = \alpha \quad \text{for all } x \in P.$$

The above shows that the open subset

$$\{(x, \mu) \in \mathbb{R}^{m+1} \mid x \in \mathrm{int}(P),\ \alpha < \mu\}$$

is contained in $\mathbf{epi}(f)$. See Figure (2) of Figure 51.17. In particular, for every $\mu > \alpha$, we have

$$(z, \mu) \in \mathrm{int}(\mathbf{epi}(f)).$$

Thus for any $\beta \in \mathbb{R}$ such that $\beta > f(z)$, we see that (z, β) belongs to the relative interior of the vertical line segment $\{(z, \mu) \in \mathbb{R}^{m+1} \mid f(z) \leq \mu \leq \alpha + \beta + 1\}$ which meets $\mathrm{int}(\mathbf{epi}(f))$. See Figure (3) of Figure 51.17. By Proposition 51.12, $(z, \beta) \in \mathrm{int}(\mathbf{epi}(f))$. ☐

We can now prove the following important theorem.

Theorem 51.14. *Let f be a proper convex function on $\mathbb{R}^n$. For any $x \in \mathbf{relint}(\mathrm{dom}(f))$, there is a nonvertical supporting hyperplane $\mathcal{H}$ to $\mathbf{epi}(f)$ at $(x, f(x))$. Consequently f is subdiffentiable for all $x \in \mathbf{relint}(\mathrm{dom}(f))$, and there is some affine form $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ such that $f(y) \geq \varphi(y)$ for all $y \in \mathbb{R}^n$.*

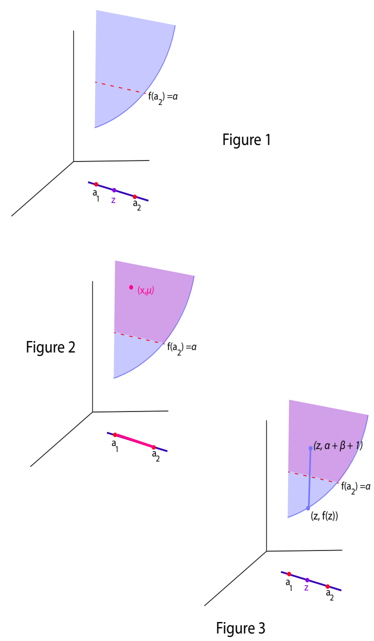


Figure 51.17: Figure (1) illustrates $\mathbf{epi}(f)$, where $\mathbf{epi}(f)$ is contained in a vertical plane of affine dimension 2. Figure (2) illustrates the magenta open subset $\{(x, \mu) \in \mathbb{R}^2 \mid x \in \mathrm{int}(P),\ \alpha < \mu\}$ of $\mathbf{epi}(f)$. Figure (3) illustrates the vertical line segment $\{(z, \mu) \in \mathbb{R}^2 \mid f(z) \leq \mu \leq \alpha + \beta + 1\}$.

Proof. By Proposition 51.14, for any $x \in \mathbf{relint}(\mathrm{dom}(f))$, we have $(x, \mu) \in \mathbf{relint}(\mathbf{epi}(f))$ for all $\mu \in \mathbb{R}$ such that $f(x) < \mu$. Since by definition of $\mathbf{epi}(f)$ we have $(x, f(x)) \in \mathbf{epi}(f) - \mathbf{relint}(\mathbf{epi}(f))$, by Minkowski's theorem (Theorem 51.11), there is a supporting hyperplane $\mathcal{H}$ to $\mathbf{epi}(f)$ through $(x, f(x))$. Since $x \in \mathbf{relint}(\mathrm{dom}(f))$ and f is proper, the hyperplane $\mathcal{H}$ is not a vertical hyperplane. By Definition 51.14, the function f is subdifferentiable at any $x \in \mathbf{relint}(\mathrm{dom}(f))$, and the subgradient inequality shows that if we pick some $x \in \mathbf{relint}(\mathrm{dom}(f))$ and if we let $\varphi(z) = f(x) + \langle z - x, u\rangle$, then φ is an affine form such that $f(z) \geq \varphi(z)$ for all $z \in \mathbb{R}^n$. $\square$

Intuitively, a proper convex function can't decrease faster than an affine function. It is surprising how much work it takes to prove such an "obvious" fact.

Remark: Consider the proper convex function $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ given by

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0 \\ +\infty & \text{if } x < 0. \end{cases}$$

We have $\mathrm{dom}(f) = [0, +\infty)$, f is differentiable for all $x > 0$, but it is not subdifferentiable at $x = 0$. The only supporting hyperplane to $\mathbf{epi}(f)$ at $(0, 0)$ is the vertical line of equation $x = 0$ (the y-axis) as illustrated by Figure 51.18.

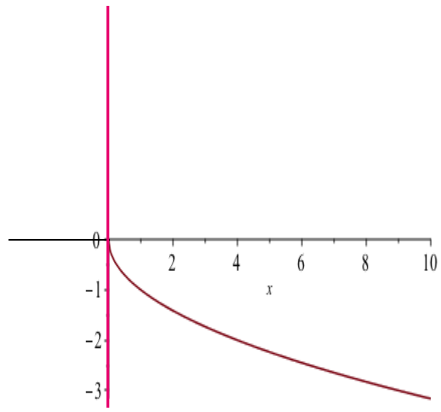


Figure 51.18: The graph of the partial function $f(x) = -\sqrt{x}$ and its red vertical supporting hyperplane at $x = 0$.

51.3 Basic Properties of Subgradients and Subdifferentials

A major tool to prove properties of subgradients is a variant of the notion of directional derivative.

Definition 51.15. Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be any function. For any $x \in \mathbb{R}^n$ such that $f(x)$ is finite ($f(x) \in \mathbb{R}$), for any $u \in \mathbb{R}^n$, the *one-sided directional derivative* $f'(x; u)$ is defined to be the limit

$$f'(x; u) = \lim_{\lambda \downarrow 0} \frac{f(x + \lambda u) - f(x)}{\lambda}$$

if it exists ($-\infty$ and $+\infty$ being allowed as limits). See Figure 51.19. The above notation for the limit means that we consider the limit when $\lambda > 0$ tends to 0.

Note that

$$\lim_{\lambda \uparrow 0} \frac{f(x + \lambda u) - f(x)}{\lambda}$$

denotes the one-sided limit when $\lambda < 0$ tends to zero, and that

$$-f'(x; -u) = \lim_{\lambda \uparrow 0} \frac{f(x + \lambda u) - f(x)}{\lambda},$$

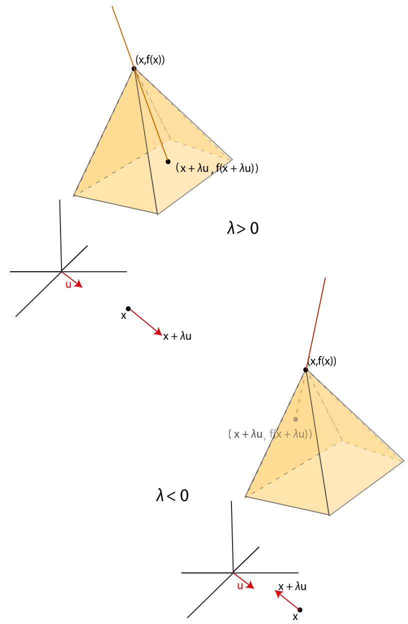


Figure 51.19: Let $f\colon \mathbb{R}^2 \to \mathbb{R}\cup\{-\infty,+\infty\}$ be the function whose graph (in $\mathbb{R}^3$) is the surface of the peach pyramid. The top figure illustrates that $f'(x;u)$ is the slope of the slanted burnt orange line, while the bottom figure depicts the line associated with $\lim_{\lambda\uparrow 0}\frac{f(x+\lambda u)-f(x)}{\lambda}$.

so *the (two-sided) directional derivative* $\mathrm{D}_u f(x)$ *exists iff* $-f'(x;-u) = f'(x;u)$. Also, if f is differentiable at x, then

$$f'(x;u) = \langle \nabla f_x, u\rangle, \quad \text{for all } u \in \mathbb{R}^n,$$

where ∇f_x is the gradient of f at x. Here is the first remarkable result.

Proposition 51.15. *Let $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{-\infty,+\infty\}$ be a convex function. For any $x\in\mathbb{R}^n$, if $f(x)$ is finite, then the function*

$$\lambda \mapsto \frac{f(x+\lambda u) - f(x)}{\lambda}$$

is a nondecreasing function of $\lambda > 0$, so that $f'(x;u)$ exists for any $u\in\mathbb{R}^n$, and

$$f'(x;u) = \inf_{\lambda>0}\frac{f(x+\lambda u) - f(x)}{\lambda}.$$

Furthermore, $f'(x;u)$ is a positively homogeneous convex function of u (which means that $f'(x;\alpha u) = \alpha f'(x;u)$ for all $\alpha\in\mathbb{R}$ with $\alpha > 0$ and all $u\in\mathbb{R}^n$), $f'(x;0) = 0$, and

$$-f'(x;-u) \leq f'(x;u) \quad \textit{for all } u \in \mathbb{R}^n$$

Proposition 51.15 is proven in Rockafellar [138] (Theorem 23.1). The proof is not difficult but not very informative.

Remark: As a convex function of u, it can be shown that the effective domain of the function $u \mapsto f'(x; u)$ is the convex cone generated by $\mathrm{dom}(f) - x$.

We will now state without proof some of the most important properties of subgradients and subdifferentials. Complete details can be found in Rockafellar [138] (Part V, Section 23).

In order to state the next proposition, we need the following definition.

Definition 51.16. For any convex set C in $\mathbb{R}^n$, the *support function* $\delta^*(-|C)$ of C is defined by

$$\delta^*(x|C) = \sup_{y \in C} \langle x, y \rangle, \quad x \in \mathbb{R}^n.$$

According to Definition 50.11, the conjugate of the indicator function I_C of a convex set C is given by

$$I_C^*(x) = \sup_{y \in \mathbb{R}^n} (\langle x, y \rangle - I_C(y)) = \sup_{y \in C} \langle x, y \rangle = \delta^*(x|C).$$

Thus $\delta^*(-|C) = I_C^*$, *the conjugate of the indicator function* I_C.

The following proposition relates directional derivatives at x and the subdifferential at x.

Proposition 51.16. *Let $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be a convex function. For any $x \in \mathbb{R}^n$, if $f(x)$ is finite, then a vector $u \in \mathbb{R}^n$ is a subgradient to f at x if and only if*

$$f'(x; y) \geq \langle y, u \rangle \quad \text{for all } y \in \mathbb{R}^n.$$

Furthermore, the closure of the convex function $y \mapsto f'(x; y)$ is the support function of the closed convex set $\partial f(x)$, the subdifferential of f at x:

$$\mathrm{cl}(f'(x; -)) = \delta^*(-|\partial f(x)).$$

Sketch of proof. Proposition 51.16 is proven in Rockafellar [138] (Theorem 23.2). We prove the inequality. If we write $z = x + \lambda y$ with $\lambda > 0$, then the subgradient inequality implies

$$f(x + \lambda u) \geq f(x) + \langle z - x, u \rangle = f(x) + \lambda \langle y, u \rangle,$$

so we get

$$\frac{f(x + \lambda y) - f(x)}{\lambda} \geq \langle y, u \rangle.$$

Since the expression on the left tends to $f'(x; y)$ as $\lambda > 0$ tends to zero, we obtain the desired inequality. The second part follows from Corollary 13.2.1 in Rockafellar [138]. $\square$

If f is a proper function on $\mathbb{R}$, then its effective domain being convex is an interval whose relative interior is an open interval (a,b). In Proposition 51.16, we can pick $y = 1$ so $\langle y, u\rangle = u$, and for any $x \in (a,b)$, since the limits $f'_-(x) = -f'(x; -1)$ and $f'_+(x) = f'(x; 1)$ exist, with $f'_-(x) \leq f'_+(x)$, we deduce that $\partial f(x) = [f'_-(x), f'_+(x)]$. The numbers $\alpha \in [f'_-(x), f'_+(x)]$ are the slopes of nonvertical lines in $\mathbb{R}^2$ passing through $(x, f(x))$ that are supporting lines to the epigraph $\mathbf{epi}(f)$ of f.

Example 51.10. If f is the celebrated **ReLU** function (ramp function) from deep learning defined so that

$$\mathrm{ReLU}(x) = \max\{x, 0\} = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0, \end{cases}$$

then $\partial\,\mathrm{ReLU}(0) = [0,1]$. See Figure 51.20. The function ReLU is differentiable for $x \neq 0$, with $\mathrm{ReLU}'(x) = 0$ if $x < 0$ and $\mathrm{ReLU}'(x) = 1$ if $x > 0$.

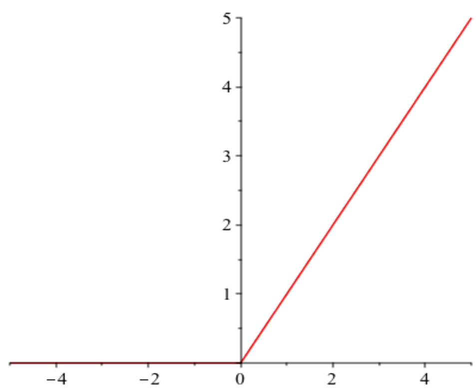


Figure 51.20: The graph of the ReLU function.

Proposition 51.16 has several interesting consequences.

Proposition 51.17. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ be a convex function. For any $x \in \mathbb{R}^n$, if $f(x)$ is finite and if f is subdifferentiable at x, then f is proper. If f is not subdifferentiable at x, then there is some $y \neq 0$ such that*

$$f'(x; y) = -f'(x; -y) = -\infty.$$

Proposition 51.17 is proven in Rockafellar [138] (Theorem 23.3). It confirms that improper convex functions are rather pathological objects, because if a convex function is subdifferentiable for some x such that $f(x)$ is finite, then f must be proper. This is because if $f(x)$ is finite, then the subgradient inequality implies that f majorizes an affine function, which is proper.

The next theorem is one of the most important results about the connection between one-sided directional derivatives and subdifferentials. It sharpens the result of Theorem 51.14.

Theorem 51.18. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. For any $x \notin \mathrm{dom}(f)$, we have $\partial f(x) = \emptyset$. For any $x \in \mathbf{relint}(\mathrm{dom}(f))$, we have $\partial f(x) \neq \emptyset$, the map $y \mapsto f'(x; y)$ is convex, closed and proper, and*

$$f'(x; y) = \sup_{u \in \partial f(x)} \langle y, u\rangle = \delta^*(y|\partial f(x)) \quad \text{for all } y \in \mathbb{R}^n.$$

The subdifferential $\partial f(x)$ is nonempty and bounded (also closed and convex) if and only if $x \in \mathrm{int}(\mathrm{dom}(f))$, in which case $f'(x; y)$ is finite for all $y \in \mathbb{R}^n$.

Theorem 51.18 is proven in Rockafellar [138] (Theorem 23.4). If we write

$$\mathrm{dom}(\partial f) = \{x \in \mathbb{R}^n \mid \partial f(x) \neq \emptyset\},$$

then Theorem 51.18 implies that

$$\mathbf{relint}(\mathrm{dom}(f)) \subseteq \mathrm{dom}(\partial f) \subseteq \mathrm{dom}(f).$$

However, $\mathrm{dom}(\partial f)$ is not necessarily convex as shown by the following counterexample.

Example 51.11. Consider the proper convex function defined on $\mathbb{R}^2$ given by

$$f(x, y) = \max\{g(x), |y|\},$$

where

$$g(x) = \begin{cases} 1 - \sqrt{x} & \text{if } x \geq 0 \\ +\infty & \text{if } x < 0. \end{cases}$$

See Figure 51.21. It is easy to see that $\mathrm{dom}(f) = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0\}$, but $\mathrm{dom}(\partial f) = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0\} - \{(0, y) \mid -1 < y < 1\}$, which is not convex.

The following theorem is important because it tells us when a convex function is differentiable in terms of its subdifferential, as shown in Rockafellar [138] (Theorem 25.1).

Theorem 51.19. *Let f be a convex function on $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$ such that $f(x)$ is finite. If f is differentiable at x then $\partial f(x) = \{\nabla f_x\}$ (where ∇f_x is the gradient of f at x) and we have*

$$f(z) \geq f(x) + \langle z - x, \nabla f_x\rangle \quad \text{for all } z \in \mathbb{R}^n.$$

Conversely, if $\partial f(x)$ consists of a single vector, then $\partial f(x) = \{\nabla f_x\}$ and f is differentiable at x.

The first direction is easy to prove. Indeed, if f is differentiable at x, then

$$f'(x; y) = \langle y, \nabla f_x\rangle \quad \text{for all } y \in \mathbb{R}^n,$$

so by Proposition 51.16, a vector u is a subgradient at x iff

$$\langle y, \nabla f_x\rangle \geq \langle y, u\rangle \quad \text{for all } y \in \mathbb{R}^n,$$

so $\langle y, \nabla f_x - u\rangle \geq 0$ for all y, which implies that $u = \nabla f_x$.

We obtain the following corollary.

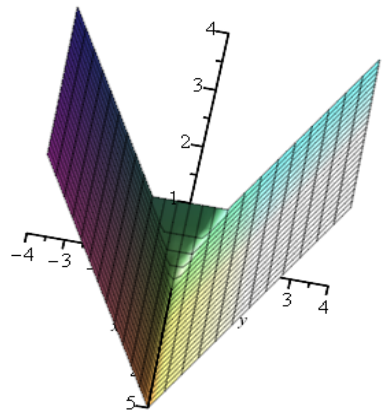


Figure 51.21: The graph of the function from Example 51.11 with a view along the positive x axis.

Corollary 51.20. *Let f be a convex function on $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$ such that $f(x)$ is finite. If f is differentiable at x, then f is proper and $x \in \mathrm{int}(\mathrm{dom}(f))$.*

The following theorem shows that proper convex functions are differentiable almost everywhere.

Theorem 51.21. *Let f be a proper convex function on $\mathbb{R}^n$, and let D be the set of vectors where f is differentiable. Then D is a dense subset of $\mathrm{int}(\mathrm{dom}(f))$, and its complement in $\mathrm{int}(\mathrm{dom}(f))$ has measure zero. Furthermore, the gradient map $x \mapsto \nabla f_x$ is continuous on D.*

Theorem 51.21 is proven in Rockafellar [138] (Theorem 25.5).

Remark: If $f \colon (a, b) \to \mathbb{R}$ is a finite convex function on an open interval of $\mathbb{R}$, then the set D where f is differentiable is dense in (a, b), and $(a, b) - D$ is at most countable. The map f' is continuous and nondecreasing on D. See Rockafellar [138] (Theorem 25.3).

We also have the following result showing that in "most cases" the subdifferential $\partial f(x)$ can be constructed from the gradient map; see Rockafellar [138] (Theorem 25.6).

Theorem 51.22. *Let f be a closed proper convex function on $\mathbb{R}^n$. If $\mathrm{int}(\mathrm{dom}(f)) \neq \emptyset$, then for every $x \in \mathrm{dom}(f)$, we have*

$$\partial f(x) = \overline{\mathrm{conv}(S(x))} + N_{\mathrm{dom}(f)}(x)$$

where $N_{\mathrm{dom}(f)}(x)$ is the normal cone to $\mathrm{dom}(f)$ at x, and $S(x)$ is the set of all limits of sequences of the form $\nabla f_{x_1}, \nabla f_{x_2}, \ldots, \nabla f_{x_p}, \ldots$, where $x_1, x_2, \ldots, x_p, \ldots$ is a sequence in $\mathrm{dom}(f)$ converging to x such that each ∇f_{x_p} is defined.

The next two results generalize familiar results about derivatives to subdifferentials.

Proposition 51.23. *Let $f_1, \ldots, f_n$ be proper convex functions on $\mathbb{R}^n$, and let $f = f_1 + \cdots + f_n$. For $x \in \mathbb{R}^n$, we have*

$$\partial f(x) \supseteq \partial f_1(x) + \cdots + \partial f_n(x).$$

If $\bigcap_{i=1}^n \mathbf{relint}(\mathrm{dom}(f_i)) \neq \emptyset$, then

$$\partial f(x) = \partial f_1(x) + \cdots + \partial f_n(x).$$

Proposition 51.23 is proven in Rockafellar [138] (Theorem 23.8).

The next result can be viewed as a generalization of the chain rule.

Proposition 51.24. *Let f be the function given by $f(x) = h(Ax)$ for all $x \in \mathbb{R}^n$, where h is a proper convex function on $\mathbb{R}^m$ and A is an $m \times n$ matrix. Then*

$$\partial f(x) \supseteq A^\top(\partial h(Ax)) \quad \textit{for all } x \in \mathbb{R}^n.$$

If the range of A contains a point of $\mathbf{relint}(\mathrm{dom}(h))$, then

$$\partial f(x) = A^\top(\partial h(Ax)).$$

Proposition 51.24 is proven in Rockafellar [138] (Theorem 23.9).

51.4 Additional Properties of Subdifferentials

In general, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is a function (not necessarily convex) and f is differentiable at x, we expect that the gradient ∇f_x of f at x is normal to the level set $\{z \in \mathbb{R}^n \mid f(z) = f(x)\}$ at $f(x)$. An analogous result, as illustrated in Figure 51.22, holds for proper convex functions in terms of subdifferentials.

Proposition 51.25. *Let f be a proper convex function on $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$ be a vector such that f is subdifferentiable at x but f does not achieve its minimum at x. Then the normal cone $N_C(x)$ at x to the sublevel set $C = \{z \in \mathbb{R}^n \mid f(z) \leq f(x)\}$ is the closure of the convex cone spanned by $\partial f(x)$.*

Proposition 51.25 is proven in Rockafellar [138] (Theorem 23.7).

The following result sharpens Proposition 51.8.

Proposition 51.26. *Let f be a closed proper convex function on $\mathbb{R}^n$, and let S be a nonempty closed and bounded subset of $\mathrm{int}(\mathrm{dom}(f))$. Then*

$$\partial f(S) = \bigcup_{x \in S} \partial f(x)$$

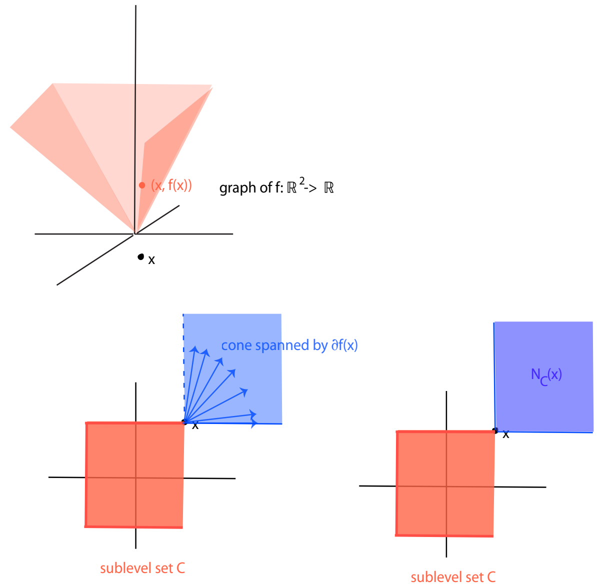


Figure 51.22: Let f be the proper convex function whose graph in $\mathbb{R}^3$ is the peach polyhedral surface. The sublevel set $C = \{z \in \mathbb{R}^2 \mid f(z) \leq f(x)\}$ is the orange square which is closed on three sides. Then the normal cone $N_C(x)$ is the closure of the convex cone spanned by $\partial f(x)$.

is nonempty, closed and bounded. If

$$\alpha = \sup_{y \in \partial f(S)} \|y\|_2 < +\infty,$$

then f is Lipschitizan on S, and we have

$$\begin{aligned} f'(x; z) &\leq \alpha \|z\|_2 && \text{for all } x \in S \text{ and all } z \in \mathbb{R}^n \\ |f(y) - f(x)| &\leq \alpha \|y - z\|_2 && \text{for all } x, y \in S. \end{aligned}$$

Proposition 51.24 is proven in Rockafellar [138] (Theorem 24.7).

The subdifferentials of a proper convex function f and its conjugate f^* are closely related. First, we have the following proposition from Rockafellar [138] (Theorem 12.2).

Proposition 51.27. *Let f be convex function on $\mathbb{R}^n$. The conjugate function f^* of f is a closed and convex function, proper iff f is proper. Furthermore, $(\mathrm{cl}(f))^* = f^*$, and $f^{**} = \mathrm{cl}(f)$.*

As a corollary of Proposition 51.27, it can be shown that

$$f^*(y) = \sup_{x \in \mathbf{relint}(\mathrm{dom}(f))} (\langle x, y \rangle - f(x)).$$

The following result is proven in Rockafellar [138] (Theorem 23.5).

Proposition 51.28. *For any proper convex function f on $\mathbb{R}^n$ and for any vector $x \in \mathbb{R}^n$, the following conditions on a vector $y \in \mathbb{R}^n$ are equivalent.*

(a) $y \in \partial f(x)$.

(b) *The function* $\langle z, y\rangle - f(z)$ *achieves its supremum in* z *at* $z = x$.

(c) $f(x) + f^*(y) \leq \langle x, y\rangle$.

(d) $f(x) + f^*(y) = \langle x, y\rangle$.

If $(\mathrm{cl}(f))(x) = f(x)$, *then there are three more conditions all equivalent to the above conditions.*

(a^*) $x \in \partial f^*(y)$.

(b^*) *The function* $\langle x, z\rangle - f^*(z)$ *achieves its supremum in* z *at* $z = y$.

(a^{**}) $y \in \partial(\mathrm{cl}(f))(x)$.

The following results are corollaries of Proposition 51.28; see Rockafellar [138] (Corollaries 23.5.1, 23.5.2, 23.5.3).

Corollary 51.29. *For any proper convex function f on $\mathbb{R}^n$, if f is closed, then $y \in \partial f(x)$ iff $x \in \partial f^*(y)$, for all $x, y \in \mathbb{R}^n$.*

Corollary 51.29 states a sort of adjunction property.

Corollary 51.30. *For any proper convex function f on $\mathbb{R}^n$, if f is subdifferentiable at $x \in \mathbb{R}^n$, then $(\mathrm{cl}(f))(x) = f(x)$ and $\partial(\mathrm{cl}(f))(x) = \partial f(x)$.*

Corollary 51.30 shows that the closure of a proper convex function f agrees with f whereever f is subdifferentiable.

Corollary 51.31. *For any proper convex function f on $\mathbb{R}^n$, for any nonempty closed convex subset C of $\mathbb{R}^n$, for any $y \in \mathbb{R}^n$, the set $\partial\delta^*(y|C) = \partial I_C^*(y)$ consists of the vectors $x \in \mathbb{R}^n$ (if any) where the linear form $z \mapsto \langle z, y\rangle$ achieves its maximum over C.*

There is a notion of approximate subgradient which turns out to be useful in optimization theory; see Bertsekas [19, 17].

Definition 51.17. Let $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any proper convex function. For any $\epsilon > 0$, for any $x \in \mathbb{R}^n$, if $f(x)$ is finite, then an *ϵ-subgradient of f at x* is any vector $u \in \mathbb{R}^n$ such that

$$f(z) \geq f(x) - \epsilon + \langle z - x, u\rangle, \quad \text{for all } z \in \mathbb{R}^n.$$

See Figure 51.23. The set of all ϵ-subgradients of f at x is denoted $\partial_\epsilon f(x)$ and is called the *ϵ-subdifferential* of f at x.

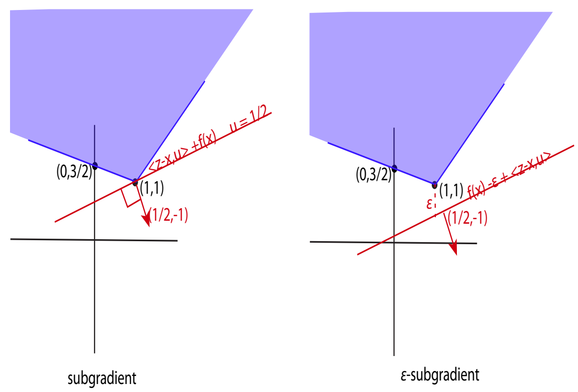


Figure 51.23: Let $f\colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$ be the piecewise function defined by $f(x) = x + 1$ for $x \geq 1$ and $f(x) = -\frac{1}{2}x + \frac{3}{2}$ for $x < 1$. Its epigraph is the shaded blue region in $\mathbb{R}^2$. The line $\frac{1}{2}(x-1)+1$ (with normal $(\frac{1}{2}, -1)$ is a supporting hyperplane to the graph of $f(x)$ at $(1,1)$ while the line $\frac{1}{2}(x-1)+1-\epsilon$ is the hyperplane associated with the ϵ-subgradient at $x=1$ and shows that $u = \frac{1}{2} \in \partial_\epsilon f(x)$.

The set $\partial_\epsilon f(x)$ can be defined in terms of the conjugate of the function h_x given by

$$h_x(y) = f(x+y) - f(x), \quad \text{for all } y \in \mathbb{R}^n.$$

Proposition 51.32. *Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any proper convex function. For any $\epsilon > 0$, if h_x is given by*

$$h_x(y) = f(x+y) - f(x), \quad \textit{for all } y \in \mathbb{R}^n,$$

then

$$h_x^*(y) = f^*(y) + f(x) - \langle x, y\rangle \quad \textit{for all } y \in \mathbb{R}^n$$

and

$$\partial_\epsilon f(x) = \{u \in \mathbb{R}^n \mid h_x^*(u) \leq \epsilon\}.$$

Proof. We have

$$\begin{aligned}
h_x^*(y) &= \sup_{z \in \mathbb{R}^n} (\langle y, z\rangle - h_x(z)) \\
&= \sup_{z \in \mathbb{R}^n} (\langle y, z\rangle - f(x+z) + f(x)) \\
&= \sup_{x+z \in \mathbb{R}^n} (\langle y, x+z\rangle - f(x+z) - \langle y, x\rangle + f(x)) \\
&= f^*(y) + f(x) - \langle x, y\rangle.
\end{aligned}$$

Observe that $u \in \partial_\epsilon f(x)$ iff for every $y \in \mathbb{R}^n$,

$$f(x+y) \geq f(x) - \epsilon + \langle y, u\rangle$$

iff

$$\epsilon \geq \langle y, u\rangle - f(x+y) + f(x) = \langle y, u\rangle - h_x(y).$$

Since by definition

$$h_x^*(u) = \sup_{y\in\mathbb{R}^n} (\langle y, u\rangle - h_x(y)),$$

we conclude that

$$\partial_\epsilon f(x) = \{u \in \mathbb{R}^n \mid h_x^*(u) \leq \epsilon\},$$

as claimed. $\square$

Remark: By Fenchel's inequality $h_x^*(y) \geq 0$, and by Proposition 51.28(d), the set of vectors where h_x^* vanishes is $\partial f(x)$.

The equation $\partial_\epsilon f(x) = \{u \in \mathbb{R}^n \mid h_x^*(u) \leq \epsilon\}$ shows that $\partial_\epsilon f(x)$ is a closed convex set. As ϵ gets smaller, the set $\partial_\epsilon f(x)$ decreases, and we have

$$\partial f(x) = \bigcap_{\epsilon > 0} \partial_\epsilon f(x).$$

However $\delta^*(y|\partial_\epsilon f(x)) = I^*_{\partial_\epsilon f(x)}(y)$ does not necessarily decrease to $\delta^*(y|\partial f(x)) = I^*_{\partial f(x)}(y)$ as ϵ decreases to zero. The discrepancy corresponds to the discrepancy between $f'(x; y)$ and $\delta^*(y|\partial f(x)) = I^*_{\partial f(x)}(y)$ and is due to the fact that f is not necessarily closed (see Proposition 51.16) as shown by the following result proven in Rockafellar [138] (Theorem 23.6).

Proposition 51.33. *Let f be a closed and proper convex function, and let $x \in \mathbb{R}^n$ such that $f(x)$ is finite. Then*

$$f'(x; y) = \lim_{\epsilon \downarrow 0} \delta^*(y|\partial_\epsilon f(x)) = \lim_{\epsilon\downarrow 0} I^*_{\partial_\epsilon f(x)}(y) \quad \textit{for all } y \in \mathbb{R}^n.$$

The theory of convex functions is rich and we have only given a sample of some of the most significant results that are relevant to optimization theory. There are a few more results regarding the minimum of convex functions that are particularly important due to their applications to optimization theory.

51.5 The Minimum of a Proper Convex Function

Let h be a proper convex function on $\mathbb{R}^n$. The general problem is to study the minimum of h over a nonempty convex set C in $\mathbb{R}^n$, possibly defined by a set of inequality and equality constraints. We already observed that minimizing h over C is equivalent to minimizing the proper convex function f given by

$$f(x) = h(x) + I_C(x) = \begin{cases} h(x) & \text{if } x \in C \\ +\infty & \text{if } x \notin C. \end{cases}$$

Therefore it makes sense to begin by considering the problem of minimizing a proper convex function f over $\mathbb{R}^n$. Of course, minimizing over $\mathbb{R}^n$ is equivalent to minimizing over $\mathrm{dom}(f)$.

Definition 51.18. Let f be a proper convex function on $\mathbb{R}^n$. We denote by $\inf f$ the quantity

$$\inf f = \inf_{x \in \mathrm{dom}(f)} f(x).$$

This is the minimum of the function f over $\mathbb{R}^n$ (it may be equal to $-\infty$).

For every $\alpha \in \mathbb{R}$, we have the sublevel set

$$\mathrm{sublev}_\alpha(f) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}.$$

By Proposition 51.2, we know that the sublevel sets $\mathrm{sublev}_\alpha(f)$ are convex and that

$$\mathrm{dom}(f) = \bigcup_{\alpha \in \mathbb{R}} \mathrm{sublev}_\alpha(f).$$

Observe that $\mathrm{sublev}_\alpha(f) = \emptyset$ if $\alpha < \inf f$. If $\inf f > -\infty$, then for $\alpha = \inf f$, the sublevel set $\mathrm{sublev}_\alpha(f)$ consists of the set of vectors where f achieves it minimum.

Definition 51.19. Let f be a proper convex function on $\mathbb{R}^n$. If $\inf f > -\infty$, then the sublevel set $\mathrm{sublev}_{\inf f}(f)$ is called the *minimum set* of f (this set may be empty). See Figure 51.24.

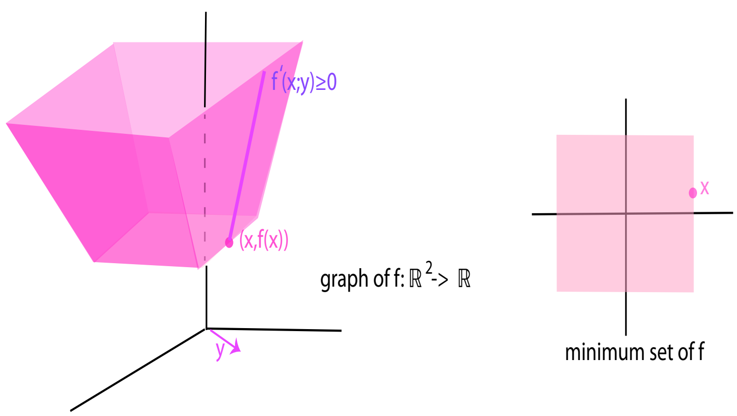


Figure 51.24: Let f be the proper convex function whose graph is the surface of the upward facing pink trough. The minimum set of f is the light pink square of $\mathbb{R}^2$ which maps to the bottom surface of the trough in $\mathbb{R}^3$. For any x in the minimum set, $f'(x; y) \geq 0$, a fact substantiated by Proposition 51.34.

It is important to determine whether the minimum set is empty or nonempty, or whether it contains a single point. As we noted in Theorem 40.13(2), if f is strictly convex then the minimum set contains at most one point.

In any case, we know from Proposition 51.2 and Proposition 51.3 that the minimum set of f is convex, and closed iff f is closed.

Subdifferentials provide the first criterion for deciding whether a vector $x \in \mathbb{R}^n$ belongs to the minimum set of f. Indeed, the very definition of a subgradient says that $x \in \mathbb{R}^n$ belongs to the minimum set of f iff $0 \in \partial f(x)$. Using Proposition 51.16, we obtain the following result.

Proposition 51.34. *Let f be a proper convex function over $\mathbb{R}^n$. A vector $x \in \mathbb{R}^n$ belongs to the minimum set of f iff*

$$0 \in \partial f(x)$$

iff $f(x)$ is finite and

$$f'(x; y) \geq 0 \quad \textit{for all } y \in \mathbb{R}^n.$$

Of course, if f is differentiable at x, then $\partial f(x) = \{\nabla f_x\}$, and we obtain the well-known condition $\nabla f_x = 0$.

There are many ways of expressing the conditions of Proposition 51.34, and the minimum set of f can even be characterized in terms of the conjugate function f^*. The notion of direction of recession plays a key role.

Definition 51.20. Let $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any function. A *direction of recession* of f is any non-zero vector $u \in \mathbb{R}^n$ such that for every $x \in \mathrm{dom}(f)$, the function $\lambda \mapsto f(x + \lambda u)$ is nonincreasing (this means that for all $\lambda_1, \lambda_2 \in \mathbb{R}$, if $\lambda_1 < \lambda_2$, then $x + \lambda_1 u \in \mathrm{dom}(f)$, $x + \lambda_2 u \in \mathrm{dom}(f)$, and $f(x + \lambda_2 u) \leq f(x + \lambda_1 u)$).

Example 51.12. Consider the function $f \colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = 2x + y^2$. Since

$$f(x + \lambda u, y + \lambda v) = 2(x + \lambda u) + (y + \lambda v)^2 = 2x + y^2 + 2(u + yv)\lambda + v^2\lambda^2,$$

if $v \neq 0$, we see that the above quadratic function of λ increases for $\lambda \geq -(u + yv)/v^2$. If $v = 0$, then the function $\lambda \mapsto 2x + y^2 + 2u\lambda$ decreases to $-\infty$ when λ goes to $+\infty$ if $u < 0$, so all vectors $(-u, 0)$ with $u > 0$ are directions of recession. See Figure 51.25.

The function $f(x, y) = 2x + x^2 + y^2$ does not have any direction of recession, because

$$f(x + \lambda u, y + \lambda v) = 2x + x^2 + y^2 + 2(u + ux + yv)\lambda + (u^2 + v^2)\lambda^2,$$

and since $(u, v) \neq (0, 0)$, we have $u^2 + v^2 > 0$, so as a function of λ, the above quadratic function increases for $\lambda \geq -(u + ux + yv)/(u^2 + v^2)$. See Figure 51.25.

In fact, the above example is typical. For any symmetric positive definite $n \times n$ matrix A and any vector $b \in \mathbb{R}^n$, the quadratic strictly convex function q given by $q(x) = x^\top Ax + b^\top x$ has no directions of recession. For any $u \in \mathbb{R}^n$, with $u \neq 0$, we have

$$\begin{aligned} q(x + \lambda u) &= (x + \lambda u)^\top A(x + \lambda u) + b^\top(x + \lambda u) \\ &= x^\top Ax + b^\top x + (2x^\top Au + b^\top u)\lambda + (u^\top Au)\lambda^2. \end{aligned}$$

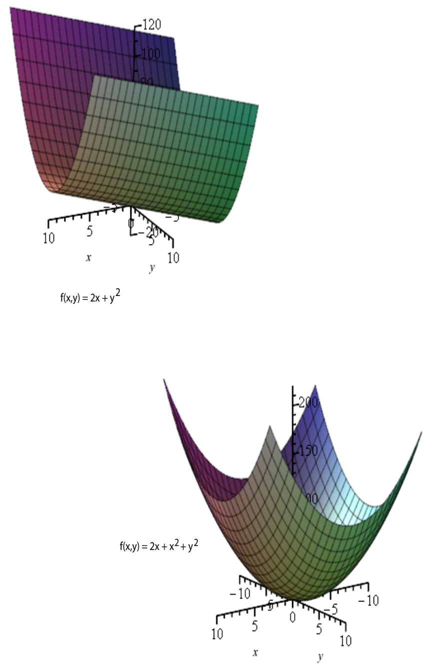


Figure 51.25: The graphs of the two functions discussed in Example 51.12. The graph of $f(x,y) = 2x+y^2$ slopes "downward" along the negative x-axis, reflecting the fact that $(-u,0)$ is a direction of recession.

Since $u \neq 0$ and A is SPD, we have $u^\top Au > 0$, and the above quadratic function increases for $\lambda \geq -(2x^\top Au + b^\top u)/(2u^\top Au)$.

The above fact yields an important trick of convex optimization. If f is any proper closed and convex function, then for any quadratic strictly convex function q, the function $h = f+q$ is a proper and closed strictly convex function that has a minimum which is attained for a *unique* vector. This trick is at the core of the method of augmented Lagrangians, and in particular ADMM. Surprisingly, a rigorous proof requires the deep theorem below.

One should be careful not to conclude hastily that if a convex function is proper and closed, then $\mathrm{dom}(f)$ and $\mathrm{Im}(f)$ are also closed. Also, a closed and proper convex function may not attain its minimum. For example, the function $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ given by

$$f(x) = \begin{cases} \dfrac{1}{x} & \text{if } x > 0 \\ +\infty & \text{if } x \leq 0 \end{cases}$$

is a proper, closed and convex function, but $\mathrm{dom}(f) = (0,+\infty)$ and $\mathrm{Im}(f) = (0,+\infty)$. Note that $\inf f = 0$ is not attained. The problem is that f has 1 has a direction of recession as evidenced by the graph provided in Figure 51.26.

The following theorem is proven in Rockafellar [138] (Theorem 27.1).

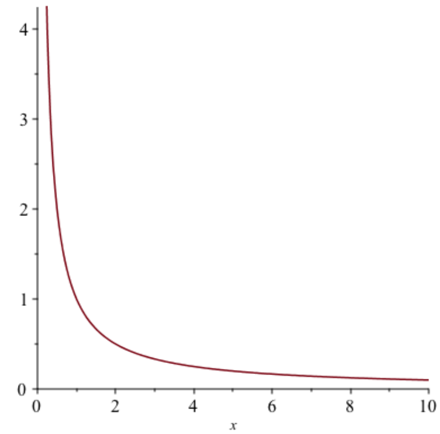


Figure 51.26: The graph of the partial function $f(x) = \frac{1}{x}$ for $x > 0$. The graph of this function decreases along the x-axis since 1 is a direction of recession.

Theorem 51.35. *Let f be a proper and closed convex function over $\mathbb{R}^n$. The following statements hold:*

(1) We have $\inf f = -f^*(0)$*. Thus f is bounded below iff* $0 \in \mathrm{dom}(f^*)$*.*

(2) The minimum set of f is equal to $\partial f^*(0)$. Thus the infimum of f is attained (which means that there is some $x \in \mathbb{R}^n$ such that $f(x) = \inf f$) iff f^* is subdifferentiable at 0. This condition holds in particular when $0 \in \mathbf{relint}(\mathrm{dom}(f^*))$. Moreover, $0 \in \mathbf{relint}(\mathrm{dom}(f^*))$ iff every direction of recession of f is a direction in which f is constant.

(3) For the infimum of f to be finite but unattained, it is necessary and sufficient that $f^*(0)$ be finite and $(f^*)'(0; y) = -\infty$ for some $y \in \mathbb{R}^n$.

(4) The minimum set of f is a nonempty bounded set iff $0 \in \mathrm{int}(\mathrm{dom}(f^*))$. This condition holds iff f has no directions of recession.

(5) The minimum set of f consists of a unique vector x iff f^* is differentiable at x and $x = \nabla f^*_0$.

(6) For each $\alpha \in \mathbb{R}$, the support function of $\mathrm{sublev}_\alpha(f)$ is the closure of the positively homogeneous convex function generated by $f^* + \alpha$. If f is bounded below, then the support function of the minimum set of f is the closure of the directional derivative map $y \mapsto (f^*)'(0; y)$.

In view of the importance of Theorem 51.35(4), we state this property as the following corollary.

Corollary 51.36. *Let f be a closed proper convex function on $\mathbb{R}^n$. Then the minimal set of f is a non-empty bounded set iff f has no directions of recession. In particular, if f has no directions of recession, then the minimum* $\inf f$ *of f is finite and attained for some $x \in \mathbb{R}^n$.*

Theorem 51.14 implies the following result which is very important for the design of optimization procedures.

Proposition 51.37. *Let f be a proper and closed convex function over $\mathbb{R}^n$. The function h given by $h(x) = f(x) + q(x)$ obtained by adding any strictly convex quadratic function q of the form $q(x) = x^\top Ax + b^\top x$ (where A is symmetric positive definite) is a proper closed strictly convex function such that* $\inf h$ *is finite, and there is a unique $x^* \in \mathbb{R}^n$ such that h attains its minimum in x^* (that is, $h(x^*) = \inf h$).*

Proof. By Theorem 51.14 there is some affine form φ given by $\varphi(x) = c^\top x + \alpha$ (with $\alpha \in \mathbb{R}$) such that $f(x) \geq \varphi(x)$ for all $x \in \mathbb{R}^n$. Then we have

$$h(x) = f(x) + q(x) \geq x^\top Ax + (b^\top + c^\top)x + \alpha \quad \text{for all } x \in \mathbb{R}^n.$$

Since A is symmetric positive definite, by Example 51.12, the quadratic function Q given by $Q(x) = x^\top Ax + (b^\top + c^\top)x + \alpha$ has no directions of recession. Since $h(x) \geq Q(x)$ for all $x \in \mathbb{R}^n$, we claim that h has no directions of recession. Otherwise, there would be some nonzero vector u, such that the function $\lambda \mapsto h(x + \lambda u)$ is nonincreasing for all $x \in \mathrm{dom}(h)$, so $h(x + \lambda u) \leq \beta$ for some β for all λ. But we showed that for λ large enough, the function $\lambda \mapsto Q(x + \lambda u)$ increases like λ^2, so for λ large enough, we will have $Q(x + \lambda u) > \beta$, contradicting the fact that h majorizes Q. By Corollary 51.36, h has a finite minimum x^* which is attained.

If f and g are proper convex functions and if g is strictly convex, then $f + g$ is a proper function. For all $x, y \in \mathbb{R}^n$, for any λ such that $0 < \lambda < 1$, since f is convex and g is strictly convex, we have

$$\begin{aligned} f((1-\lambda)x + \lambda y) &\leq (1-\lambda)f(x) + \lambda f(y) \\ g((1-\lambda)x + \lambda y) &< (1-\lambda)g(x) + \lambda g(y), \end{aligned}$$

so we deduce that

$$f((1-\lambda)x + \lambda y) + g((1-\lambda)x + \lambda y) < ((1-\lambda)(f(x) + g(x)) + \lambda(f(x) + g(x))),$$

which shows that $f + g$ is strictly convex. Then, as $f + q$ is strictly convex, it has a unique minimum at x^*. ∎

We now come back to the problem of minimizing a proper convex function h over a nonempty convex subset C . Here is a nice characterization.

Proposition 51.38. *Let h be a proper convex function on $\mathbb{R}^n$, and let C be a nonempty convex subset of $\mathbb{R}^n$.*

(1) *For any $x \in \mathbb{R}^n$, if there is some $y \in \partial h(x)$ such that $-y \in N_C(x)$, that is, $-y$ is normal to C at x, then h attains its minimum on C at x.*

(2) *If* $\mathbf{relint}(\mathrm{dom}(h)) \cap \mathbf{relint}(C) \neq \emptyset$*, then the converse of (1) holds. This means that if h attains its minimum on C at x, then there is some $y \in \partial h(x)$ such that $-y \in N_C(x)$.*

Proposition 51.38 is proven in Rockafellar [138] (Theorem 27.4). The proof is actually quite simple.

Proof. (1) By Proposition 51.34, h attains its minimum on C at x iff

$$0 \in \partial(h + I_C)(x).$$

By Proposition 51.23, since

$$\partial h(x) + \partial I_C(x) \subseteq \partial(h + I_C)(x),$$

if $0 \in \partial h(x) + \partial I_C(x)$, then h attains its minimum on C at x. But we saw in Section 51.2 that $\partial I_C(x) = N_C(x)$, the normal cone to C at x. Then the condition $0 \in \partial h(x) + \partial I_C(x)$ says that there is some $y \in \partial h(x)$ such that $y + z = 0$ for some $z \in N_C(x)$, and this is equivalent to $-y \in N_C(x)$.

(2) By definition of I_C, the condition $\mathbf{relint}(\mathrm{dom}(h)) \cap \mathbf{relint}(C) \neq \emptyset$ is the hypothesis of Proposition 51.23 to have

$$\partial(h + I_C)(x) = \partial h(x) + \partial I_C(x).$$

If h attains its minimum on C at x, then by Proposition 51.34 we have $0 \in \partial(h + I_C)(x)$, so $0 \in \partial h(x) + \partial I_C(x) = \partial h(x) + N_C(x)$, and by the reasoning of Part (1), this means that there is some $y \in \partial h(x)$ such that $-y \in N_C(x)$. □

Remark: A *polyhedral function* is a convex function whose epigraph is a polyhedron. It is easy to see that Proposition 51.38(2) also holds in the following cases

(1) C is a $\mathcal{H}$-polyhedron and $\mathbf{relint}(\mathrm{dom}(h)) \cap C \neq \emptyset$

(2) h is polyhedral and $\mathrm{dom}(h) \cap \mathbf{relint}(C) \neq \emptyset$.

(3) Both h and C are polyhedral, and $\mathrm{dom}(h) \cap C \neq \emptyset$.

51.6 Generalization of the Lagrangian Framework

Essentially all the results presented in Section 50.3, Section 50.7, Section 50.8, and Section 50.9 about Lagrangians and Lagrangian duality generalize to programs involving a proper and convex objective function J, proper and convex inequality constraints, and affine equality constraints. The extra generality is that it is no longer assumed that these functions are differentiable. This theory is thoroughly discussed in Part VI, Section 28, of Rockafellar [138], for programs called ordinary convex programs. We do not have the space to even sketch this theory but we will spell out some of the key results.

We will be dealing with programs consisting of an objective function $J\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ which is convex and proper, subject to $m \geq 0$ inequality contraints $\varphi_i(v) \leq 0$, and $p \geq 0$ affine equality constraints $\psi_j(v) = 0$. The constraint functions φ_i are also convex and proper, and we assume that

$$\mathbf{relint}(\mathrm{dom}(J)) \subseteq \mathbf{relint}(\mathrm{dom}(\varphi_i)), \qquad \mathrm{dom}(J) \subseteq \mathrm{dom}(\varphi_i), \quad i = 1, \ldots, m.$$

Such programs are *called ordinary convex programs*. Let

$$U = \mathrm{dom}(J) \cap \{v \in \mathbb{R}^n \mid \varphi_i(v) \leq 0,\ \psi_j(v) = 0,\ 1 \leq i \leq m,\ 1 \leq j \leq p\},$$

be the set of *feasible solutions*. We are seeking elements in $u \in U$ that minimize J over U.

A generalized version of Theorem 50.18 holds under the above hypotheses on J and the constraints φ_i and ψ_j, except that in the KKT conditions, the equation involving gradients must be replaced by the following condition involving subdifferentials:

$$0 \in \partial \left(J + \sum_{i=1}^{m} \lambda_i \varphi_i + \sum_{j=1}^{p} \mu_j \psi_j \right)(u),$$

with $\lambda_i \geq 0$ for $i = 1, \ldots, m$ and $\mu_j \in \mathbb{R}$ for $j = 1, \ldots, p$ (where $u \in U$ and J attains its minimum at u).

The *Lagrangian* $L(v, \lambda, \nu)$ of our problem is defined as follows: Let

$$E_m = \{x \in \mathbb{R}^{m+p} \mid x_i \geq 0,\ 1 \leq i \leq m\}.$$

Then

$$L(v, \lambda, \mu) = \begin{cases} J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) + \sum_{j=1}^p \mu_j \psi_j(v) & \text{if } (\lambda, \mu) \in E_m,\ v \in \mathrm{dom}(J) \\ -\infty & \text{if } (\lambda, \mu) \notin E_m,\ v \in \mathrm{dom}(J) \\ +\infty & \text{if } v \notin \mathrm{dom}(J). \end{cases}$$

For *fixed values* $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, we also define the function $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ given by

$$h(x) = J(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x) + \sum_{j=1}^{p} \mu_j \psi_j(x),$$

whose effective domain is $\mathrm{dom}(J)$ (since we are assuming that $\mathrm{dom}(J) \subseteq \mathrm{dom}(\varphi_i)$, $i = 1, \ldots, m$). Thus $h(x) = L(x, \lambda, \mu)$, but h is a *function only of* x, so we denote it differently to avoid confusion (also, technically, $L(x, \lambda, \mu)$ may take the value $-\infty$, but h does not). Since J and the φ_i are proper convex functions and the ψ_j are affine, the function h is a proper convex function.

A proof of a generalized version of Theorem 50.18 can be obtained by putting together Theorem 28.1, Theorem 28.2, and Theorem 28.3, in Rockafellar [138]. For the sake of completeness, we state these theorems. Here is Theorem 28.1.

Theorem 51.39. *(Theorem 28.1, Rockafellar) Let (P) be an ordinary convex program. Let $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ be Lagrange multipliers such that the infimum of the function $h = J + \sum_{i=1}^m \lambda_i \varphi_i + \sum_{j=1}^p \mu_j \psi_j$ is finite and equal to the optimal value of J over U. Let D be the minimal set of h over $\mathbb{R}^n$, and let $I = \{i \in \{1, \ldots, m\} \mid \lambda_i = 0\}$. If D_0 is the subset of D consisting of vectors x such that*

$$\begin{aligned} \varphi_i(x) &\leq 0 && \textit{for all } i \in I \\ \varphi_i(x) &= 0 && \textit{for all } i \notin I \\ \psi_j(x) &= 0 && \textit{for all } j = 1, \ldots, p, \end{aligned}$$

then D_0 is the set of minimizers of (P) over U.

And now here is Theorem 28.2.

Theorem 51.40. *(Theorem 28.2, Rockafellar) Let (P) be an ordinary convex program, and let $I \subseteq \{1, \ldots, m\}$ be the subset of indices of inequality constraints that are not affine. Assume that the optimal value of (P) is finite, and that (P) has at least one feasible solution $x \in$* **relint**$(\mathrm{dom}(J))$ *such that*

$$\varphi_i(x) < 0 \quad \textit{for all } i \in I.$$

Then there exist some Lagrange multipliers $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ (not necessarily unique) such that

(a) The infimum of the function $h = J + \sum_{i=1}^m \lambda_i \varphi_i + \sum_{j=1}^p \mu_j \psi_j$ is finite and equal to the optimal value of J over U.

The hypotheses of Theorem 51.40 are qualification conditions on the constraints, essentially Slater's conditions from Definition 50.6.

Definition 51.21. Let (P) be an ordinary convex program, and let $I \subseteq \{1, \ldots, m\}$ be the subset of indices of inequality constraints that are not affine. The constraints are *qualified* is there is a feasible solution $x \in$ **relint**$(\mathrm{dom}(J))$ such that

$$\varphi_i(x) < 0 \quad \text{for all } i \in I.$$

Finally, here is Theorem 28.3 from Rockafellar [138].

Theorem 51.41. *(Theorem 28.3, Rockafellar) Let (P) be an ordinary convex program. If $x \in \mathbb{R}^n$ and $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, then (λ, μ) and x have the property that*

(a) The infimum of the function $h = J + \sum_{i=1}^m \lambda_i \varphi_i + \sum_{j=1}^p \mu_j \psi_j$ is finite and equal to the optimal value of J over U, and

(b) The vector x is an optimal solution of (P) (so $x \in U$),

iff (x, λ, μ) is a saddle point of the Lagrangian $L(x, \lambda, \mu)$ of (P).

Moreover, this condition holds iff the following KKT conditions hold:

(1) $\lambda \in \mathbb{R}^m_+$, $\varphi_i(x) \leq 0$, and $\lambda_i \varphi_i(x) = 0$ for $i = 1, \ldots, m$.

(2) $\psi_j(x) = 0$ for $j = 1, \ldots, p$.

(3) $0 \in \partial J(x) + \sum_{i=1}^m \lambda_i \partial \varphi_i(x) + \sum_{j=1}^p \mu_j \partial \psi_j(x)$.

Observe that by Theorem 51.40, if the optimal value of (P) is finite and if the constraints are qualified, then Condition (a) of Theorem 51.41 holds for (λ, μ). As a consequence we obtain the following corollary of Theorem 51.41 attributed to Kuhn and Tucker, which is one of the main results of the theory. It is a generalized version of Theorem 50.18.

Theorem 51.42. *(Theorem 28.3.1, Rockafellar) Let (P) be an ordinary convex program satisfying the hypothesis of Theorem 51.40, which means that the optimal value of (P) is finite, and that the constraints are qualified. In order that a vector $x \in \mathbb{R}^n$ be an optimal solution to (P), it is necessary and sufficient that there exist Lagrange multipliers $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ such that (x, λ, μ) is a saddle point of $L(x, \lambda, \mu)$. Equivalently, x is an optimal solution of (P) if and only if there exist Lagrange multipliers $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, which, together with x, satisfy the KKT conditions from Theorem 51.41.*

Theorem 51.42 has to do with the existence of an optimal solution for (P), but it does not say anything about the optimal value of (P). To establish such a result, we need the notion of dual function.

The *dual function* G is defined by

$$G(\lambda, \mu) = \inf_{v \in \mathbb{R}^n} L(v, \lambda, \mu).$$

It is a concave function (so $-G$ is convex) which may take the values $\pm\infty$. Note that maximizing G, which is equivalent to minimizing $-G$, runs into troubles if $G(\lambda, \mu) = +\infty$ for some λ, μ, but that $G(\lambda, \mu) = -\infty$ does not cause a problem. At first glance, this seems counterintuitive, but remember that G is *concave*, not *convex*. It is $-G$ that is convex, and $-\infty$ and $+\infty$ get flipped.

Then a generalized and stronger version of Theorem 50.19(2) also holds. A proof can be obtained by putting together Corollary 28.3.1, Theorem 28.4, and Corollary 28.4.1, in Rockafellar [138]. For the sake of completeness, we state the following results from Rockafellar [138].

Theorem 51.43. *(Theorem 28.4, Rockafellar) Let (P) be an ordinary convex program with Lagrangian $L(x, \lambda, \mu)$. If the Lagrange multipliers $(\lambda^*, \mu^*) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ and the vector $x^* \in \mathbb{R}^n$ have the property that*

(a) The infimum of the function $h = J + \sum_{i=1}^m \lambda_i^* \varphi_i + \sum_{j=1}^p \mu_j^* \psi_j$ is finite and equal to the optimal value of J over U, and

(b) The vector x^* is an optimal solution of (P) (so $x^* \in U$),

then the saddle value $L(x^*, \lambda^*, \mu^*)$ is the optimal value $J(x^*)$ of (P).

More generally, the Lagrange multipliers $(\lambda^*, \mu^*) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ have Property (a) iff

$$-\infty < \inf_x L(x, \lambda^*, \mu^*) \leq \sup_{\lambda,\mu} \inf_x L(x, \lambda, \mu) = \inf_x \sup_{\lambda,\mu} L(x, \lambda, \mu),$$

in which case, the common value of the extremum value is the optimal value of (P). In particular, if x^* is an optimal solution for (P), then $\sup_{\lambda,\mu} G(\lambda, \mu) = L(x^*, \lambda^*, \mu^*) = J(x^*)$ (zero duality gap).

Observe that Theorem 51.43 gives sufficient Conditions (a) and (b) for the duality gap to be zero. In view of Theorem 51.41, these conditions are equivalent to the fact that (x^*, λ^*, μ^*) is a saddle point of L, or equivalently that the KKT conditions hold.

Again, by Theorem 51.40, if the optimal value of (P) is finite and if the constraints are qualified, then Condition (a) of Theorem 51.43 holds for (λ, μ). Then the following corollary of Theorem 51.43 holds.

Theorem 51.44. *(Theorem 28.4.1, Rockafellar) Let (P) be an ordinary convex program satisfying the hypothesis of Theorem 51.40, which means that the optimal value of (P) is finite, and that the constraints are qualified. The Lagrange multipliers $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ that have the property that the infimum of the function $h = J + \sum_{i=1}^m \lambda_i \varphi_i + \sum_{j=1}^p \mu_j \psi_j$ is finite and equal to the optimal value of J over U are exactly the vectors where the dual function G attains is supremum over $\mathbb{R}^n$.*

Theorem 51.44 is a generalized and stronger version of Theorem 50.19(2). Part (1) of Theorem 50.19 requires J and the φ_i to be differentiable, so it does not generalize.

More results can shown about ordinary convex programs, and another class of programs called *generalized convex programs*. However, we do not need such resuts for our purposes, in particular to discuss the ADMM method. The interested reader is referred to Rockafellar [138] (Part VI, Sections 28 and 29).

51.7 Summary

The main concepts and results of this chapter are listed below:

- Extended real-valued functions.

- Epigraph ($\mathbf{epi}(f)$).
- Convex and concave (extended real-valued) functions.
- Effective domain ($\mathrm{dom}(f)$).
- Proper and improper convex functions.
- Sublevel sets.
- Lower semi-continuous functions.
- Lower semi-continuous hull; closure of a convex function.
- Relative interior ($\mathbf{relint}(C)$).
- Indicator function.
- Lipschitz condition.
- Affine form, affine hyperplane.
- Half spaces.
- Supporting hyperplane.
- Normal cone at a.
- Subgradient, subgradient inequality, subdifferential.
- Minkowski's supporting hyperplane theorem.
- One-sided directional derivative.
- Support function.
- ReLU function.
- ϵ-subgradient.
- Minimum set of a convex function.
- Direction of recession.
- Ordinary convex programs.
- Set of feasible solutions.
- Lagrangian.

- Saddle point.
- KKT conditions.
- Qualified constraints.
- Duality gap.

51.8 Problems

Problem 51.1. Prove Proposition 51.1.

Problem 51.2. Prove Proposition 51.2.

Problem 51.3. Prove Proposition 51.3.

Problem 51.4. Prove that the convex function defined in Example 51.4 has the property that the limit along any line segment from $(0,0)$ to a point in the open right half-plane is 0.

Problem 51.5. Check that the normal cone to C at a is a convex cone.

Problem 51.6. Prove that $\partial f(x)$ is closed and convex.

Problem 51.7. For Example 51.6, with $f(x) = \|x\|_\infty$, prove that $\partial f(0)$ is the polyhedron

$$\partial f(0) = \mathrm{conv}\{\pm e_1, \ldots, \pm e_n\}.$$

Problem 51.8. For Example 51.7, with

$$f(x) = \begin{cases} -(1 - |x|^2)^{1/2} & \text{if } |x| \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

prove that f is subdifferentiable (in fact differentiable) at x when $|x| < 1$, but $\partial f(x) = \emptyset$ when $|x| \geq 1$, even though $x \in \mathrm{dom}(f)$ for $|x| = 1$

Problem 51.9. Prove Proposition 51.15.

Problem 51.10. Prove that as a convex function of u, the effective domain of the function $u \mapsto f'(x; u)$ is the convex cone generated by $\mathrm{dom}(f) - x$.

Problem 51.11. Prove Proposition 51.28.

Problem 51.12. Prove Proposition 51.33.

Problem 51.13. Prove that Proposition 51.38(2) also holds in the following cases:

(1) C is a $\mathcal{H}$-polyhedron and $\mathbf{relint}(\mathrm{dom}(h)) \cap C \neq \emptyset$

(2) h is polyhedral and $\mathrm{dom}(h) \cap \mathbf{relint}(C) \neq \emptyset$.

(3) Both h and C are polyhedral, and $\mathrm{dom}(h) \cap C \neq \emptyset$.

Chapter 52

Dual Ascent Methods; ADMM

This chapter is devoted to the presentation of one of the best methods known at the present for solving optimization problems involving equality constraints. In fact, this method can also handle more general constraints, namely, membership in a convex set. It can also be used to solve a range of problems arising in machine learning including *lasso minimization*, *elastic net regression*, *support vector machine (SVM)*, and *ν-SV regression*. In order to obtain a good understanding of this method, called the *alternating direction method of multipliers*, for short *ADMM*, we review two precursors of ADMM, the *dual ascent method* and the *method of multipliers*.

ADMM is not a new method. In fact, it was developed in the 1970's. It has been revived as a very effective method to solve problems in statistical and machine learning dealing with very large data because it is well suited to distributed (convex) optimization. An extensive presentation of ADMM, its variants, and its applications, is given in the excellent paper by Boyd, Parikh, Chu, Peleato and Eckstein [28]. This paper is essentially a book on the topic of ADMM, and our exposition is deeply inspired by it.

In this chapter, we consider the problem of minimizing a convex function J (not necessarily differentiable) under the equality constraints $Ax = b$. In Section 52.1 we discuss the dual ascent method. It is essentially gradient descent applied to the dual function G, but since G is maximized, gradient descent becomes gradient ascent.

In order to make the minimization step of the dual ascent method more robust, one can use the trick of adding the penalty term $(\rho/2)\left\|Au - b\right\|_2^2$ to the Lagrangian. We obtain the *augmented Lagrangian*

$$L_\rho(u, \lambda) = J(u) + \lambda^\top (Au - b) + (\rho/2)\left\|Au - b\right\|_2^2,$$

with $\lambda \in \mathbb{R}^m$, and where $\rho > 0$ is called the *penalty parameter*. We obtain the minimization Problem (P_ρ),

$$\begin{aligned} \text{minimize} \quad & J(u) + (\rho/2)\left\|Au - b\right\|_2^2 \\ \text{subject to} \quad & Au = b, \end{aligned}$$

which is equivalent to the original problem.

The benefit of adding the penalty term $(\rho/2)\left\|Au-b\right\|_2^2$ is that by Proposition 51.37, Problem (P_ρ) has a unique optimal solution under mild conditions on A. Dual ascent applied to the dual of (P_ρ) is called the *method of multipliers* and is discussed in Section 52.2.

The alternating direction method of multipliers, for short ADMM, combines the decomposability of dual ascent with the superior convergence properties of the method of multipliers. The idea is to split the function J into two independent parts, as $J(x,z)=f(x)+g(z)$, and to consider the Minimization Problem (P_{admm}),

$$\begin{aligned} &\text{minimize} && f(x)+g(z)\\ &\text{subject to} && Ax+Bz=c, \end{aligned}$$

for some $p\times n$ matrix A, some $p\times m$ matrix B, and with $x\in\mathbb{R}^n$, $z\in\mathbb{R}^m$, and $c\in\mathbb{R}^p$. We also assume that f and g are convex. Further conditions will be added later.

As in the method of multipliers, we form the augmented Lagrangian

$$L_\rho(x,z,\lambda)=f(x)+g(z)+\lambda^\top(Ax+Bz-c)+(\rho/2)\left\|Ax+Bz-c\right\|_2^2,$$

with $\lambda\in\mathbb{R}^p$ and for some $\rho>0$. The major difference with the method of multipliers is that instead of performing a minimization step jointly over x and z, ADMM *first performs an x-minimization step and then a z-minimization step*. Thus x and z are updated in an alternating or sequential fashion, which accounts for the term *alternating direction*. Because the Lagrangian is augmented, some mild conditions on A and B imply that these minimization steps are guaranteed to terminate. ADMM is presented in Section 52.3.

In Section 52.4 we prove the convergence of ADMM under the following assumptions:

(1) The functions $f\colon\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ and $g\colon\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ are proper and closed convex functions (see Section 51.1) such that $\mathbf{relint}(\mathrm{dom}(f))\cap\mathbf{relint}(\mathrm{dom}(g))\neq\emptyset$.

(2) The $n\times n$ matrix $A^\top A$ is invertible and the $m\times m$ matrix $B^\top B$ is invertible. Equivalently, the $p\times n$ matrix A has rank n and the $p\times m$ matrix has rank m.

(3) The unaugmented Lagrangian $L_0(x,z,\lambda)=f(x)+g(z)+\lambda^\top(Ax+Bz-c)$ has a saddle point, which means there exists x^*,z^*,λ^* (not necessarily unique) such that

$$L_0(x^*,z^*,\lambda)\leq L_0(x^*,z^*,\lambda^*)\leq L_0(x,z,\lambda^*)$$

for all x,z,λ.

By Theorem 51.41, Assumption (3) is equivalent to the fact that the KKT equations are satisfied by some triple (x^*,z^*,λ^*), namely

$$Ax^*+Bz^*-c=0 \tag{$*$}$$

and

$$0 \in \partial f(x^*) + \partial g(z^*) + A^\top \lambda^* + B^\top \lambda^*, \tag{$\dagger$}$$

Assumption (3) is also equivalent to Conditions (a) and (b) of Theorem 51.41. In particular, our program has an optimal solution (x^*, z^*). By Theorem 51.43, λ^* is maximizer of the dual function $G(\lambda) = \inf_{x,z} L_0(x, z, \lambda)$ and strong duality holds, that is, $G(\lambda^*) = f(x^*) + g(z^*)$ (the duality gap is zero).

We will show after the proof of Theorem 52.1 that Assumption (2) is actually implied by Assumption (3). This allows us to prove a convergence result stronger than the convergence result proven in Boyd et al. [28] (under the exact same assumptions (1) and (3)). In particular, we prove that *all* of the sequences (x^k), (z^k), and (λ^k) converge to optimal solutions $(\widetilde{x}, \widetilde{z})$, and $\widetilde{\lambda}$. The core of our proof is due to Boyd et al. [28], but there are new steps because we have the stronger hypothesis (2).

In Section 52.5, we discuss stopping criteria.

In Section 52.6 we present some applications of ADMM, in particular, minimization of a proper closed convex function f over a closed convex set C in $\mathbb{R}^n$ and quadratic programming. The second example provides one of the best methods for solving quadratic problems, including the SVM problems discussed in Chapter 54, the elastic net method in Section 55.6, and ν-SV regression in Chapter 56.

Section 52.8 gives applications of ADMM to ℓ^1-norm problems, in particular, lasso regularization, which plays an important role in machine learning.

52.1 Dual Ascent

Our goal is to solve the minimization problem, Problem (P),

$$\begin{aligned} &\text{minimize} && J(u) \\ &\text{subject to} && Au = b, \end{aligned}$$

with affine equality constraints (with A an $m \times n$ matrix and $b \in \mathbb{R}^m$). The Lagrangian $L(u, \lambda)$ of Problem (P) is given by

$$L(u, \lambda) = J(u) + \lambda^\top (Au - b).$$

with $\lambda \in \mathbb{R}^m$. From Proposition 50.20, the dual function $G(\lambda) = \inf_{u \in \mathbb{R}^n} L(u, \lambda)$ is given by

$$G(\lambda) = \begin{cases} -b^\top \lambda - J^*(-A^\top \lambda) & \text{if } -A^\top \lambda \in \mathrm{dom}(J^*), \\ -\infty & \text{otherwise}, \end{cases}$$

for all $\lambda \in \mathbb{R}^m$, where J^* is the conjugate of J. Recall that by Definition 50.11, the *conjugate* f^* of a function $f\colon U \to \mathbb{R}$ defined on a subset U of $\mathbb{R}^n$ is the partial function $f^*\colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f^*(y) = \sup_{x \in U} (y^\top x - f(x)), \quad y \in \mathbb{R}^n.$$

If the conditions of Theorem 50.19(1) hold, which in our case means that for every $\lambda \in \mathbb{R}^m$, there is a unique $u_\lambda \in \mathbb{R}^n$ such that

$$G(\lambda) = L(u_\lambda, \lambda) = \inf_{u \in \mathbb{R}^n} L(u, \lambda),$$

and that the function $\lambda \mapsto u_\lambda$ is continuous, then G is differentiable. Furthermore, we have

$$\nabla G_\lambda = Au_\lambda - b,$$

and for any solution $\mu = \lambda^*$ of the dual problem

$$\begin{aligned} &\text{maximize} && G(\lambda) \\ &\text{subject to} && \lambda \in \mathbb{R}^m, \end{aligned}$$

the vector $u^* = u_\mu$ is a solution of the primal Problem (P). Furthermore, $J(u^*) = G(\lambda^*)$, that is, the duality gap is zero.

The dual ascent method is essentially gradient descent applied to the dual function G. But since G is maximized, gradient descent becomes gradient ascent. Also, we no longer worry that the minimization problem $\inf_{u \in \mathbb{R}^n} L(u, \lambda)$ has a unique solution, so we denote by u^+ some minimizer of the above problem, namely

$$u^+ = \arg\min_u L(u, \lambda).$$

Given some initial dual variable λ^0, the *dual ascent method* consists of the following two steps:

$$\begin{aligned} u^{k+1} &= \arg\min_u L(u, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \alpha^k(Au^{k+1} - b), \end{aligned}$$

where $\alpha^k > 0$ is a step size. The first step is used to compute the "new gradient" (indeed, if the minimizer u^{k+1} is unique, then $\nabla G_{\lambda^k} = Au^{k+1} - b$), and the second step is a dual variable update.

Example 52.1. Let us look at a very simple example of the gradient ascent method applied to a problem we first encountered in Section 42.1, namely minimize $J(x, y) = (1/2)(x^2 + y^2)$ subject to $2x - y = 5$. The Lagrangian is

$$L(x, y, \lambda) = \frac{1}{2}(x^2 + y^2) + \lambda(2x - y - 5).$$

See Figure 52.1.

The method of Lagrangian duality says first calculate

$$G(\lambda) = \inf_{(x,y) \in \mathbb{R}^2} L(x, y, \lambda).$$

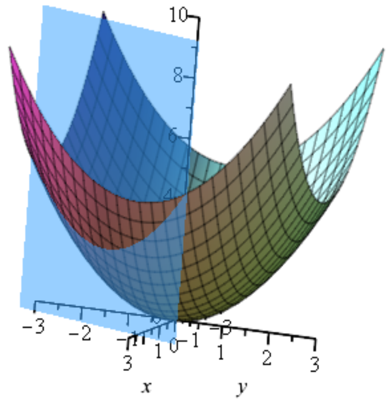


Figure 52.1: The graph of $J(x,y) = (1/2)(x^2+y^2)$ is the parabolic surface while the graph of $2x - y = 5$ is the transparent blue plane. The solution to Example 52.1 is apex of the intersection curve, namely the point $(2, -1, \frac{5}{2})$.

Since

$$J(x,y) = \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},$$

we observe that $J(x,y)$ is a quadratic function determined by the positive definite matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and hence to calculate $G(\lambda)$, we must set $\nabla L_{x,y} = 0$. By calculating $\frac{\partial J}{\partial x} = 0$ and $\frac{\partial J}{\partial y} = 0$, we find that $x = -2\lambda$ and $y = \lambda$. Then $G(\lambda) = -5/2\lambda^2 - 5\lambda$, and we must calculate the maximum of $G(\lambda)$ with respect to $\lambda \in \mathbb{R}$. This means calculating $G'(\lambda) = 0$ and obtaining $\lambda = -1$ for the solution of $(x, y, \lambda) = (-2\lambda, \lambda, \lambda) = (2, -1, -1)$.

Instead of solving *directly* for λ in terms of (x,y), the method of dual assent begins with a *numerical* estimate for λ, namely λ^0, and forms the "numerical" Lagrangian

$$L(x, y, \lambda^0) = \frac{1}{2}(x^2 + y^2) + \lambda^0(2x - y - 5).$$

With this numerical value λ^0, we minimize $L(x, y, \lambda^0)$ with respect to (x,y). This calculation will be identical to that used to form $G(\lambda)$ above, and as such, we obtain the iterative step $(x^1, y^1) = (-2\lambda^0, \lambda^0)$. So if we replace λ^0 by λ^k, we have the first step of the dual ascent method, namely

$$u^{k+1} = \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}\lambda^k.$$

The second step of the dual ascent method refines the numerical estimate of λ by calculating

$$\lambda^{k+1} = \lambda^k + \alpha^k\left(\begin{pmatrix} 2 & -1 \end{pmatrix}\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} - 5\right).$$

(Recall that in our original problem the constraint is $2x - y = 5$ or $\begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - 5$, so $A = \begin{pmatrix} 2 & -1 \end{pmatrix}$ and $b = 5$.) By simplifying the above equation, we find that

$$\lambda^{k+1} = (1 - \beta)\lambda^k - \beta, \qquad \beta = 5\alpha^k.$$

Back substituting for λ^k in the preceding equation shows that

$$\lambda^{k+1} = (1 - \beta)^{k+1}\lambda^0 + (1 - \beta)^{k+1} - 1.$$

If $0 < \beta \leq 1$, the preceding line implies that λ^{k+1} converges to $\lambda = -1$, which coincides with the answer provided by the original Lagrangian duality method. Observe that if $\beta = 1$ or $\alpha^k = \frac{1}{5}$, the dual ascent method terminates in one step.

With an appropriate choice of α^k, we have $G(\lambda^{k+1}) > G(\lambda^k)$, so the method makes progress. Under certain assumptions, for example, that J is strictly convex and some conditions of the α^k, it can be shown that dual ascent converges to an optimal solution (both for the primal and the dual). However, the main flaw of dual ascent is that the minimization step may diverge. For example, this happens is J is a nonzero affine function of one of its components. The remedy is to add a penalty term to the Lagrangian.

On the positive side, the dual ascent method leads to a decentralized algorithm if the function J is separable. Suppose that u can be split as $u = \sum_{i=1}^N u_i$, with $u_i \in \mathbb{R}^{n_i}$ and $n = \sum_{i=1}^N n_i$, that

$$J(u) = \sum_{i=1}^N J_i(u_i),$$

and that A is split into N blocks A_i (with A_i a $m \times n_i$ matrix) as $A = [A_1 \ \cdots \ A_N]$, so that $Au = \sum_{k=1}^N A_i u_i$. Then the Lagrangian can be written as

$$L(u, \lambda) = \sum_{i=1}^N L_i(u_i, \lambda),$$

with

$$L_i(u_i, \lambda) = J_i(u_i) + \lambda^\top \left(A_i u_i - \frac{1}{N} b \right).$$

it follows that the minimization of $L(u, \lambda)$ with respect to the primal variable u can be split into N separate minimization problems that can be solved in parallel. The algorithm then performs the N updates

$$u_i^{k+1} = \operatorname*{arg\,min}_{u_i} L_i(u_i, \lambda^k)$$

in parallel, and then the step

$$\lambda^{k+1} = \lambda^k + \alpha^k (Au^{k+1} - b).$$

52.2 Augmented Lagrangians and the Method of Multipliers

In order to make the minimization step of the dual ascent method more robust, one can use the trick of adding the penalty term $(\rho/2)\left\|Au-b\right\|_2^2$ to the Lagrangian.

Definition 52.1. Given the Optimization Problem (P),

$$\begin{aligned} &\text{minimize} \quad J(u) \\ &\text{subject to} \quad Au = b, \end{aligned}$$

the *augmented Lagrangian* is given by

$$L_\rho(u,\lambda) = J(u) + \lambda^\top(Au-b) + (\rho/2)\left\|Au-b\right\|_2^2,$$

with $\lambda \in \mathbb{R}^m$, and where $\rho > 0$ is called the *penalty parameter*.

The augmented Lagrangian $L_\rho(u,\lambda)$ can be viewed as the ordinary Lagrangian of the Minimization Problem (P_ρ),

$$\begin{aligned} &\text{minimize} \quad J(u) + (\rho/2)\left\|Au-b\right\|_2^2 \\ &\text{subject to} \quad Au = b. \end{aligned}$$

The above problem is equivalent to Program (P), since for any feasible solution of (P_ρ), we must have $Au - b = 0$.

The benefit of adding the penalty term $(\rho/2)\left\|Au-b\right\|_2^2$ is that by Proposition 51.37, Problem (P_ρ) has a unique optimal solution under mild conditions on A.

Dual ascent applied to the dual of (P_ρ) yields the the *method of multipliers*, which consists of the following steps, given some initial λ^0:

$$\begin{aligned} u^{k+1} &= \arg\min_u L_\rho(u,\lambda^k) \\ \lambda^{k+1} &= \lambda^k + \rho(Au^{k+1} - b). \end{aligned}$$

Observe that the second step uses the parameter ρ. The reason is that it can be shown that choosing $\alpha^k = \rho$ guarantees that (u^{k+1},λ^{k+1}) satisfies the equation

$$\nabla J_{u^{k+1}} + A^\top\lambda^{k+1} = 0,$$

which means that (u^{k+1},λ^{k+1}) is dual feasible; see Boyd, Parikh, Chu, Peleato and Eckstein [28], Section 2.3.

Example 52.2. Consider the minimization problem

$$\begin{aligned} &\text{minimize} \quad y^2 + 2x \\ &\text{subject to} \quad 2x - y = 0. \end{aligned}$$

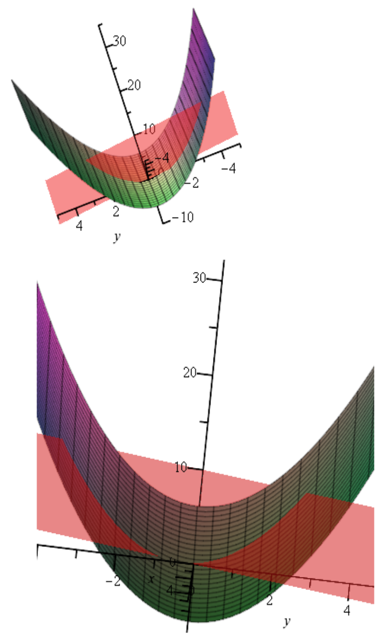


Figure 52.2: Two views of the graph of $y^2 + 2x$ intersected with the transparent red plane $2x - y = 0$. The solution to Example 52.2 is apex of the intersection curve, namely the point $(-\frac{1}{4}, -\frac{1}{2}, -\frac{15}{16})$.

See Figure 52.2.

The quadratic function

$$J(x, y) = y^2 + 2x = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

is convex but not strictly convex. Since $y = 2x$, the problem is equivalent to minimizing $y^2 + 2x = 4x^2 + 2x$, whose minimum is achieved for $x = -1/4$ (since setting the derivative of the function $x \mapsto 4x^2 + 2$ yields $8x + 2 = 0$). Thus, the unique minimum of our problem is achieved for $(x = -1/4, y = -1/2)$. The Lagrangian of our problem is

$$L(x, y, \lambda) = y^2 + 2x + \lambda(2x - y).$$

If we apply the dual ascent method, minimization of $L(x, y, \lambda)$ with respect to x and y holding λ constant yields the equations

$$\begin{aligned} 2 + 2\lambda &= 0 \\ 2y - \lambda &= 0, \end{aligned}$$

obtained by setting the gradient of L (with respect to x and y) to zero. If $\lambda \neq -1$, the problem has no solution. Indeed, if $\lambda \neq -1$, minimizing $L(x, y, \lambda) = y^2 + 2x + \lambda(2x - y)$ with respect to x and y yields $-\infty$.

The augmented Lagrangian is

$$\begin{aligned} L_\rho(x, y, \lambda) &= y^2 + 2x + \lambda(2x - y) + (\rho/2)(2x - y)^2 \\ &= 2\rho x^2 - 2\rho xy + 2(1+\lambda)x - \lambda y + \left(1 + \frac{\rho}{2}\right) y^2, \end{aligned}$$

which in matrix form is

$$L_\rho(x, y, \lambda) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 2\rho^2 & -\rho \\ -\rho & 1 + \dfrac{\rho}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2(1+\lambda) & -\lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The trace of the above matrix is $1 + \frac{\rho}{2} + 2\rho^2 > 0$, and the determinant is

$$2\rho^2 \left(1 + \frac{\rho}{2}\right) - \rho^2 = \rho^2(1 + \rho) > 0,$$

since $\rho > 0$. Therefore, the above matrix is symmetric positive definite. Minimizing $L_\rho(x, y, \lambda)$ with respect to x and y, we set the gradient of $L_\rho(x, y, \lambda)$ (with respect to x and y) to zero, and we obtain the equations:

$$\begin{aligned} 2\rho x - \rho y + (1 + \lambda) &= 0 \\ -2\rho x + (2 + \rho) y - \lambda &= 0. \end{aligned}$$

The solution is

$$x = -\frac{1}{4} - \frac{1 + \lambda}{2\rho}, \quad y = -\frac{1}{2}.$$

Thus the steps for the method of multipliers are

$$\begin{aligned} x^{k+1} &= -\frac{1}{4} - \frac{1 + \lambda^k}{2\rho} \\ y^{k+1} &= -\frac{1}{2} \\ \lambda^{k+1} &= \lambda^k + \rho \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} -\frac{1}{4} - \frac{1+\lambda^k}{2\rho} \\ -\frac{1}{2} \end{pmatrix}, \end{aligned}$$

and the second step simplifies to

$$\lambda^{k+1} = -1.$$

Consequently, we see that the method converges after two steps for any initial value of λ^0, and we get

$$x = -\frac{1}{4} \quad y = -\frac{1}{2}, \quad \lambda = -1.$$

The method of multipliers also converges for functions J that are not even convex, as illustrated by the next example.

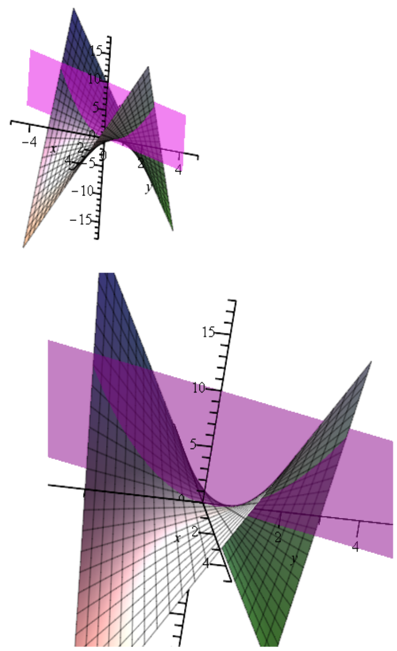


Figure 52.3: Two views of the graph of the saddle of $2xy$ ($\beta = 1$) intersected with the transparent magenta plane $2x - y = 0$. The solution to Example 52.3 is apex of the intersection curve, namely the point $(0, 0, 0)$.

Example 52.3. Consider the minimization problem

$$\begin{aligned} &\text{minimize} \quad 2\beta xy \\ &\text{subject to} \;\; 2x - y = 0, \end{aligned}$$

with $\beta > 0$. See Figure 52.3.

The quadratic function

$$J(x, y) = 2\beta xy = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 0 & \beta \\ \beta & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

is not convex because the above matrix is not even positive semidefinite (the eigenvalues of the matrix are $-\beta$ and $+\beta$). The augmented Lagrangian is

$$\begin{aligned} L_\rho(x, y, \lambda) &= 2\beta xy + \lambda(2x - y) + (\rho/2)(2x - y)^2 \\ &= 2\rho x^2 + 2(\beta - \rho)xy + 2\lambda x - \lambda y + \frac{\rho}{2} y^2, \end{aligned}$$

which in matrix form is

$$L_\rho(x, y, \lambda) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 2\rho & \beta - \rho \\ \beta - \rho & \dfrac{\rho}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2\lambda & -\lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The trace of the above matrix is $2\rho + \frac{\rho}{2} = \frac{5}{2}\rho > 0$, and the determinant is

$$\rho^2 - (\beta - \rho)^2 = \beta(2\rho - \beta).$$

This determinant is positive if $\rho > \beta/2$, in which case the matrix is symmetric positive definite. Minimizing $L_\rho(x, y, \lambda)$ with respect to x and y, we set the gradient of $L_\rho(x, y, \lambda)$ (with respect to x and y) to zero, and we obtain the equations:

$$\begin{aligned} 2\rho x + (\beta - \rho)y + \lambda &= 0 \\ 2(\beta - \rho)x + \rho y - \lambda &= 0. \end{aligned}$$

Since we are assuming that $\rho > \beta/2$, the solutions are

$$x = -\frac{\lambda}{2(2\rho - \beta)}, \quad y = \frac{\lambda}{(2\rho - \beta)}.$$

Thus the steps for the method of multipliers are

$$\begin{aligned} x^{k+1} &= -\frac{\lambda^k}{2(2\rho - \beta)} \\ y^{k+1} &= \frac{\lambda^k}{(2\rho - \beta)} \\ \lambda^{k+1} &= \lambda^k + \frac{\rho}{2(2\rho - \beta)} \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} -\lambda^k \\ 2\lambda^k \end{pmatrix}, \end{aligned}$$

and the second step simplifies to

$$\lambda^{k+1} = \lambda^k + \frac{\rho}{2(2\rho - \beta)}(-4\lambda^k),$$

that is,

$$\lambda^{k+1} = -\frac{\beta}{2\rho - \beta}\lambda^k.$$

If we pick $\rho > \beta > 0$, which implies that $\rho > \beta/2$, then

$$\frac{\beta}{2\rho - \beta} < 1,$$

and the method converges for any intial value λ^0 to the solution

$$x = 0, \quad y = 0, \quad \lambda = 0.$$

Indeed, since the constraint $2x - y = 0$ holds, $2\beta xy = 4\beta x^2$, and the minimum of the function $x \mapsto 4\beta x^2$ is achieved for $x = 0$ (since $\beta > 0$).

As an exercise, the reader should verify that dual ascent (with $\alpha^k = \rho$) yields the equations

$$\begin{aligned} x^{k+1} &= \frac{\lambda^k}{2\beta} \\ y^{k+1} &= -\frac{\lambda^k}{\beta} \\ \lambda^{k+1} &= \left(1 + \frac{2\rho}{\beta}\right)\lambda^k, \end{aligned}$$

and so the method diverges, except for $\lambda^0 = 0$, which is the optimal solution.

The method of multipliers converges under conditions that are far more general than the dual ascent. However, the addition of the penalty term has the negative effect that even if J is separable, then the Lagrangian L_ρ is not separable. Thus the basic method of multipliers cannot be used for decomposition and is not parallelizable. The next method deals with the problem of separability.

52.3 ADMM: Alternating Direction Method of Multipliers

The alternating direction method of multipliers, for short ADMM, combines the decomposability of dual ascent with the superior convergence properties of the method of multipliers. It can be viewed as an approximation of the method of multipliers, but it is generally superior.

The idea is to split the function J into two independent parts, as $J(x, z) = f(x) + g(z)$, and to consider the Minimization Problem (P_{admm}),

$$\begin{aligned} &\text{minimize} \quad f(x) + g(z) \\ &\text{subject to} \quad Ax + Bz = c, \end{aligned}$$

for some $p \times n$ matrix A, some $p \times m$ matrix B, and with $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, and $c \in \mathbb{R}^p$. We also assume that f and g are convex. Further conditions will be added later.

As in the method of multipliers, we form the *augmented Lagrangian*

$$L_\rho(x, z, \lambda) = f(x) + g(z) + \lambda^\top (Ax + Bz - c) + (\rho/2) \left\| Ax + Bz - c \right\|_2^2,$$

with $\lambda \in \mathbb{R}^p$ and for some $\rho > 0$.

Given some initial values (z^0, λ^0), the *ADMM method* consists of the following iterative steps:

$$\begin{aligned} x^{k+1} &= \arg\min_x L_\rho(x, z^k, \lambda^k) \\ z^{k+1} &= \arg\min_z L_\rho(x^{k+1}, z, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \rho(Ax^{k+1} + Bz^{k+1} - c). \end{aligned}$$

Instead of performing a minimization step jointly over x and z, as the method of multipliers would in the step

$$(x^{k+1}, z^{k+1}) = \arg\min_{x,z} L_\rho(x, z, \lambda^k),$$

ADMM first performs an x-minimization step, and then a z-minimization step. Thus x and z are updated in an alternating or sequential fashion, which accounts for the term *alternating direction*.

The algorithm state in ADMM is (z^k, λ^k), in the sense that (z^{k+1}, λ^{k+1}) is a function of (z^k, λ^k). The variable x^{k+1} is an auxiliary variable which is used to compute z^{k+1} from (z^k, λ^k). The roles of x and z are not quite symmetric, since the update of x is done before the update of λ. By switching x and z, f and g and A and B, we obtain a variant of ADMM in which the order of the x-update step and the z-update step are reversed.

Example 52.4. Let us reconsider the problem of Example 52.2 to solve it using ADMM. We formulate the problem as

$$\begin{aligned} &\text{minimize} && 2x + z^2 \\ &\text{subject to} && 2x - z = 0, \end{aligned}$$

with $f(x) = 2x$ and $g(z) = z^2$. The augmented Lagrangian is given by

$$L_\rho(x, z, \lambda) = 2x + z^2 + 2\lambda x - \lambda z + 2\rho x^2 - 2\rho xz + \frac{\rho}{2} z^2.$$

The ADMM steps are as follows. The x-update is

$$x^{k+1} = \arg\min_x \left(2\rho x^2 - 2\rho x z^k + 2\lambda^k x + 2x\right),$$

and since this is a quadratic function in x, its minimum is achieved when the derivative of the above function (with respect to x) is zero, namely

$$x^{k+1} = \frac{1}{2} z^k - \frac{1}{2\rho}\lambda^k - \frac{1}{2\rho}. \tag{1}$$

The z-update is

$$z^{k+1} = \arg\min_z \left(z^2 + \frac{\rho}{2} z^2 - 2\rho x^{k+1} z - \lambda^k z\right),$$

and as for the x-step, the minimum is achieved when the derivative of the above function (with respect to z) is zero, namely

$$z^{k+1} = \frac{2\rho x^{k+1}}{\rho + 2} + \frac{\lambda^k}{\rho + 2}. \tag{2}$$

The λ-update is

$$\lambda^{k+1} = \lambda^k + \rho(2x^{k+1} - z^{k+1}). \tag{3}$$

Substituting the right hand side of (1) for x^{k+1} in (2) yields

$$z^{k+1} = \frac{\rho z^k}{\rho+2} - \frac{1}{\rho+2}. \tag{4}$$

Using (2), we obtain

$$2x^{k+1} - z^{k+1} = \frac{4x^{k+1}}{\rho+2} - \frac{\lambda^k}{\rho+2}, \tag{5}$$

and then using (3) we get

$$\lambda^{k+1} = \frac{2\lambda^k}{\rho+2} + \frac{4\rho x^{k+1}}{\rho+2}. \tag{6}$$

Substituting the right hand side of (1) for x^{k+1} in (6), we obtain

$$\lambda^{k+1} = \frac{2\rho z^k}{\rho+2} - \frac{2}{\rho+2}. \tag{7}$$

Equation (7) shows that z^k determines λ^{k+1}, and Equation (1) for $k+2$, along with Equation (4), shows that z^k also determines x^{k+2}. In particular, we find that

$$\begin{aligned} x^{k+2} &= \frac{1}{2}z^{k+1} - \frac{1}{2\rho}\lambda^{k+1} - \frac{1}{2\rho} \\ &= \frac{(\rho-2)z^k}{2(\rho+2)} - \frac{1}{\rho+2}. \end{aligned}$$

Thus is suffices to find the limit of the sequence (z^k). Since we already know from Example 52.2 that this limit is $-1/2$, using (4), we write

$$z^{k+1} = -\frac{1}{2} + \frac{\rho z^k}{\rho+2} - \frac{1}{\rho+2} + \frac{1}{2} = -\frac{1}{2} + \frac{\rho}{\rho+2}\left(\frac{1}{2} + z^k\right).$$

By induction, we deduce that

$$z^{k+1} = -\frac{1}{2} + \left(\frac{\rho}{\rho+2}\right)^{k+1}\left(\frac{1}{2} + z^0\right),$$

and since $\rho > 0$, we have $\rho/(\rho+2) < 1$, so the limit of the sequence (z^{k+1}) is indeed $-1/2$, and consequently the limit of (λ^{k+1}) is -1 and the limit of x^{k+2} is $-1/4$.

For ADMM to be practical, the x-minimization step and the z-minimization step have to be doable efficiently.

It is often convenient to write the ADMM updates in terms of the *scaled dual variable* $\mu = (1/\rho)\lambda$. If we define the *residual* as

$$r = Ax + bz - c,$$

then we have

$$\begin{aligned}\lambda^\top r + (\rho/2)\,\|r\|_2^2 &= (\rho/2)\,\|r + (1/\rho)\lambda\|_2^2 - (1/(2\rho))\,\|\lambda\|_2^2 \\ &= (\rho/2)\,\|r + \mu\|_2^2 - (\rho/2)\,\|\mu\|_2^2 .\end{aligned}$$

The *scaled form of ADMM* consists of the following steps:

$$\begin{aligned}x^{k+1} &= \arg\min_x \Big(f(x) + (\rho/2)\,\big\|Ax + Bz^k - c + \mu^k\big\|_2^2\Big) \\ z^{k+1} &= \arg\min_z \Big(g(z) + (\rho/2)\,\big\|Ax^{k+1} + Bz - c + \mu^k\big\|_2^2\Big) \\ \mu^{k+1} &= \mu^k + Ax^{k+1} + Bz^{k+1} - c.\end{aligned}$$

If we define the *residual* r^k *at step* k as

$$r^k = Ax^k + Bz^k - c = \mu^k - \mu^{k-1} = (1/\rho)(\lambda^k - \lambda^{k-1}),$$

then we see that

$$r = u^0 + \sum_{j=1}^k r^j.$$

The formulae in the scaled form are often shorter than the formulae in the unscaled form.

We now discuss the convergence of ADMM.

52.4 Convergence of ADMM ⊛

Let us repeat the steps of ADMM: Given some initial (z^0, λ^0), do:

$$x^{k+1} = \arg\min_x L_\rho(x, z^k, \lambda^k) \qquad (x\text{-update})$$

$$z^{k+1} = \arg\min_z L_\rho(x^{k+1}, z, \lambda^k) \qquad (z\text{-update})$$

$$\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + Bz^{k+1} - c). \qquad (\lambda\text{-update})$$

The convergence of ADMM can be proven under the following three assumptions:

(1) The functions $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ are proper and closed convex functions (see Section 51.1) such that $\mathbf{relint}(\mathrm{dom}(f)) \cap \mathbf{relint}(\mathrm{dom}(g)) \neq \emptyset$.

(2) The $n \times n$ matrix $A^\top A$ is invertible and the $m \times m$ matrix $B^\top B$ is invertible. Equivalently, the $p \times n$ matrix A has rank n and the $p \times m$ matrix has rank m.

(3) The unaugmented Lagrangian $L_0(x, z, \lambda) = f(x) + g(z) + \lambda^\top (Ax + Bz - c)$ has a saddle point, which means there exists x^*, z^*, λ^* (not necessarily unique) such that

$$L_0(x^*, z^*, \lambda) \leq L_0(x^*, z^*, \lambda^*) \leq L_0(x, z, \lambda^*)$$

for all x, z, λ.

Recall that the augmented Lagrangian is given by

$$L_\rho(x,z,\lambda) = f(x) + g(z) + \lambda^\top(Ax+Bz-c) + (\rho/2)\left\|Ax+Bz-c\right\|_2^2.$$

For z (and λ) fixed, we have

$$\begin{aligned}
L_\rho(x,z,\lambda) &= f(x) + g(z) + \lambda^\top(Ax+Bz-c) + (\rho/2)(Ax+Bz-c)^\top(Ax+Bz-c)\\
&= f(x) + (\rho/2)x^\top A^\top Ax + (\lambda^\top + \rho(Bz-c)^\top)Ax\\
&\quad + g(z) + \lambda^\top(Bz-c) + (\rho/2)(Bz-c)^\top(Bz-c).
\end{aligned}$$

Assume that (1) and (2) hold. Since $A^\top A$ is invertible, then it is symmetric positive definite, and by Proposition 51.37 the x-minimization step has a unique solution (the minimization problem succeeds with a unique minimizer).

Similarly, for x (and λ) fixed, we have

$$\begin{aligned}
L_\rho(x,z,\lambda) &= f(x) + g(z) + \lambda^\top(Ax+Bz-c) + (\rho/2)(Ax+Bz-c)^\top(Ax+Bz-c)\\
&= g(z) + (\rho/2)z^\top B^\top Bz + (\lambda^\top + \rho(Ax-c)^\top)Bz\\
&\quad + f(x) + \lambda^\top(Ax-c) + (\rho/2)(Ax-c)^\top(Ax-c).
\end{aligned}$$

Since $B^\top B$ is invertible, then it is symmetric positive definite, and by Proposition 51.37 the z-minimization step has a unique solution (the minimization problem succeeds with a unique minimizer).

By Theorem 51.41, Assumption (3) is equivalent to the fact that the KKT equations are satisfied by some triple (x^*, z^*, λ^*), namely

$$Ax^* + Bz^* - c = 0 \tag{$*$}$$

and

$$0 \in \partial f(x^*) + \partial g(z^*) + A^\top\lambda^* + B^\top\lambda^*, \tag{$\dagger$}$$

Assumption (3) is also equivalent to Conditions (a) and (b) of Theorem 51.41. In particular, our program has an optimal solution (x^*, z^*). By Theorem 51.43, λ^* is maximizer of the dual function $G(\lambda) = \inf_{x,z} L_0(x,z,\lambda)$ and strong duality holds, that is, $G(\lambda^*) = f(x^*) + g(z^*)$ (the duality gap is zero).

We will see after the proof of Theorem 52.1 that Assumption (2) is actually implied by Assumption (3). This allows us to prove a convergence result stronger than the convergence result proven in Boyd et al. [28] under the exact same assumptions (1) and (3).

Let p^* be the minimum value of $f+g$ over the convex set $\{(x,z) \in \mathbb{R}^{m+p} \mid Ax+Bz-c = 0\}$, and let (p^k) be the sequence given by $p^k = f(x^k) + g(z^k)$, and recall that $r^k = Ax^k + Bz^k - c$.

Our main goal is to prove the following result.

Theorem 52.1. *Suppose the following assumptions hold:*

(1) The functions $f\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ are proper and closed convex functions (see Section 51.1) such that $\mathbf{relint}(\mathrm{dom}(f)) \cap \mathbf{relint}(\mathrm{dom}(g)) \neq \emptyset$.

(2) The $n \times n$ matrix $A^\top A$ is invertible and the $m \times m$ matrix $B^\top B$ is invertible. Equivalently, the $p \times n$ matrix A has rank n and the $p \times m$ matrix has rank m. (This assumption is actually redundant, because it is implied by Assumption (3)).

(3) The unaugmented Lagrangian $L_0(x, z, \lambda) = f(x) + g(z) + \lambda^\top (Ax + Bz - c)$ has a saddle point, which means there exists x^*, z^*, λ^* (not necessarily unique) such that

$$L_0(x^*, z^*, \lambda) \leq L_0(x^*, z^*, \lambda^*) \leq L_0(x, z, \lambda^*)$$

for all x, z, λ.

Then for any initial values (z^0, λ^0), the following properties hold:

(1) The sequence (r^k) converges to 0 (residual convergence).

(2) The sequence (p^k) converge to p^* (objective convergence).

(3) The sequences (x^k) and (z^k) converge to an optimal solution $(\widetilde{x}, \widetilde{z})$ of Problem (P_{admm}) and the sequence (λ^k) converges an optimal solution $\widetilde{\lambda}$ of the dual problem (primal and dual variable convergence).

Proof. The core of the proof is due to Boyd et al. [28], but there are new steps because we have the stronger hypothesis (2), which yield the stronger result (3).

The proof consists of several steps. It is not possible to prove directly that the sequences (x^k), (z^k), and (λ^k) converge, so first we prove that the sequence (r^{k+1}) converges to zero, and that the sequences (Ax^{k+1}) and (Bz^{k+1}) also converge.

Step 1. Prove the inequality (A1) below.

Consider the sequence of reals (V^k) given by

$$V^k = (1/\rho) \left\|\lambda^k - \lambda^*\right\|_2^2 + \rho \left\|B(z^k - z^*)\right\|_2^2.$$

It can be shown that the V^k satisfy the following inequality:

$$V^{k+1} \leq V^k - \rho \left\|r^{k+1}\right\|_2^2 - \rho \left\|B(z^{k+1} - z^k)\right\|_2^2. \tag{A1}$$

This is rather arduous. Since a complete proof is given in Boyd et al. [28], we will only provide some of the key steps later.

Inequality (A1) shows that the sequence (V^k) in nonincreasing. If we write these inequalities for $k, k-1, \ldots, 0$, we have

$$\begin{aligned}
V^{k+1} &\leq V^k - \rho \left\|r^{k+1}\right\|_2^2 - \rho \left\|B(z^{k+1} - z^k)\right\|_2^2 \\
V^k &\leq V^{k-1} - \rho \left\|r^k\right\|_2^2 - \rho \left\|B(z^k - z^{k-1})\right\|_2^2 \\
&\vdots \\
V^1 &\leq V^0 - \rho \left\|r^1\right\|_2^2 - \rho \left\|B(z^1 - z^0)\right\|_2^2,
\end{aligned}$$

and by adding up these inequalities, we obtain

$$V^{k+1} \leq V^0 - \rho \sum_{j=0}^{k} \left(\left\| r^{j+1} \right\|_2^2 + \left\| B(z^{j+1} - z^j) \right\|_2^2 \right),$$

which implies that

$$\rho \sum_{j=0}^{k} \left(\left\| r^{j+1} \right\|_2^2 + \left\| B(z^{j+1} - z^j) \right\|_2^2 \right) \leq V_0 - V^{k+1} \leq V^0, \tag{B}$$

since $V^{k+1} \leq V^0$.

Step 2. Prove that the sequence (r^k) converges to 0, and that the sequences (Ax^{k+1}) and (Bz^{k+1}) also converge.

Inequality (B) implies that the series $\sum_{k=1}^{\infty} r^k$ and $\sum_{k=0}^{\infty} B(z^{k+1} - z^k)$ converge absolutely. In particular, the sequence (r^k) converges to 0.

The nth partial sum of the series $\sum_{k=0}^{\infty} B(z^{k+1} - z^k)$ is

$$\sum_{k=0}^{n} B(z^{k+1} - z^k) = B(z^{n+1} - z^0),$$

and since the series $\sum_{k=0}^{\infty} B(z^{k+1} - z^k)$ converges, we deduce that the sequence (Bz^{k+1}) converges. Since $Ax^{k+1} + Bz^{k+1} - c = r^{k+1}$, the convergence of (r^{k+1}) and (Bz^{k+1}) implies that the sequence (Ax^{k+1}) also converges.

Step 3. Prove that the sequences (x^{k+1}) and (z^{k+1}) converge. By Assumption (2), the matrices $A^\top A$ and $B^\top B$ are invertible, so multiplying each vector Ax^{k+1} by $(A^\top A)^{-1}A^\top$, if the sequence (Ax^{k+1}) converges to u, then the sequence (x^{k+1}) converges to $(A^\top A)^{-1}A^\top u$. Siimilarly, if the sequence (Bz^{k+1}) converges to v, then the sequence (z^{k+1}) converges to $(B^\top B)^{-1}B^\top v$.

Step 4. Prove that the sequence (λ^k) converges.

Recall that

$$\lambda^{k+1} = \lambda^k + \rho r^{k+1}.$$

It follows by induction that

$$\lambda^{k+p} = \lambda^k + \rho(r^{k+1} + \cdots + \rho^{k+p}), \quad p \geq 2.$$

As a consequence, we get

$$\left\| \lambda^{k+p} - \lambda^k \right\| \leq \rho(\left\| r^{k+1} \right\| + \cdots + \left\| r^{k+p} \right\|).$$

Since the series $\sum_{k=1}^{\infty} \left\| r^k \right\|$ converges, the partial sums form a Cauchy sequence, and this immediately implies that for any $\epsilon > 0$ we can find $N > 0$ such that

$$\rho(\left\| r^{k+1} \right\| + \cdots + \left\| r^{k+p} \right\|) < \epsilon, \quad \text{for all } k, p + k \geq N,$$

so the sequence (λ^k) is also a Cauchy sequence, thus it converges.

Step 5. Prove that the sequence (p^k) converges to p^*.

For this, we need two more inequalities. Following Boyd et al. [28], we need to prove that

$$p^{k+1} - p^* \leq -(\lambda^{k+1})^\top r^{k+1} - \rho(B(z^{k+1} - z^k))^\top (-r^{k+1} + B(z^{k+1} - z^*)) \tag{A2}$$

and

$$p^* - p^{k+1} \leq (\lambda^*)^\top r^{k+1}. \tag{A3}$$

Since we proved that the sequence (r^k) and $B(z^{k+1} - z^k)$ converge to 0, and that the sequence (λ^{k+1}) converges, from

$$(\lambda^{k+1})^\top r^{k+1} + \rho(B(z^{k+1} - z^k))^\top (-r^{k+1} + B(z^{k+1} - z^*)) \leq p^* - p^{k+1} \leq (\lambda^*)^\top r^{k+1},$$

we deduce that in the limit, p^{k+1} converges to p^*.

Step 6. Prove (A3).

Since (x^*, y^*, λ^*) is a saddle point, we have

$$L_0(x^*, z^*, \lambda^*) \leq L_0(x^{k+1}, z^{k+1}, \lambda^*).$$

Since $Ax^* + Bz^* = c$, we have $L_0(x^*, z^*, \lambda^*) = p^*$, and since $p^{k+1} = f(x^{k+1}) + g(z^{k+1})$, we have

$$L_0(x^{k+1}, z^{k+1}, \lambda^*) = p^{k+1} + (\lambda^*)^\top r^{k+1},$$

so we obtain

$$p^* \leq p^{k+1} + (\lambda^*)^\top r^{k+1},$$

which yields (A3).

Step 7. Prove (A2).

By Proposition 51.34, z^{k+1} minimizes $L_\rho(x^{k+1}, z, \lambda^k)$ iff

$$\begin{aligned} 0 &\in \partial g(z^{k+1}) + B^\top \lambda^k + \rho B^\top (Ax^{k+1} + Bz^{k+1} - c) \\ &= \partial g(z^{k+1}) + B^\top \lambda^k + \rho B^\top r^{k+1} \\ &= \partial g(z^{k+1}) + B^\top \lambda^{k+1}, \end{aligned}$$

since $r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$ and $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$.

In summary, we have

$$0 \in \partial g(z^{k+1}) + B^\top \lambda^{k+1}, \tag{$\dagger_1$}$$

which shows that z^{k+1} minimizes the function

$$z \mapsto g(z) + (\lambda^{k+1})^\top Bz.$$

Consequently, we have

$$g(z^{k+1}) + (\lambda^{k+1})^\top Bz^{k+1} \leq g(z^*) + (\lambda^{k+1})^\top Bz^*. \tag{B1}$$

Similarly, x^{k+1} minimizes $L_\rho(x, z^k, \lambda^k)$ iff

$$\begin{aligned} 0 &\in \partial f(x^{k+1}) + A^\top \lambda^k + \rho A^\top (Ax^{k+1} + Bz^k - c) \\ &= \partial f(x^{k+1}) + A^\top (\lambda^k + \rho r^{k+1} + \rho B(z^k - z^{k+1})) \\ &= \partial f(x^{k+1}) + A^\top \lambda^{k+1} + \rho A^\top B(z^k - z^{k+1}) \end{aligned}$$

since $r^{k+1} - Bz^{k+1} = Ax^{k+1} - c$ and $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + Bz^{k+1} - c) = \lambda^k + \rho r^{k+1}$.

Equivalently, the above derivation shows that

$$0 \in \partial f(x^{k+1}) + A^\top(\lambda^{k+1} - \rho B(z^{k+1} - z^k)), \tag{$\dagger_2$}$$

which shows that x^{k+1} minimizes the function

$$x \mapsto f(x) + (\lambda^{k+1} - \rho B(z^{k+1} - z^k))^\top Ax.$$

Consequently, we have

$$f(x^{k+1}) + (\lambda^{k+1} - \rho B(z^{k+1} - z^k))^\top Ax^{k+1} \leq f(x^*) + (\lambda^{k+1} - \rho B(z^{k+1} - z^k))^\top Ax^*. \tag{B2}$$

Adding up Inequalities (B1) and (B2), using the equation $Ax^* + Bz^* = c$, and rearranging, we obtain inequality (A2).

Step 8. Prove that $(x^k), (z^k)$, and (λ^k) converge to optimal solutions.

Recall that (r^k) converges to 0, and that (x^k), (z^k), and (λ^k) converge to limits $\widetilde{x}$, $\widetilde{z}$, and $\widetilde{\lambda}$. Since $r^k = Ax^k + Bz^k - c$, in the limit, we have

$$A\widetilde{x} + B\widetilde{z} - c = 0. \tag{$*_1$}$$

Using $(\dagger_1)$, in the limit, we obtain

$$0 \in \partial g(\widetilde{z}) + B^\top \widetilde{\lambda}. \tag{$*_2$}$$

Since $(B(z^{k+1} - z^k))$ converges to 0, using $(\dagger_2)$, in the limit, we obtain

$$0 \in \partial f(\widetilde{x}) + A^\top \widetilde{\lambda}. \tag{$*_3$}$$

From $(*_2)$ and $(*_3)$, we obtain

$$0 \in \partial f(\widetilde{x}) + \partial g(\widetilde{z}) + A^\top \widetilde{\lambda} + B^\top \widetilde{\lambda}. \tag{$*_4$}$$

But $(*_1)$ and $(*_4)$ are exactly the KKT equations, and by Theorem 51.41, we conclude that $\widetilde{x}, \widetilde{z}, \widetilde{\lambda}$ are optimal solutions.

Step 9. Prove (A1). This is the most tedious step of the proof. We begin by adding up (A2) and (A3), and then perform quite a bit or rewriting and manipulation. The complete derivation can be found in Boyd et al. [28]. □

Remarks:

(1) In view of Theorem 51.42, we could replace Assumption (3) by the slightly stronger assumptions that the optimum value of our program is finite and that the constraints are qualified. Since the constraints are affine, this means that there is some feasible solution in $\mathbf{relint}(\mathrm{dom}(f)) \cap \mathbf{relint}(\mathrm{dom}(g))$. These assumptions are more practical than Assumption (3).

(2) Actually, Assumption (3) implies Assumption (2). Indeed, we know from Theorem 51.41 that the existence of a saddle point implies that our program has a finite optimal solution. However, if either $A^\top A$ or $B^\top B$ is not invertible, then Program (P) may not have a finite optimal solution, as shown by the following counterexample.

Example 52.5. Let

$$f(x, y) = x, \quad g(z) = 0, \quad y - z = 0.$$

Then

$$L_\rho(x, y, z, \lambda) = x + \lambda(y - z) + (\rho/2)(y - z)^2,$$

but minimizing over (x, y) with z held constant yields $-\infty$, which implies that the above program has no finite optimal solution. See Figure 52.4.

The problem is that

$$A = \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 \end{pmatrix},$$

but

$$A^\top A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

is not invertible.

(3) Proving (A1), (A2), (A3), and the convergence of (r^k) to 0 and of (p^k) to p^* does not require Assumption (2). The proof, using the ingeneous Inequality (A1) (and (B)) is the proof given in Boyd et al. [28]. We were also able to prove that (λ^k), (Ax^k) and (Bz^k) converge without Assumption (2), but to prove that (x^k), (y^k), and (λ^k) converge to optimal solutions, we had to use Assumption (2).

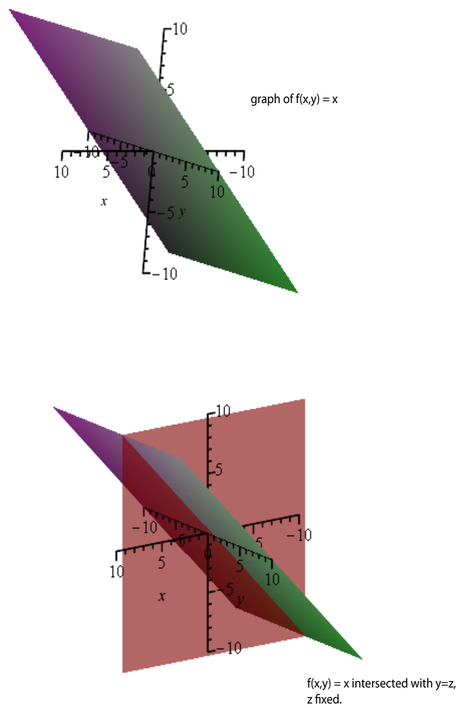


Figure 52.4: A graphical representation of the Example 52.5. This is an illustration of the x minimization step when z is held fixed. Since the intersection of the two planes is an unbounded line, we "see" that minimizing over x yields $-\infty$.

(4) Bertsekas discusses ADMM in [17], Sections 2.2 and 5.4. His formulation of ADMM is slightly different, namely

$$\begin{aligned} \text{minimize} \quad & f(x) + g(z) \\ \text{subject to} \quad & Ax = z. \end{aligned}$$

Bertsekas states a convergence result for this version of ADMM under the hypotheses that either $\mathrm{dom}(f)$ is compact or that $A^\top A$ is invertible, and that a saddle point exists; see Proposition 5.4.1. The proof is given in Bertsekas [20], Section 3.4, Proposition 4.2. It appears that the proof makes use of gradients, so it is not clear that it applies in the more general case where f and g are not differentiable.

(5) Versions of ADMM are discussed in Gabay [69] (Sections 4 and 5). They are more general than the version discussed here. Some convergence proofs are given, but because Gabay's framework is more general, it is not clear that they apply to our setting. Also, these proofs rely on earlier result by Lions and Mercier, which makes the comparison difficult.

(5) Assumption (2) does not imply that the system $Ax + Bz = c$ has any solution. For example, if

$$A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

the system

$$\begin{aligned} x - z &= 1 \\ x - z &= 0 \end{aligned}$$

has no solution. However, since Assumption (3) implies that the program has an optimal solution, it implies that c belongs to the column space of the $p \times (n + m)$ matrix $\begin{pmatrix} A & B \end{pmatrix}$.

Here is an example where ADMM diverges for a problem whose optimum value is $-\infty$.

Example 52.6. Consider the problem given by

$$f(x) = x, \quad g(z) = 0, \quad x - z = 0.$$

Since $f(x) + g(z) = x$, and $x = z$, the variable x is unconstrained and the above function goes to $-\infty$ when x goes to $-\infty$. The augmented Lagrangian is

$$\begin{aligned} L_\rho(x, z, \lambda) &= x + \lambda(x - z) + \frac{\rho}{2}(x - z)^2 \\ &= \frac{\rho}{2}x^2 - \rho xz + \frac{\rho}{2}z^2 + x + \lambda x - \lambda z. \end{aligned}$$

The matrix

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

is singular and $L_\rho(x, z, \lambda)$ goes to $-\infty$ in when $(x, z) = t(1, 1)$ and t goes to $-\infty$. The ADMM steps are:

$$\begin{aligned} x^{k+1} &= z^k - \frac{1}{\rho}\lambda^k - \frac{1}{\rho} \\ z^{k+1} &= x^{k+1} + \frac{1}{\rho}\lambda^k \\ \lambda^{k+1} &= \lambda^k + \rho(x^{k+1} - z^{k+1}), \end{aligned}$$

and these equations hold for all $k \geq 0$. From the last two equations we deduce that

$$\lambda^{k+1} = \lambda^k + \rho(x^{k+1} - z^{k+1}) = \lambda^k + \rho(-\frac{1}{\rho}\lambda^k) = 0, \quad \text{for all } k \geq 0,$$

so

$$z^{k+2} = x^{k+2} + \frac{1}{\rho}\lambda^{k+1} = x^{k+2}, \quad \text{for all } k \geq 0.$$

Consequently we find that

$$x^{k+3} = z^{k+2} + \frac{1}{\rho}\lambda^{k+2} - \frac{1}{\rho} = x^{k+2} - \frac{1}{\rho}.$$

By induction, we obtain

$$x^{k+3} = x^2 - \frac{k+1}{\rho}, \quad \text{for all } k \geq 0,$$

which shows that x^{k+3} goes to $-\infty$ when k goes to infinity, and since $x^{k+2} = z^{k+2}$, similarly z^{k+3} goes to $-\infty$ when k goes to infinity.

52.5 Stopping Criteria

Going back to Inequality (A2),

$$p^{k+1} - p^* \leq -(\lambda^{k+1})^\top r^{k+1} - \rho(B(z^{k+1} - z^k))^\top(-r^{k+1} + B(z^{k+1} - z^*)), \tag{A2}$$

using the fact that $Ax^* + Bz^* - c = 0$ and $r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$, we have

$$\begin{aligned} -r^{k+1} + B(z^{k+1} - z^*) &= -Ax^{k+1} - Bz^{k+1} + c + B(z^{k+1} - z^*) \\ &= -Ax^{k+1} + c - Bz^* \\ &= -Ax^{k+1} + Ax^* = -A(x^{k+1} - x^*), \end{aligned}$$

so (A2) can be rewritten as

$$p^{k+1} - p^* \leq -(\lambda^{k+1})^\top r^{k+1} + \rho(B(z^{k+1} - z^k))^\top A(x^{k+1} - x^*),$$

or equivalently as

$$p^{k+1} - p^* \leq -(\lambda^{k+1})^\top r^{k+1} + (x^{k+1} - x^*)^\top \rho A^\top B(z^{k+1} - z^k). \tag{s_1}$$

We define the *dual residual* as

$$s^{k+1} = \rho A^\top B(z^{k+1} - z^k),$$

the quantity $r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$ being the *primal residual*. Then (s_1) can be written as

$$p^{k+1} - p^* \leq -(\lambda^{k+1})^\top r^{k+1} + (x^{k+1} - x^*)^\top s^{k+1}. \tag{s}$$

Inequality (s) shows that when the residuals r^k and s^k are small, then p^k is close to p^* from below. Since x^* is unknown, we can't use this inequality, but if we have a guess that $\left\|x^k - x^*\right\| \leq d$, then using Cauchy–Schwarz we obtain

$$p^{k+1} - p^* \leq \left\|\lambda^{k+1}\right\| \left\|r^{k+1}\right\| + d\left\|s^{k+1}\right\|.$$

The above suggests that a reasonable termination criterion is that $\|r^k\|$ and $\|s^k\|$ should be small, namely that

$$\|r^k\| \leq \epsilon^{\mathrm{pri}} \quad \text{and} \quad \|s^k\| \leq \epsilon^{\mathrm{dual}},$$

for some chosen feasibility tolerances ϵ^{pri} and ϵ^{dual}. Further discussion for choosing these parameters can be found in Boyd et al. [28] (Section 3.3.1).

Various extensions and variations of ADMM are discussed in Boyd et al. [28] (Section 3.4). In order to accelerate convergence of the method, one may choose a different ρ at each step (say ρ^k), although proving the convergence of such a method may be difficult. If we assume that ρ^k becomes constant after a number of iterations, then the proof that we gave still applies. A simple scheme is this:

$$\rho^{k+1} = \begin{cases} \tau^{\mathrm{incr}}\rho^k & \text{if } \|r^k\| > \mu \|s^k\| \\ \rho^k/\tau^{\mathrm{decr}} & \text{if } \|s^k\| > \mu \|r^k\| \\ \rho^k & \text{otherwise,} \end{cases}$$

where $\tau^{\mathrm{incr}} > 1, \tau^{\mathrm{decr}} > 1$, and $\mu > 1$ are some chosen parameters. Again, we refer the interested reader to Boyd et al. [28] (Section 3.4).

52.6 Some Applications of ADMM

Structure in f, g, A, and B can often be exploited to yield more efficient methods for performing the x-update and the z-update. We focus on the x-update, but the discussion applies just as well to the z-update. Since z and λ are held constant during minimization over x, it is more convenient to use the scaled form of ADMM. Recall that

$$x^{k+1} = \arg\min_x \left(f(x) + (\rho/2) \left\|Ax + Bz^k - c + u^k\right\|_2^2 \right)$$

(here we use u instead of μ), so we can express the x-update step as

$$x^+ = \arg\min_x \left(f(x) + (\rho/2) \|Ax - v\|_2^2 \right),$$

with $v = -Bz^k + c - u^k$.

Example 52.7. A first simplification arises when $A = I$, in which case the x-update is

$$x^+ = \arg\min_x \left(f(x) + (\rho/2) \|x - v\|_2^2 \right) = \mathbf{prox}_{f,\rho}(v).$$

The map $v \mapsto \mathbf{prox}_{f,\rho}(v)$ is known as the *proximity operator of f with penalty ρ*. The above minimization is generally referred to as *proximal minimization*.

Example 52.8. When the function f is simple enough, the proximity operator can be computed analytically. This is the case in particular when $f = I_C$, the indicator function of a nonempty closed convex set C. In this case, it is easy to see that

$$x^+ = \arg\min_x \left(I_C(x) + (\rho/2)\|x - v\|_2^2\right) = \Pi_C(v),$$

the orthogonal projection of v onto C. In the special case where $C = \mathbb{R}^n_+$ (the first orthant), then

$$x^+ = (v)_+,$$

the vector obtained by setting the negative components of v to zero.

Example 52.9. A second case where simplifications arise is the case where f is a convex quadratic functional of the form

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r,$$

where P is an $n \times n$ symmetric positive semidefinite matrix, $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$. In this case the gradient of the map

$$x \mapsto f(x) + (\rho/2)\|Ax - v\|_2^2 = \frac{1}{2}x^\top Px + q^\top x + r + \frac{\rho}{2}x^\top(A^\top A)x - \rho x^\top A^\top v + \frac{\rho}{2}v^\top v$$

is given by

$$(P + \rho A^\top A)x + q - \rho A^\top v,$$

and since A has rank n, the matrix $A^\top A$ is symmetric positive definite, so we get

$$x^+ = (P + \rho A^\top A))^{-1}(\rho A^\top v - q).$$

Methods from numerical linear algebra can be used so compute x^+ fairly efficiently; see Boyd et al. [28] (Section 4).

Example 52.10. A third case where simplifications arise is the variation of the previous case where f is a convex quadratic functional of the form

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r,$$

except that f is constrained by equality constraints $Cx = b$, as in Section 50.4, which means that $\mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid Cx = b\}$, and $A = I$. The x-minimization step consists in minimizing the function

$$J(x) = \frac{1}{2}x^\top Px + q^\top x + r + \frac{\rho}{2}x^\top x - \rho x^\top v + \frac{\rho}{2}v^\top v$$

subject to the constraint

$$Cx = b,$$

so by the results of Section 50.4, x^+ is a component of the solution of the KKT-system

$$\begin{pmatrix} P + \rho I & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} x^+ \\ y \end{pmatrix} = \begin{pmatrix} -q + \rho v \\ b \end{pmatrix}.$$

The matrix $P + \rho I$ is symmetric positive definite, so the KKT-matrix is invertible.

We can now describe how ADMM is used to solve two common problems of convex optimization.

(1) *Minimization of a proper closed convex function f over a closed convex set C in $\mathbb{R}^n$.* This is the following problem

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in C, \end{aligned}$$

which can be rewritten in ADMM form as

$$\begin{aligned} &\text{minimize} && f(x) + I_C(z) \\ &\text{subject to} && x - z = 0. \end{aligned}$$

Using the scaled dual variable $u = \lambda/\rho$, the augmented Lagrangian is

$$L_\rho(x, z, u) = f(x) + I_C(z) + \frac{\rho}{2} \|x - z + u\|_2^2 - \frac{\rho}{2} \|u\|^2.$$

In view of Example 52.8, the scaled form of ADMM for this problem is

$$\begin{aligned} x^{k+1} &= \arg\min_x \left(f(x) + (\rho/2) \left\|x - z^k + u^k\right\|_2^2 \right) \\ z^{k+1} &= \Pi_C(x^{k+1} + u^k) \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}. \end{aligned}$$

The x-update involves evaluating a proximal operator. Note that the function f need not be differentiable. Of course, these minimizations depend on having efficient computational procedures for the proximal operator and the projection operator.

(2) *Quadratic Programming, Version 1.* Here the problem is

$$\begin{aligned} &\text{minimize} && \frac{1}{2} x^\top P x + q^\top x + r \\ &\text{subject to} && Ax = b,\ x \geq 0, \end{aligned}$$

where P is an $n \times n$ symmetric positive semidefinite matrix, $q \in \mathbb{R}^n$, $r \in \mathbb{R}$, and A is an $m \times n$ matrix of rank m.

The above program is converted in ADMM form as follows:

$$\begin{aligned} &\text{minimize} && f(x) + g(z) \\ &\text{subject to} && x - z = 0, \end{aligned}$$

with

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r, \quad \mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid Ax = b\},$$

and

$$g = I_{\mathbb{R}^n_+},$$

the indicator function of the positive orthant $\mathbb{R}^n_+$. In view of Example 52.8 and Example 52.10, the scaled form of ADMM consists of the following steps:

$$\begin{aligned} x^{k+1} &= \arg\min_x \left(f(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \right) \\ z^{k+1} &= (x^{k+1} + u^k)_+ \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}. \end{aligned}$$

The x-update involves solving the KKT equations

$$\begin{pmatrix} P + \rho I & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y \end{pmatrix} = \begin{pmatrix} -q + \rho(z^k - u^k) \\ b \end{pmatrix}.$$

This is an important example because it provides one of the best methods for solving quadratic problems, in particular, the SVM problems discussed in Chapter 54.

(3) *Quadratic Programming, Version 2*. This problem is similar to the previous one, except that the variable $x \in \mathbb{R}^n$ is *not restricted* to be nonnegative. The problem is

$$\begin{aligned} &\text{minimize} && \frac{1}{2}x^\top Px + q^\top x + r \\ &\text{subject to} && Ax = b, \end{aligned}$$

where P is an $n \times n$ symmetric positive semidefinite matrix, $q \in \mathbb{R}^n$, $r \in \mathbb{R}$, and A is an $m \times n$ matrix of rank m.

The above program is converted in ADMM form as follows:

$$\begin{aligned} &\text{minimize} && f(x) + g(z) \\ &\text{subject to} && x - z = 0, \end{aligned}$$

with

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r, \quad \mathrm{dom}(f) = \{x \in \mathbb{R}^n \mid Ax = b\},$$

and

$$g = 1,$$

the constant function which is the indicator function of the convex set $C = \mathbb{R}^n$. In view of Example 52.8 and (1), since $\Pi_{\mathbb{R}^n}(x^{k+1} + u^k) = x^{k+1} + u^k$, the scaled form of ADMM consists of the following steps:

$$\begin{aligned}
x^{k+1} &= \arg\min_x \left(f(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \right) \\
z^{k+1} &= x^{k+1} + u^k \\
u^{k+1} &= u^k + x^{k+1} - z^{k+1} = 0,
\end{aligned}$$

for all $k \geq 0$, so

$$\begin{aligned}
u^k &= 0 \\
z^{k+1} &= x^{k+1}
\end{aligned}$$

for all $k \geq 1$. Consequently we have

$$\begin{aligned}
x^{k+1} &= \arg\min_x \left(f(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \right) \\
z^{k+1} &= x^{k+1} + u^k \\
u^1 &= 0,
\end{aligned}$$

for $k = 0, 1$, and for $k \geq 2$ we have $u^k = 0$ and $z^k = x^k$, with

$$x^{k+1} = \arg\min_x \left(f(x) + (\rho/2) \left\| x - x^k \right\|_2^2 \right).$$

As before, the x-update involves solving the KKT equations

$$\begin{pmatrix} P + \rho I & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y \end{pmatrix} = \begin{pmatrix} -q + \rho(z^k - u^k) \\ b \end{pmatrix},$$

with $u^k = 0$ if $k \geq 1$ and $z^k = x^k$ if $k \geq 2$.

We programmed the above method in `Matlab` as the function `qsolve1`, see Appendix B, Section B.1. Here are two examples.

Example 52.11. Consider the quadratic program for which

$$P_1 = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \qquad q_1 = -\begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \qquad b_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We see immediately that the constraints

$$\begin{aligned} x + y - z &= 0 \\ x - y - z &= 0 \end{aligned}$$

imply that $z = x$ and $y = 0$. Then it is easy using calculus to find that the unique minimum is given by $(x, y, z) = (1, 0, 1)$. Running `qsolve1` on P_1, q_1, A_1, b_1 with $\rho = 10$, $tolr = tols = 10^{-12}$ and $iternum = 10000$, we find that after 83 iterations the primal and the dual residuals are less than 10^{-12}, and we get

$$(x, y, z) = (1.000000000000149, 0.000000000000000, 1.000000000000148).$$

Example 52.12. Consider the quadratic program for which

$$P_2 = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix} \qquad q_2 = -\begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 1 & 1 & -1 & 0 \\ 1 & -1 & -1 & 0 \end{pmatrix} \qquad b_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Again, we see immediately that the constraints imply that $z = x$ and $y = 0$. Then it is easy using calculus to find that the unique minimum is given by $(x, y, z) = (28/31, 0, 28/31, 24/31)$. Running `qsolve1` on P_2, q_2, A_2, b_2 with $\rho = 10$, $tolr = tols = 10^{-12}$ and $iternum = 10000$, we find that after 95 iterations the primal and the dual residuals are less than 10^{-12}, and we get

$$(x, y, z, t) = (0.903225806451495, 0.000000000000000, 0.903225806451495, 0.774193548387264),$$

which agrees with the answer found earlier up to 11 decimals.

As an illustration of the wide range of applications of ADMM we show in the next section how to solve the hard margin SVM (SVM_{h2}) discussed in Section 50.6.

52.7 Solving Hard Margin (SVM_{h2}) Using ADMM

Recall that we would like to solve the following optimization problem (see Section 50.6):

Hard margin SVM (SVM_{h2}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 \qquad j = 1, \ldots, q. \end{aligned}$$

The margin is $\delta = 1/\|w\|$. The separating hyperplane $H_{w,b}$ is the hyperplane of equation $w^\top x - b = 0$, and the margin hyperplanes are the hyperplanes $H_{w,b+1}$ (the blue hyperplane) of equation $w^\top x - b - 1 = 0$ and $H_{w,b-1}$ (the red hyperplane) of equation $w^\top x - b + 1 = 0$. The dual program derived in Section 50.10 is the following program:

Dual of the Hard margin SVM (SVM_{h2}):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{1}_{p+q}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0\\
&\qquad \lambda \geq 0,\ \mu \geq 0,
\end{aligned}
$$

where X is the $n \times (p+q)$ matrix given by

$$X = \begin{pmatrix}-u_1 & \cdots & -u_p & v_1 & \cdots & v_q\end{pmatrix}.$$

Then w is determined as follows:

$$w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j.$$

To solve the dual using ADMM we need to determine the matrices P, q A and c as in Section 52.6(2). We renamed b as c to avoid a clash since b is already used. We have

$$P = X^\top X, \quad q = -\mathbf{1}_{p+q},$$

and since the only constraint is

$$\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0,$$

the matrix A is the $1 \times (p+q)$ row vector

$$A = \begin{pmatrix}\mathbf{1}_p^\top & -\mathbf{1}_q^\top\end{pmatrix},$$

and the right-hand side c is the scalar

$$c = 0.$$

Obviously the matrix A has rank 1. We obtain b using any i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$. Since the corresponding constraints are active, we have

$$w^\top u_{i_0} - b = 1, \quad -w^\top v_{j_0} + b = 1,$$

so we obtain

$$b = w^\top (u_{i_0} + v_{j_0})/2.$$

For improved numerical stability, we can average over the sets of indices defined as $I_{\lambda>0} = \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\}$ and $I_{\mu>0} = \{j \in \{1, \ldots, q\} \mid \mu_j > 0\}$. We obtain

$$b = w^\top \left(\Bigg(\sum_{i \in I_{\lambda>0}} u_i \Bigg) / |I_{\lambda>0}| + \Bigg(\sum_{j \in I_{\mu>0}} v_j \Bigg) / |I_{\mu>0}| \right) /2.$$

The `Matlab` programs implementing the above method are given in Appendix B, Section B.1. This should convince the reader that there is very little gap between theory and practice, although it is quite consuming to tune the tolerance parameters needed to deal with floating-point arithmetric.

Figure 52.5 shows the result of running the `Matlab` program implementing the above method using ADMM on two sets of points of 50 points each generated at random using the following `Matlab` code.

```
u14 = 10.1*randn(2,50)+18;
v14 = -10.1*randn(2,50)-18;
```

The function `SVMhard2` is called with $\rho = 10$ as follows

```
[lamb,mu,w] = SVMhard2(10,u14,v14)
```

and produces the output shown in Figure 52.5. Observe that there is one blue support vector and two red support vectors.

52.8 Applications of ADMM to ℓ^1-Norm Problems

Another important application of ADMM is to ℓ^1-norm minimization problems, especially lasso minimization, discussed below and in Section 55.4. This involves the special case of ADMM where $f(x) = \tau \left\|x\right\|_1$ and $A = I$. In particular, in the one-dimensional case, we need to solve the minimization problem: find

$$x^* = \arg\min_x \left(\tau |x| + (\rho/2)(x - v)^2 \right),$$

with $x, v \in \mathbb{R}$, and $\rho, \tau > 0$. Let $c = \tau/\rho$ and write

$$f(x) = \frac{\tau}{2c} \left(2c|x| + (x - v)^2 \right).$$

Minimizing f over x is equivalent to minimizing

$$g(x) = 2c|x| + (x - v)^2 = 2c|x| + x^2 - 2xv + v^2,$$

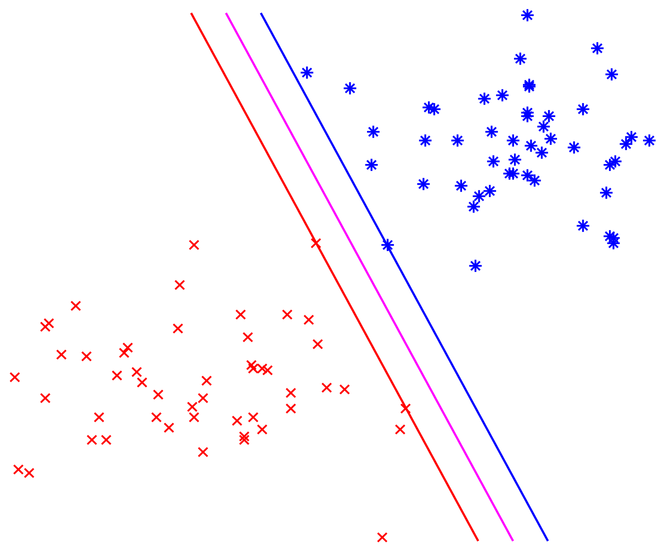


Figure 52.5: An example of hard margin SVM.

which is equivalent to minimizing

$$h(x) = x^2 + 2(c|x| - xv)$$

over x. If $x \geq 0$, then

$$h(x) = x^2 + 2(cx - xv) = x^2 + 2(c - v)x = (x - (v - c))^2 - (v - c)^2.$$

If $v - c > 0$, that is, $v > c$, since $x \geq 0$, the function $x \mapsto (x - (v - c))^2$ has a minimum for $x = v - c > 0$, else if $v - c \leq 0$, then the function $x \mapsto (x - (v - c))^2$ has a minimum for $x = 0$.

If $x \leq 0$, then

$$h(x) = x^2 + 2(-cx - xv) = x^2 - 2(c + v)x = (x - (v + c))^2 - (v + c)^2.$$

if $v + c < 0$, that is, $v < -c$, since $x \leq 0$, the function $x \mapsto (x - (v + c))^2$ has a minimum for $x = v + c$, else if $v + c \geq 0$, then the function $x \mapsto (x - (v + c))^2$ has a minimum for $x = 0$.

In summary, $\inf_x h(x)$ is the function of v given by

$$S_c(v) = \begin{cases} v - c & \text{if } v > c \\ 0 & \text{if } |v| \leq c \\ v + c & \text{if } v < -c. \end{cases}$$

The function S_c is known as a *soft thresholding operator*. The graph of S_c shown in Figure 52.6.

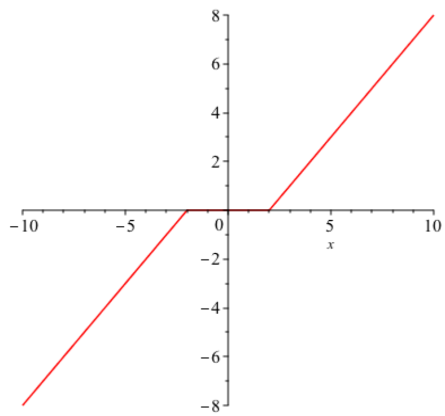


Figure 52.6: The graph of S_c (when $c = 2$).

One can check that

$$S_c(v) = (v - c)_+ - (-v - c)_+,$$

and also

$$S_c(v) = (1 - c/|v|)_+ v, \quad v \neq 0,$$

which shows that S_c is a *shrinkage operator* (it moves a point toward zero).

The operator S_c is extended to vectors in $\mathbb{R}^n$ component wise, that is, if $x = (x_1, \ldots, x_n)$, then

$$S_c(x) = (S_c(x_1), \ldots, S_c(x_n)).$$

We now consider several ℓ^1-norm problems.

(1) *Least absolute deviation.*

This is the problem of minimizing $\|Ax - b\|_1$, rather than $\|Ax - b\|_2$. Least absolute deviation is more robust than least squares fit because it deals better with outliers. The problem can be formulated in ADMM form as follows:

$$\begin{aligned} &\text{minimize} && \|z\|_1 \\ &\text{subject to} && Ax - z = b, \end{aligned}$$

with $f = 0$ and $g = \| \, \|_1$. As usual, we assume that A is an $m \times n$ matrix of rank n, so that $A^\top A$ is invertible. ADMM (in scaled form) can be expressed as

$$\begin{aligned} x^{k+1} &= (A^\top A)^{-1} A^\top (b + z^k - u^k) \\ z^{k+1} &= S_{1/\rho}(Ax^{k+1} - b + u^k) \\ u^{k+1} &= u^k + Ax^{k+1} - z^{k+1} - b. \end{aligned}$$

(2) *Basis pursuit.*

This is the following minimization problem:

$$\begin{aligned} &\text{minimize} && \|x\|_1 \\ &\text{subject to} && Ax = b, \end{aligned}$$

where A is an $m \times n$ matrix of rank $m < n$, and $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. The problem is to find a sparse solution to an underdetermined linear system, which means a solution x with many zero coordinates. This problem plays a central role in compressed sensing and statistical signal processing.

Basis pursuit can be expressed in ADMM form as the problem

$$\begin{aligned} &\text{minimize} && I_C(x) + \|z\|_1 \\ &\text{subject to} && x - z = 0, \end{aligned}$$

with $C = \{x \in \mathbb{R}^n \mid Ax = b\}$. It is easy to see that the ADMM procedure (in scaled form) is

$$\begin{aligned} x^{k+1} &= \Pi_C(z^k - u^k) \\ z^{k+1} &= S_{1/\rho}(x^{k+1} + u^k) \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}, \end{aligned}$$

where Π_C is the orthogonal projection onto the subspace C. In fact, it is not hard to show that

$$x^{k+1} = (I - A^\top (AA^\top)^{-1}A)(z^k - u^k) + A^\top (AA^\top)^{-1}b.$$

In some sense, an ℓ^1-minimization problem is reduced to a sequence of ℓ^2-norm problems. There are ways of improving the efficiency of the method; see Boyd et al. [28] (Section 6.2)

(3) *General ℓ^1-regularized loss minimization.*

This is the following minimization problem:

$$\text{minimize} \quad l(x) + \tau \|x\|_1,$$

where l is any proper closed and convex loss function, and $\tau > 0$. We convert the problem to the ADMM problem:

$$\begin{aligned} &\text{minimize} && l(x) + \tau \|z\|_1 \\ &\text{subject to} && x - z = 0. \end{aligned}$$

The ADMM procedure (in scaled form) is

$$\begin{aligned}
x^{k+1} &= \arg\min_x \Big(l(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \Big) \\
z^{k+1} &= S_{\tau/\rho}(x^{k+1} + u^k) \\
u^{k+1} &= u^k + x^{k+1} - z^{k+1}.
\end{aligned}$$

The x-update is a proximal operator evaluation. In general, one needs to apply a numerical procedure to compute x^{k+1}, for example, a version of Newton's method. The special case where $l(x) = (1/2) \|Ax - b\|_2^2$ is particularly important.

(4) *Lasso regularization.*

This is the following minimization problem:

$$\text{minimize} \quad (1/2) \|Ax - b\|_2^2 + \tau \|x\|_1 .$$

This is a linear regression with the regularizing term $\tau \|x\|_1$ instead of $\tau \|x\|_2$, to encourage a sparse solution. This method was first proposed by Tibshirani around 1996, under the name *lasso*, which stands for "least absolute selection and shrinkage operator." This method is also known as *ℓ^1-regularized regression*, but this is not as cute as "lasso," which is used predominantly. This method is discussed extensively in Hastie, Tibshirani, and Wainwright [89].

The lasso minimization is converted to the following problem in ADMM form:

$$\begin{aligned}
&\text{minimize} && \frac{1}{2} \|Ax - b\|_2^2 + \tau \|z\|_1 \\
&\text{subject to} && x - z = 0.
\end{aligned}$$

Then the ADMM procedure (in scaled form) is

$$\begin{aligned}
x^{k+1} &= (A^\top A + \rho I)^{-1}(A^\top b + \rho(z^k - u^k)) \\
z^{k+1} &= S_{\tau/\rho}(x^{k+1} + u^k) \\
u^{k+1} &= u^k + x^{k+1} - z^{k+1}.
\end{aligned}$$

Since $\rho > 0$, the matrix $A^\top A + \rho I$ is symmetric positive definite. Note that the x-update looks like a *ridge regression step* (see Section 55.1).

There are various generalizations of lasso.

(5) *Generalized Lasso regularization.*

This is the following minimization problem:

$$\text{minimize} \quad (1/2) \|Ax - b\|_2^2 + \tau \|Fx\|_1 ,$$

where A is an $m \times n$ matrix, F is a $p \times n$ matrix, and either A has rank n or F has rank n. This problem is converted to the ADMM problem

$$\begin{aligned} &\text{minimize} \quad \|Ax - b\|_2^2 + \tau \|z\|_1 \\ &\text{subject to} \quad Fx - z = 0, \end{aligned}$$

and the corresponding ADMM procedure (in scaled form) is

$$\begin{aligned} x^{k+1} &= (A^\top A + \rho F^\top F)^{-1}(A^\top b + \rho F^\top (z^k - u^k)) \\ z^{k+1} &= S_{\tau/\rho}(Fx^{k+1} + u^k) \\ u^{k+1} &= u^k + Fx^{k+1} - z^{k+1}. \end{aligned}$$

(6) *Group Lasso.*

This a generalization of (3). Here we assume that x is split as $x = (x_1, \ldots, x_N)$, with $x_i \in \mathbb{R}^{n_i}$ and $n_1 + \cdots + x_N = n$, and the regularizing term $\|x\|_1$ is replaced by $\sum_{i=1}^N \|x_i\|_2$. When $n_i = 1$, this reduces to (3). The z-update of the ADMM procedure needs to modified. We define the soft thresholding operator $\mathcal{S}_c \colon \mathbb{R}^m \to \mathbb{R}^m$ given by

$$\mathcal{S}_c(v) = \left(1 - \frac{c}{\|v\|_2}\right)_+ v,$$

with $\mathcal{S}_c(0) = 0$. Then the z-update consists of the N updates

$$z_i^{k+1} = \mathcal{S}_{\tau/\rho}(x_i^{k+1} + u^k), \quad i = 1, \ldots, N.$$

The method can be extended to deal with overlapping groups; see Boyd et al. [28] (Section 6.4).

There are many more applications of ADMM discussed in Boyd et al. [28], including consensus and sharing. See also Strang [171] for a brief overview.

52.9 Summary

The main concepts and results of this chapter are listed below:

- Dual ascent.
- Augmented Lagrangian.
- Penalty parameter.
- Method of multipliers.
- ADMM (alternating direction method of multipliers).

- x-update, z-update, λ-update.
- Scaled form of ADMM.
- Residual, dual residual.
- Stopping criteria.
- Proximity operator, proximal minimization.
- Quadratic programming.
- KKT equations.
- Soft thresholding operator.
- Shrinkage operator.
- Least absolute deviation.
- Basis pursuit.
- General ℓ^1-regularized loss minimization.
- Lasso regularization.
- Generalized lasso regularization.
- Group lasso.

52.10 Problems

Problem 52.1. In the method of multipliers described in Section 52.2, prove that choosing $\alpha^k = \rho$ guarantees that (u^{k+1}, λ^{k+1}) satisfies the equation

$$\nabla J_{u^{k+1}} + A^\top \lambda^{k+1} = 0.$$

Problem 52.2. Prove that the Inequality (A1) follows from the Inequalities (A2) and (A3) (see the proof of Theorem 52.1). For help consult Appendix A of Boyd et al. [28].

Problem 52.3. Consider Example 52.8. Prove that if $f = I_C$, the indicator function of a nonempty closed convex set C, then

$$x^+ = \arg\min_x \left(I_C(x) + (\rho/2)\left\|x - v\right\|_2^2\right) = \Pi_C(v),$$

the orthogonal projection of v onto C. In the special case where $C = \mathbb{R}^n_+$ (the first orthant), then

$$x^+ = (v)_+,$$

the vector obtained by setting the negative components of v to zero.

Problem 52.4. Prove that the soft thresholding operator S_c from Section 52.8 satisfies the equations

$$S_c(v) = (v - c)_+ - (-v - c)_+,$$

and

$$S_c(v) = (1 - c/|v|)_+ v, \quad v \neq 0.$$

Problem 52.5. Rederive the formula

$$S_c(v) = \begin{cases} v - c & \text{if } v > c \\ 0 & \text{if } |v| \leq c \\ v + c & \text{if } v < -c \end{cases}$$

using subgradients.

Problem 52.6. In basis pursuit (see Section 52.8 (2)) prove that

$$x^{k+1} = (I - A^\top (AA^\top)^{-1} A)(z^k - u^k) + A^\top (AA^\top)^{-1} b.$$

Problem 52.7. Implement (in `Matlab`) ADMM applied to lasso regularization as described in Section 52.6 (4). The stopping criterion should be based on feasibility tolerances ϵ^{pri} and ϵ^{dual}, say 10^{-4}, and on a maximum number of iteration steps, say 10000. There is a build in `Matlab` function `wthresh` implementing soft thresholding. You may use the `Matlab` command `randn` to create a random data set X and a random response vector y (see the help menu in `Matlab` under `lasso`). Try various values of ρ and τ. You will observe that the choice of ρ greatly affects the rate of convergence of the procedure.

Part IX

Applications to Machine Learning

Chapter 53

Positive Definite Kernels

This chapter is an introduction to positive definite kernels and the use of kernel functions in machine learning.

Let X be a nonempty set. If the set X represents a set of highly nonlinear data, it may be advantageous to map X into a space F of much higher dimension called the *feature space*, using a function $\varphi\colon X \to F$ called a *feature map*. This idea is that φ "unwinds" the description of the objects in F in an attempt to make it linear. The space F is usually a vector space equipped with an inner product $\langle -, - \rangle$. If F is infinite dimensional, then we assume that it is a Hilbert space.

Many algorithms that analyze or classify data make use of the inner products $\langle \varphi(x), \varphi(y) \rangle$, where $x, y \in X$. These algorithms make use of the function $\kappa\colon X \times X \to \mathbb{C}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X,$$

called a *kernel function*.

The kernel trick is to pretend that we have a feature embedding $\varphi\colon X \to F$ (actually unknown), but to only use inner products $\langle \varphi(x), \varphi(y) \rangle$ that can be evaluated using the original data through the known kernel function κ. It turns out that the functions of the form κ as above can be defined in terms of a condition which is reminiscent of positive semidefinite matrices (see Definition 53.2). Furthermore, every function satisfying Definition 53.2 arises from a suitable feature map into a Hilbert space; see Theorem 53.8.

We illustrate the kernel methods on kernel PCA (see Section 53.4).

53.1 Feature Maps and Kernel Functions

Definition 53.1. Let X be a nonempty set, let H be a (complex) Hilbert space, and let $\varphi\colon X \to H$ be a function called a *feature map*. The function $\kappa\colon X \times X \to \mathbb{C}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X,$$

is called a *kernel function*.

Remark: A *feature map* is often called a *feature embedding*, but this terminology is a bit misleading because it suggests that such a map is injective, which is not necessarily the case. Unfortunately this terminology is used by most people.

Example 53.1. Suppose we have two feature maps $\varphi_1 \colon X \to \mathbb{R}^{n_1}$ and $\varphi_2 \colon X \to \mathbb{R}^{n_2}$, and let $\kappa_1(x, y) = \langle \varphi_1(x), \varphi_1(y)\rangle$ and $\kappa_2(x, y) = \langle \varphi_2(x), \varphi_2(y)\rangle$ be the corresponding kernel functions (where $\langle -, -\rangle$ is the standard inner product on $\mathbb{R}^n$). Define the feature map $\varphi \colon X \to \mathbb{R}^{n_1+n_2}$ by

$$\varphi(x) = (\varphi_1(x), \varphi_2(x)),$$

an $(n_1 + n_2)$-tuple. We have

$$\begin{aligned}\langle \varphi(x), \varphi(y)\rangle &= \langle (\varphi_1(x), \varphi_2(x)), (\varphi_1(y), \varphi_2(y))\rangle = \langle \varphi_1(x), \varphi_1(y)\rangle + \langle \varphi_2(x), \varphi_2(y)\rangle \\ &= \kappa_1(x, y) + \kappa_2(x, y),\end{aligned}$$

which shows that the map κ given by

$$\kappa(x, y) = \kappa_1(x, y) + \kappa_2(x, y)$$

is the kernel function corresponding to the feature map $\varphi \colon X \to \mathbb{R}^{n_1+n_2}$.

Example 53.2. Let X be a subset of $\mathbb{R}^2$, and let $\varphi_1 \colon X \to \mathbb{R}^3$ be the map given by

$$\varphi_1(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

Figure 53.1 illustrates $\varphi_1 \colon X \to \mathbb{R}^3$ when $X = \{((x_1, x_2) \mid -10 \leq x_1 \leq 10, -10 \leq x_2 \leq 10\}$.

Observe that linear relations in the feature space $H = \mathbb{R}^3$ correspond to quadratic relations in the input space (of data). We have

$$\begin{aligned}\langle \varphi_1(x), \varphi_1(y)\rangle &= \langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2)\rangle \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\ &= (x_1y_1 + x_2y_2)^2 = \langle x, y\rangle^2,\end{aligned}$$

where $\langle x, y\rangle$ is the usual inner product on $\mathbb{R}^2$. Hence the function

$$\kappa(x, y) = \langle x, y\rangle^2$$

is a kernel function associated with the feature space $\mathbb{R}^3$.

If we now consider the map $\varphi_2 \colon X \to \mathbb{R}^4$ given by

$$\varphi_2(x_1, x_2) = (x_1^2, x_2^2, x_1x_2, x_1x_2),$$

we check immediately that

$$\langle \varphi_2(x), \varphi_2(y)\rangle = \kappa(x, y) = \langle x, y\rangle^2,$$

which shows that the same kernel can arise from different maps into different feature spaces.

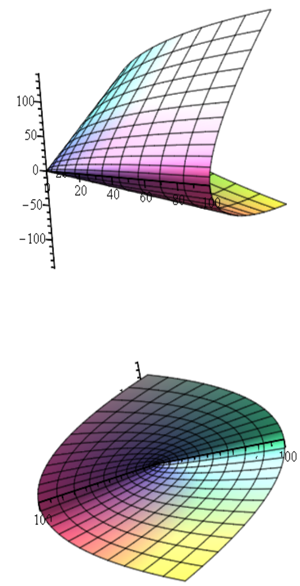


Figure 53.1: The parametric surface $\varphi_1(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ where $-10 \leq x_1 \leq 10$ and $-10 \leq x_2 \leq 10$.

Example 53.3. Example 53.2 can be generalized as follows. Suppose we have a feature map $\varphi_1 \colon X \to \mathbb{R}^n$ and let $\kappa_1(x, y) = \langle \varphi_1(x), \varphi_1(y) \rangle$ be the corresponding kernel function (where $\langle -, - \rangle$ is the standard inner product on $\mathbb{R}^n$). Define the feature map $\varphi \colon X \to \mathbb{R}^n \times \mathbb{R}^n$ by its n^2 components

$$\varphi(x)_{(i,j)} = (\varphi_1(x))_i(\varphi_1(x))_j, \qquad 1 \leq i, j \leq n,$$

with the inner product on $\mathbb{R}^n \times \mathbb{R}^n$ given by

$$\langle u, v \rangle = \sum_{i,j=1}^{n} u_{(i,j)} v_{(i,j)}.$$

Then we have

$$\begin{aligned}
\langle \varphi(x), \varphi(y) \rangle &= \sum_{i,j=1}^{n} \varphi_{(i,j)}(x)\varphi_{(i,j)}(y) \\
&= \sum_{i,j=1}^{n} (\varphi_1(x))_i(\varphi_1(x))_j(\varphi_1(y))_i(\varphi_1(y))_j \\
&= \sum_{i=1}^{n} (\varphi_1(x))_i(\varphi_1(y))_i \sum_{j=1}^{n} (\varphi_1(x))_j(\varphi_1(y))_j \\
&= (\kappa_1(x, y))^2.
\end{aligned}$$

Thus the map κ given by $\kappa(x,y) = (\kappa_1(x,y))^2$ is a kernel map associated with the feature map $\varphi\colon X \to \mathbb{R}^n \times \mathbb{R}^n$. The feature map φ is a direct generalization of the feature map φ_2 of Example 53.2.

The above argument is immediately adapted to show that if $\varphi_1\colon X \to \mathbb{R}^{n_1}$ and $\varphi_2\colon X \to \mathbb{R}^{n_2}$ are two feature maps and if $\kappa_1(x,y) = \langle \varphi_1(x), \varphi_1(y)\rangle$ and $\kappa_2(x,y) = \langle \varphi_2(x), \varphi_2(y)\rangle$ are the corresponding kernel functions, then the map defined by

$$\kappa(x,y) = \kappa_1(x,y)\kappa_2(x,y)$$

is a kernel function for the feature space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and the feature map

$$\varphi(x)_{(i,j)} = (\varphi_1(x))_i(\varphi_2(x))_j, \qquad 1 \leq i \leq n_1,\ 1 \leq j \leq n_2.$$

Example 53.4. Note that the feature map $\varphi\colon X \to \mathbb{R}^n \times \mathbb{R}^n$ is not very economical because if $i \neq j$ then the components $\varphi_{(i,j)}(x)$ and $\varphi_{(j,i)}(x)$ are both equal to $(\varphi_1(x))_i(\varphi_1(x))_j$. Therefore we can define the more economical embedding $\varphi'\colon X \to \mathbb{R}^{\binom{n+1}{2}}$ given by

$$\varphi'(x)_{(i,j)} = \begin{cases} (\varphi_1(x))_i^2 & i = j, \\ \sqrt{2}(\varphi_1(x))_i(\varphi_1(x))_j & i < j, \end{cases}$$

where the pairs (i,j) with $1 \leq i \leq j \leq n$ are ordered lexicographically. The feature map φ is a direct generalization of the feature map φ_1 of Example 53.2.

Observe that φ' can also be defined in the following way which makes it easier to come up with the generalization to any power:

$$\varphi'_{(i_1,\ldots,i_n)}(x) = \binom{2}{i_1 \cdots i_n}^{1/2} (\varphi_1(x))_1^{i_1}(\varphi_1(x))_1^{i_2} \cdots (\varphi_1(x))_1^{i_n}, \quad i_1 + i_2 + \cdots + i_n = 2,\ i_j \in \mathbb{N},$$

where the n-tuples $(i_1, \ldots, i_n)$ are ordered lexicographically. Recall that for any $m \geq 1$ and any $(i_1, \ldots, i_n) \in \mathbb{N}^m$ such that $i_1 + i_2 + \cdots + i_n = m$, we have

$$\binom{m}{i_1 \cdots i_n} = \frac{m!}{i_1! \cdots i_n!}.$$

More generally, for any $m \geq 2$, using the multinomial theorem, we can define a feature embedding $\varphi\colon X \to \mathbb{R}^{\binom{n+m-1}{m}}$ defining the kernel function κ given by $\kappa(x,y) = (\kappa_1(x,y))^m$, with φ given by

$$\varphi_{(i_1,\ldots,i_n)}(x) = \binom{m}{i_1 \cdots i_n}^{1/2} (\varphi_1(x))_1^{i_1}(\varphi_1(x))_1^{i_2} \cdots (\varphi_1(x))_1^{i_n}, \quad i_1 + i_2 + \cdots + i_n = m,\ i_j \in \mathbb{N},$$

where the n-tuples $(i_1, \ldots, i_n)$ are ordered lexicographically.

Example 53.5. For any positive real constant $R > 0$, the constant function $\kappa(x, y) = R$ is a kernel function corresponding to the feature map $\varphi\colon X \to \mathbb{R}$ given by $\varphi(x, y) = \sqrt{R}$.

By definition, the function $\kappa_1'\colon \mathbb{R}^n \to \mathbb{R}$ given by $\kappa_1'(x, y) = \langle x, y\rangle$ is a kernel function (the feature map is the identity map from $\mathbb{R}^n$ to itself). We just saw that for any positive real constant $R > 0$, the constant $\kappa_2'(x, y) = R$ is a kernel function. By Example 53.1, the function $\kappa_3'(x, y) = \kappa_1'(x, y) + \kappa_2'(x, y)$ is a kernel function, and for any integer $d \geq 1$, by Example 53.4, the function κ_d given by

$$\kappa_d(x, y) = (\kappa_3'(x, y))^d = (\langle x, y\rangle + R)^d,$$

is a kernel function on $\mathbb{R}^n$. By the binomial formula,

$$\kappa_d(x, y) = \sum_{m=0}^{d} R^{d-m}\langle x, y\rangle^m.$$

By Example 53.1, the feature map of this kernel function is the concatenation of the features of the $d+1$ kernel maps $R^{d-m}\langle x, y\rangle^m$. By Example 53.3, the components of the feature map of the kernel map $R^{d-m}\langle x, y\rangle^m$ are reweightings of the functions

$$\varphi_{(i_1,\ldots,i_n)}(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}, \quad i_1 + i_2 + \cdots + i_n = m,$$

with $(i_1, \ldots, i_n) \in \mathbb{N}^n$. Thus the components of the feature map of the kernel function κ_d are reweightings of the functions

$$\varphi_{(i_1,\ldots,i_n)}(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}, \quad i_1 + i_2 + \cdots + i_n \leq d,$$

with $(i_1, \ldots, i_n) \in \mathbb{N}^n$. It is easy to see that the dimension of this feature space is $\binom{m+d}{d}$.

There are a number of variations of the polynomial kernel κ_d; all-subsets embedding kernels, ANOVA kernels; see Shawe–Taylor and Christianini [159], Chapter III.

In the next example the set X is not a vector space.

Example 53.6. Let D be a finite set and let $X = 2^D$ be its power set. If $|D| = n$, let $H = \mathbb{R}^X \cong \mathbb{R}^{2^n}$. We are assuming that the subsets of D are enumerated in some fashion so that each coordinate of $\mathbb{R}^{2^n}$ corresponds to one of these subsets. For example, if $D = \{1, 2, 3, 4\}$, let

$$\begin{array}{llll}
U_1 = \emptyset & U_2 = \{1\} & U_3 = \{2\} & U_4 = \{3\} \\
U_5 = \{4\} & U_6 = \{1,2\} & U_7 = \{1,3\} & U_8 = \{1,4\} \\
U_9 = \{2,3\} & U_{10} = \{2,4\} & U_{11} = \{3,4\} & U_{12} = \{1,2,3\} \\
U_{13} = \{1,2,4\} & U_{14} = \{1,3,4\} & U_{15} = \{2,3,4\} & U_{16} = \{1,2,3,4\}.
\end{array}$$

Let $\varphi\colon X \to H$ be the feature map defined as follows: for any subsets $A, U \in X$,

$$\varphi(A)_U = \begin{cases} 1 & \text{if } U \subseteq A \\ 0 & \text{otherwise.} \end{cases}$$

For example, if $A_1 = \{1, 2, 3\}$, we obtain the vector

$$\varphi(\{1, 2, 3\}) = (1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0),$$

and if $A_2 = \{2, 3, 4\}$, we obtain the vector

$$\varphi(\{2, 3, 4\}) = (1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0).$$

For any two subsets A_1 and A_2 of D, it is easy to check that

$$\langle \varphi(A_1), \varphi(A_2)\rangle = 2^{|A_1 \cap A_2|},$$

the number of common subsets of A_1 and A_2. For example, $A_1 \cap A_2 = \{2, 3\}$, and

$$\langle \varphi(A_1), \varphi(A_2)\rangle = 4.$$

Therefore, the function $\kappa\colon X \times X \to \mathbb{R}$ given by

$$\kappa(A_1, A_2) = 2^{|A_1 \cap A_2|}, \qquad A_1, A_2 \subseteq D$$

is a kernel function.

Kernels on collections of sets can be defined in terms of measures.

Example 53.7. Let $(D, \mathcal{A})$ be a measurable space, where D is a nonempty set and $\mathcal{A}$ is a σ-algebra on D (the measurable sets). Let X be a subset of $\mathcal{A}$. If μ is a positive measure on $(D, \mathcal{A})$ and if μ is finite, which means that $\mu(D)$ is finite, then we can define the map $\kappa_1\colon X \times X \to \mathbb{R}$ given by

$$\kappa_1(A_1, A_2) = \mu(A_1 \cap A_2), \qquad A_1, A_2 \in X.$$

We can show that κ is a kernel function as follows. Let $H = \mathrm{L}^2_\mu(D, \mathcal{A}, \mathbb{R})$ be the Hilbert space of μ-square-integrable functions with the inner product

$$\langle f, g\rangle = \int_D f(s)g(s)\, d\mu(s),$$

and let $\varphi\colon X \to H$ be the feature embedding given by

$$\varphi(A) = \chi_A, \qquad A \in X,$$

the characteristic function of A. Then we have

$$\begin{aligned}
\kappa_1(A_1, A_2) = \mu(A_1 \cap A_2) &= \int_D \chi_{A_1 \cap A_2}(s)\, d\mu(s) \\
&= \int_D \chi_{A_1}(s)\chi_{A_2}(s)\, d\mu(s) = \langle \chi_{A_1}, \chi_{A_2}\rangle \\
&= \langle \varphi(A_1), \varphi(A_2)\rangle.
\end{aligned}$$

The above kernel is called the *intersection kernel*. If we assume that μ is normalized so that $\mu(D) = 1$, then we also have the *union complement kernel*:

$$\kappa_2(A_1, A_2) = \mu(\overline{A_1} \cap \overline{A_2}) = 1 - \mu(A_1 \cup A_2).$$

The sum κ_3 of the kernels κ_1 and κ_2 is the *agreement kernel*:

$$\kappa_s(A_1, A_2) = 1 - \mu(A_1 - A_2) - \mu(A_2 - A_1).$$

Many other kinds of kernels can be designed, in particular, graph kernels. For comprehensive presentations of kernels, see Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159].

Kernel functions have the following important property.

Proposition 53.1. *Let X be any nonempty set, let H be any (complex) Hilbert space, let $\varphi\colon X \to H$ be any function, and let $\kappa\colon X \times X \to \mathbb{C}$ be the kernel given by*

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X.$$

For any finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ matrix

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p} = (\langle \varphi(x_j), \varphi(x_i) \rangle)_{1 \leq i,j \leq p},$$

then we have

$$u^* K_S\, u \geq 0, \qquad \text{for all } u \in \mathbb{C}^p.$$

Proof. We have

$$\begin{aligned}
u^* K_S\, u = u^\top K_S^\top\, \overline{u} &= \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i \overline{u_j} \\
&= \sum_{i,j=1}^{p} \langle \varphi(x), \varphi(y) \rangle u_i \overline{u_j} \\
&= \left\langle \sum_{i=1}^{p} u_i \varphi(x_i), \sum_{j=1}^{p} u_j \varphi(x_j) \right\rangle = \left\| \sum_{i=1}^{p} u_i \varphi(x_i) \right\|^2 \geq 0,
\end{aligned}$$

as claimed. $\square$

53.2 Basic Properties of Positive Definite Kernels

Proposition 53.1 suggests a second approach to kernel functions which does not assume that a feature space and a feature map are provided. We will see in Section 53.3 that the two approaches are equivalent. The second approach is useful in practice because it is often difficult to define a feature space and a feature map in a simple manner.

Definition 53.2. Let X be a nonempty set. A function $\kappa\colon X \times X \to \mathbb{C}$ is a *positive definite kernel* if for every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ matrix

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p}$$

called a *Gram matrix*, then we have

$$u^* K_S\, u = \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i \overline{u_j} \geq 0, \qquad \text{for all } u \in \mathbb{C}^p.$$

Observe that Definition 53.2 does not require that $u^* K_S\, u > 0$ if $u \neq 0$, so the terminology *positive definite* is a bit abusive, and it would be more appropriate to use the terminology *positive semidefinite*. However, it seems customary to use the term *positive definite kernel*, or even *positive kernel*.

Proposition 53.2. *Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel. Then $\kappa(x, x) \geq 0$ for all $x \in X$, and for any finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ matrix K_S given by*

$$K_S = (\kappa(x_j, x_i))_{1 \leq i,j \leq p}$$

is Hermitian, that is, $K_S^* = K_S$.

Proof. The first property is obvious by choosing $S = \{x\}$. To prove that K_S is Hermitian, observe that we have

$$(u + v)^* K_S (u + v) = u^* K_S u + u^* K_S v + v^* K_S u + v^* K_S v,$$

and since $(u + v)^* K_S (u + v), u^* K_S u, v^* K_S v \geq 0$, we deduce that

$$2A = u^* K_S v + v^* K_S u \tag{1}$$

must be real. By replacing u by iu, we see that

$$2B = -iu^* K_S v + iv^* K_S u \tag{2}$$

must also be real. By multiplying Equation (2) by i and adding it to Equation (1) we get

$$u^* K_S v = A + iB. \tag{3}$$

By subtracting Equation (3) from Equation (1) we get

$$v^* K_S u = A - iB.$$

Then

$$u^* K_S^* v = \overline{v^* K_S u} = \overline{A - iB} = A + iB = u^* K_S v,$$

for all $u, v \in \mathbb{C}^*$, which implies $K_S^* = K_S$. $\square$

If the map $\kappa\colon X \times X \to \mathbb{R}$ is real-valued, then we have the following criterion for κ to be a positive definite kernel that only involves real vectors.

Proposition 53.3. *If $\kappa\colon X \times X \to \mathbb{R}$, then κ is a positive definite kernel iff for any finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ real matrix K_S given by*

$$K_S = (\kappa(x_k, x_j))_{1 \le j,k \le p}$$

is symmetric, that is, $K_S^\top = K_S$, and

$$u^\top K_S\, u = \sum_{j,k=1}^{p} \kappa(x_j, x_k) u_j u_k \ge 0, \qquad \text{for all } u \in \mathbb{R}^p.$$

Proof. If κ is a real-valued positive definite kernel, then the proposition is a trivial consequence of Proposition 53.2.

For the converse assume that κ is symmetric and that it satisfies the second condition of the proposition. We need to show that κ is a positive definite kernel with respect to complex vectors. If we write $u_k = a_k + i b_k$, then

$$\begin{aligned} u^* K_S\, u &= \sum_{j,k=1}^{p} \kappa(x_j, x_k)(a_j + i b_j)(a_k - i b_k) \\ &= \sum_{j,k=1}^{p} (a_j a_k + b_j b_k)\kappa(x_j, x_k) + i \sum_{j,k=1}^{p} (b_j a_k - a_j b_k)\kappa(x_j, x_k) \\ &= \sum_{j,k=1}^{p} (a_j a_k + b_j b_k)\kappa(x_j, x_k) + i \sum_{1 \le j < k \le p} b_j a_k(\kappa(x_j, x_k) - \kappa(x_k, x_j)). \end{aligned}$$

Thus $u^* K_S u$ is real iff K_S is symmetric. $\square$

Consequently we make the following definition.

Definition 53.3. Let X be a nonempty set. A function $\kappa\colon X \times X \to \mathbb{R}$ is a *(real) positive definite kernel* if $\kappa(x, y) = \kappa(y, x)$ for all $x, y \in X$, and for every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_S is the $p \times p$ real symmetric matrix

$$K_S = (\kappa(x_i, x_j))_{1 \le i,j \le p},$$

then we have

$$u^\top K_S\, u = \sum_{i,j=1}^{p} \kappa(x_i, x_j) u_i u_j \geq 0, \qquad \text{for all } u \in \mathbb{R}^p.$$

Among other things, the next proposition shows that a positive definite kernel satisfies the Cauchy–Schwarz inequality.

Proposition 53.4. *A Hermitian 2×2 matrix*

$$A = \begin{pmatrix} a & \overline{b} \\ b & d \end{pmatrix}$$

is positive semidefinite if and only if $a \geq 0$, $d \geq 0$, and $ad - |b|^2 \geq 0$.

Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel. For all $x, y \in X$, we have

$$|\kappa(x,y)|^2 \leq \kappa(x,x)\kappa(y,y).$$

Proof. For all $x, y \in \mathbb{C}$, we have

$$\begin{aligned} \begin{pmatrix} \overline{x} & \overline{y} \end{pmatrix} \begin{pmatrix} a & \overline{b} \\ b & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} \overline{x} & \overline{y} \end{pmatrix} \begin{pmatrix} ax + \overline{b}y \\ bx + dy \end{pmatrix} \\ &= a|x|^2 + bx\overline{y} + \overline{bx\overline{y}} + d|y|^2. \end{aligned}$$

If A is positive semidefinite, then we already know that $a \geq 0$ and $d \geq 0$. If $a = 0$, then we must have $b = 0$, since otherwise we can make $bx\overline{y} + \overline{bx\overline{y}}$, which is twice the real part of $bx\overline{y}$, as negative as we want. In this case, $ad - |b|^2 = 0$.

If $a > 0$, then

$$a|x|^2 + bx\overline{y} + \overline{bx\overline{y}} + d|y|^2 = a\left| x + \frac{\overline{b}}{a} y \right|^2 + \frac{|y|^2}{a}(ad - |b|^2).$$

If $ad - |b|^2 < 0$, we can pick $y \neq 0$ and $x = -(\overline{b}y)/a$, so that the above expression is negative. Therefore, $ad - |b|^2 \geq 0$. The converse is trivial.

If $x = y$, the inequality $|\kappa(x,y)|^2 \leq \kappa(x,x)\kappa(y,y)$ is trivial. If $x \neq y$, the inequality follows by applying the criterion for being positive semidefinite to the matrix

$$\begin{pmatrix} \kappa(x,x) & \overline{\kappa(x,y)} \\ \kappa(x,y) & \kappa(y,y) \end{pmatrix},$$

as claimed. ∎

The following property due to I. Schur (1911) shows that the pointwise product of two positive definite kernels is also a positive definite kernel.

Proposition 53.5. *(I. Schur) If $\kappa_1\colon X \times X \to \mathbb{C}$ and $\kappa_2\colon X \times X \to \mathbb{C}$ are two positive definite kernels, then the function $\kappa\colon X \times X \to \mathbb{C}$ given by $\kappa(x, y) = \kappa_1(x, y)\kappa_2(x, y)$ for all $x, y \in X$ is also a positive definite kernel.*

Proof. It suffices to prove that if $A = (a_{jk})$ and $B = (b_{jk})$ are two Hermitian positive semidefinite $p \times p$ matrices, then so is their pointwise product $C = A \circ B = (a_{jk}b_{jk})$ (also known as Hadamard or Schur product). Recall that a Hermitian positive semidefinite matrix A can be diagonalized as $A = U\Lambda U^*$, where Λ is a diagonal matrix with nonnegative entries and U is a unitary matrix. Let $\Lambda^{1/2}$ be the diagonal matrix consisting of the positive square roots of the diagonal entries in Λ. Then we have

$$A = U\Lambda U^* = U\Lambda^{1/2}\Lambda^{1/2}U^* = U\Lambda^{1/2}(U\Lambda^{1/2})^*.$$

Thus if we set $R = U\Lambda^{1/2}$, we have

$$A = RR^*,$$

which means that

$$a_{jk} = \sum_{h=1}^{p} r_{jh}\overline{r_{kh}}.$$

Then for any $u \in \mathbb{C}^p$, we have

$$\begin{aligned} u^*(A \circ B)u &= \sum_{j,k=1}^{p} a_{jk}b_{jk}u_j\overline{u_k} \\ &= \sum_{j,k=1}^{p}\sum_{h=1}^{p} r_{jh}\overline{r_{kh}}b_{jk}u_j\overline{u_k} \\ &= \sum_{h=1}^{p}\sum_{j,k=1}^{p} b_{jk}u_j r_{jh}\overline{u_k r_{kh}}. \end{aligned}$$

Since B is positive semidefinite, for each fixed h, we have

$$\sum_{j,k=1}^{p} b_{jk}u_j r_{jh}\overline{u_k r_{kh}} = \sum_{j,k=1}^{p} b_{jk}z_j\overline{z_k} \geq 0,$$

as we see by letting $z = (u_1r_{1h}, \ldots, u_pr_{ph})$, ☐

In contrast, the ordinary product AB of two symmetric positive semidefinite matrices A and B may not be symmetric positive semidefinite; see Section 8.9 for an example.

Here are other ways of obtaining new positive definite kernels from old ones.

Proposition 53.6. *Let $\kappa_1\colon X \times X \to \mathbb{C}$ and $\kappa_2\colon X \times X \to \mathbb{C}$ be two positive definite kernels, $f\colon X \to \mathbb{C}$ be a function, $\psi\colon X \to \mathbb{R}^N$ be a function, $\kappa_3\colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{C}$ be a positive definite kernel, and $a \in \mathbb{R}$ be any positive real number. Then the following functions are positive definite kernels:*

(1) $\kappa(x, y) = \kappa_1(x, y) + \kappa_2(x, y)$.

(2) $\kappa(x, y) = a\kappa_1(x, y)$.

(3) $\kappa(x, y) = f(x)\overline{f(y)}$.

(4) $\kappa(x, y) = \kappa_3(\psi(x), \psi(y))$.

(5) *If B is a symmetric positive semidefinite $n \times n$ matrix, then the map $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by*

$$\kappa(x, y) = x^\top B y$$

is a positive definite kernel.

Proof. (1) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K_1 is the $p \times p$ matrix

$$K_1 = (\kappa_1(x_k, x_j))_{1 \le j,k \le p}$$

and if if K_2 is the $p \times p$ matrix

$$K_2 = (\kappa_2(x_k, x_j))_{1 \le j,k \le p},$$

then for any $u \in \mathbb{C}^p$, we have

$$u^*(K_1 + K_2)u = u^* K_1 u + u^* K_2 u \ge 0,$$

since $u^* K_1 u \ge 0$ and $u^* K_2 u \ge 0$ because κ_2 and κ_2 are positive definite kernels, which means that K_1 and K_2 are positive semidefinite.

(2) We have

$$u^*(aK_1)u = au^* K_1 u \ge 0,$$

since $a > 0$ and $u^* K_1 u \ge 0$.

(3) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, if K is the $p \times p$ matrix

$$K = (\kappa(x_k, x_j))_{1 \le j,k \le p} = (\overline{f(x_k)} f(x_j))_{1 \le j,k \le p}$$

then we have

$$u^* K u = u^\top K^\top \overline{u} = \sum_{j,k=1}^{p} \kappa(x_j, x_k) u_j \overline{u_k} = \sum_{j,k=1}^{p} u_j f(x_j) \overline{u_k f(x_k)} = \left| \sum_{j=1}^{p} u_j f(x_j) \right|^2 \ge 0.$$

(4) For every finite subset $S = \{x_1, \ldots, x_p\}$ of X, the $p \times p$ matrix K given by

$$K = (\kappa(x_k, x_j))_{1 \le j,k \le p} = (\kappa_3(\psi(x_k), \psi(x_j)))_{1 \le j,k \le p}$$

is symmetric positive semidefinite since κ_3 is a positive definite kernel.

(5) As in the proof of Proposition 53.5 (adapted to the real case) there is a matrix R such that

$$B = RR^\top,$$

so

$$\kappa(x, y) = x^\top By = x^\top RR^\top y = (R^\top x)^\top R^\top y = \langle R^\top x, R^\top y\rangle,$$

so κ is the kernel function given by the feature map $\varphi(x) = R^\top x$ from $\mathbb{R}^n$ to itself, and by Proposition 53.1, it is a symmetric positive definite kernel. $\square$

Proposition 53.7. *Let $\kappa_1 \colon X \times X \to \mathbb{C}$ be a positive definite kernel, and let $p(z)$ be a polynomial with nonnegative coefficients. Then the following functions κ defined below are also positive definite kernels.*

(1) $\kappa(x, y) = p(\kappa_1(x, y))$.

(2) $\kappa(x, y) = e^{\kappa_1(x,y)}$.

(3) If X is real Hilbert space with inner product $\langle -, -\rangle_X$ and corresponding norm $\| \, \|_X$,

$$\kappa(x, y) = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

for any $\sigma > 0$.

Proof. (1) If $p(z) = a_m z^m + \cdots + a_1 z + a_0$, then

$$p(\kappa_1(x, y)) = a_m\kappa_1(x, y)^m + \cdots + a_1\kappa_1(x, y) + a_0.$$

Since $a_k \geq 0$ for $k = 0, \ldots, m$, by Proposition 53.5 and Proposition 53.6(2), each function $a_k\kappa_i(x, y)^k$ with $1 \leq k \leq m$ is a positive definite kernel, by Proposition 53.6(3) with $f(x) = \sqrt{a_0}$, the constant function a_0 is a positive definite kernel, and by Proposition 53.6(1), $p(\kappa_1(x, y))$ is a positive definite kernel.

(2) We have

$$e^{\kappa_1(x,y)} = \sum_{k=0}^{\infty} \frac{\kappa_1(x, y)^k}{k!}.$$

By (1), the partial sums

$$\sum_{k=0}^{m} \frac{\kappa_1(x, y)^k}{k!}$$

are positive definite kernels, and since $e^{\kappa_1(x,y)}$ is the (uniform) pointwise limit of positive definite kernels, it is also a positive definite kernel.

(3) By Proposition 53.6(2), since the map $(x, y) \mapsto \langle x, y\rangle_X$ is obviously a positive definite kernel (the feature map is the identity) and since $\sigma \neq 0$, the function $(x, y) \mapsto \langle x, y\rangle_X/\sigma^2$ is a positive definite kernel (by Proposition 53.6(2)), so by (2),

$$\kappa_1(x, y) = e^{\frac{\langle x,y\rangle_X}{\sigma^2}}$$

is a positive definite kernel. Let $f\colon X \to \mathbb{R}$ be the function given by

$$f(x) = e^{-\frac{\|x\|^2}{2\sigma^2}}.$$

Then by Proposition 53.6(3),

$$\kappa_2(x,y) = f(x)f(y) = e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|y\|^2}{2\sigma^2}} = e^{-\frac{\|x\|_X^2 + \|y\|_X^2}{2\sigma^2}}$$

is a positive definite kernel. By Proposition 53.5, the function $\kappa_1\kappa_2$ is a positive definite kernel, that is

$$\kappa_1(x,y)\kappa_2(x,y) = e^{\frac{\langle x,y\rangle_X}{\sigma^2}} e^{-\frac{\|x\|_X^2 + \|y\|_X^2}{2\sigma^2}} = e^{\frac{\langle x,y\rangle_X}{\sigma^2} - \frac{\|x\|_X^2 + \|y\|_X^2}{2\sigma^2}} = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

is a positive definite kernel. $\square$

Definition 53.4. The positive definite kernel

$$\kappa(x,y) = e^{-\frac{\|x-y\|_X^2}{2\sigma^2}}$$

is called a *Gaussian kernel*.

This kernel requires a feature map in an infinite-dimensional space because it is an infinite sum of distinct kernels.

Remark: If κ_1 is a positive definite kernel, the proof of Proposition 53.7(3) is immediately adapted to show that

$$\kappa(x,y) = e^{-\frac{\kappa_1(x,x) + \kappa_1(y,y) - 2\kappa_1(x,y)}{2\sigma^2}}$$

is a positive definite kernel.

Next we prove that every positive definite kernel arises from a feature map in a Hilbert space which is a function space.

53.3 Hilbert Space Representation of a Positive Definite Kernel

The following result shows how to construct a so-called *reproducing kernel Hilbert space*, for short RKHS, from a positive definite kernel.

Theorem 53.8. *Let $\kappa\colon X \times X \to \mathbb{C}$ be a positive definite kernel on a nonempty set X. For every $x \in X$, let $\kappa_x\colon X \to \mathbb{C}$ be the function given by*

$$\kappa_x(y) = \kappa(x,y), \qquad y \in X.$$

Let H_0 be the subspace of the vector space $\mathbb{C}^X$ of functions from X to $\mathbb{C}$ spanned by the family of functions $(\kappa_x)_{\in X}$, and let $\varphi\colon X \to H_0$ be the map given by $\varphi(x) = \kappa_x$. There is a Hermitian inner product $\langle -, - \rangle$ on H_0 such that

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \text{for all } x, y \in X.$$

The completion H of H_0 is a Hilbert space, and the map $\eta\colon H \to \mathbb{C}^X$ given by

$$\eta(f)(x) = \langle f, \kappa_x \rangle, \qquad x \in X,$$

is linear and injective, so H can be identified with a subspace of $\mathbb{C}^X$. We also have

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \text{for all } x, y \in X.$$

For all $f \in H_0$ and all $x \in X$,

$$\langle f, \kappa_x \rangle = f(x), \tag{$*$}$$

a property known as the **reproducing property**.

Proof.

Step 1. Define a candidate inner product.

For any two linear combinations $f = \sum_{j=1}^{p} \alpha_j \kappa_{x_j}$ and $g = \sum_{k=1}^{q} \beta_k \kappa_{y_k}$ in H_0, with $x_j, y_k \in X$ and $\alpha_j, \beta_k \in \mathbb{C}$, define $\langle f, g \rangle$ by

$$\langle f, g \rangle = \sum_{j=1}^{p} \sum_{k=1}^{q} \alpha_j \overline{\beta_k} \kappa(x_j, y_k). \tag{$\dagger$}$$

At first glance, the above expression appears to depend on the expression of f and g as linear combinations, but since $\kappa(x_j, y_k) = \overline{\kappa(y_k, x_j)}$, observe that

$$\sum_{k=1}^{q} \overline{\beta_k} f(y_k) = \sum_{j=1}^{p} \sum_{k=1}^{q} \alpha_j \overline{\beta_k} \kappa(x_j, y_k) = \sum_{j=1}^{p} \alpha_j \overline{g(x_j)}, \tag{$*$}$$

and since the first and the third term are equal for all linear combinations representing f and g, we conclude that ($\dagger$) depends only on f and g and not on their representation as a linear combination.

Step 2. Prove that the inner product defined by ($\dagger$) is Hermitian semidefinite.

Obviously ($\dagger$) defines a Hermitian sequilinear form. For every $f \in H_0$, we have

$$\langle f, f \rangle = \sum_{j,k=1}^{p} \alpha_j \overline{\alpha_k} \kappa(x_j, x_k) \geq 0,$$

since κ is a positive definite kernel.

Step 3. Prove that the inner product defined by (†) is positive definite.

For any finite subset $\{f_1, \ldots, f_n\}$ of H_0 and any $z \in \mathbb{C}^n$, we have

$$\sum_{j,k=1}^{n} \langle f_j, f_k \rangle z_j \overline{z_k} = \left\langle \sum_{j=1}^{n} z_j f_j, \sum_{j=1}^{n} z_j f_j \right\rangle \geq 0,$$

which shows that the map $(f, g) \mapsto \langle f, g \rangle$ from $H_0 \times H_0$ to $\mathbb{C}$ is a positive definite kernel.

Observe that for all $f \in H_0$ and all $x \in X$, (†) implies that

$$\langle f, \kappa_x \rangle = \sum_{j=1}^{k} \alpha_j \kappa(x_j, x) = f(x),$$

a property known as the *reproducing property*. The above implies that

$$\langle \kappa_x, \kappa_y \rangle = \kappa(x, y). \tag{$**$}$$

By Proposition 53.4 applied to the positive definite kernel $(f, g) \mapsto \langle f, g \rangle$, we have

$$|\langle f, \kappa_x \rangle|^2 \leq \langle f, f \rangle \langle \kappa_x, \kappa_x \rangle,$$

that is,

$$|f(x)|^2 \leq \langle f, f \rangle \kappa(x, x),$$

so $\langle f, f \rangle = 0$ implies that $f(x) = 0$ for all $x \in X$, which means that $\langle -, - \rangle$ as defined by (†) is positive definite. Therefore, $\langle -, - \rangle$ is a Hermitian inner product on H_0, and by $(**)$ and since $\varphi(x) = \kappa_x$, we have

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad \text{for all } x, y \in X.$$

Step 4. Define the embedding η.

Let H be the Hilbert space which is the completion of H_0, so that H_0 is dense in H. The map $\eta \colon H \to \mathbb{C}^X$ given by

$$\eta(f)(x) = \langle f, \kappa_x \rangle$$

is obviously linear, and it is injective because the family $(\kappa_x)_{x \in X}$ spans H_0 which is dense in H, thus it is also dense in H, so if $\langle f, \kappa_x \rangle = 0$ for all $x \in X$, then $f = 0$. □

Corollary 53.9. *If we identify a function $f \in H$ with the function $\eta(f)$, then we have the reproducing property*

$$\langle f, \kappa_x \rangle = f(x), \qquad \textit{for all } f \in H \textit{ and all } x \in X.$$

If X is finite, then $\mathbb{C}^X$ is finite-dimensional. If X is a separable topological space and if κ is continuous, then it can be shown that H is a separable Hilbert space.

Also, if $\kappa\colon X \times X \to \mathbb{R}$ is a real symmetric positive definite kernel, then we see immediately that Theorem 53.8 holds with H_0 a real Euclidean space and H a real Hilbert space.

$\circledast$ **Remark:** If $X = G$, where G is a locally compact group, then a function $p\colon G \to \mathbb{C}$ (not necessarily continuous) is *positive semidefinite* if for all $s_1, \ldots, s_n \in G$ and all $\xi_1, \ldots, \xi_n \in \mathbb{C}$, we have

$$\sum_{j,k=1}^{n} p(s_j^{-1}s_k)\xi_k\overline{\xi_j} \geq 0.$$

So if we define $\kappa\colon G \times G \to \mathbb{C}$ by

$$\kappa(s,t) = p(t^{-1}s),$$

then κ is a positive definite kernel on G. If p is continuous, then it is known that p arises from a unitary representation $U\colon G \to \mathbf{U}(H)$ of the group G in a Hilbert space H with inner product $\langle -, - \rangle$ (a homomorphism with a certain continuity property), in the sense that there is some vector $x_0 \in H$ such that

$$p(s) = \langle U(s)(x_0), x_0 \rangle, \qquad \text{for all } s \in G.$$

Since the $U(s)$ are unitary operators on H,

$$\begin{aligned} p(t^{-1}s) &= \langle U(t^{-1}s)(x_0), x_0 \rangle = \langle U(t^{-1})(U(s)(x_0)), x_0 \rangle \\ &= \langle U(t)^*(U(s)(x_0)), x_0 \rangle = \langle U(s)(x_0), U(t)(x_0) \rangle, \end{aligned}$$

which shows that

$$\kappa(s,t) = \langle U(s)(x_0), U(t)(x_0) \rangle,$$

so the map $\varphi\colon G \to H$ given by

$$\varphi(s) = U(s)(x_0)$$

is a feature map into the feature space H. This theorem is due to Gelfand and Raikov (1943).

The proof of Theorem 53.8 is essentially identical to part of Godement's proof of the above result about the correspondence between functions of positive type and unitary representations; see Helgason [90], Chapter IV, Theorem 1.5. Theorem 53.8 is a little more general since it does not assume that X is a group, but when G is a group, the feature map arises from a unitary representation.

53.4 Kernel PCA

As an application of kernel functions, we discuss a generalization of the method of principal component analysis (PCA). Suppose we have a set of data $S = \{x_1, \ldots, x_n\}$ in some input space $\mathcal{X}$, and pretend that we have an embedding $\varphi\colon \mathcal{X} \to F$ of $\mathcal{X}$ in a (real) feature space $(F, \langle -, - \rangle)$, but that we only have access to the kernel function $\kappa(x,y) = \langle \varphi(x), \varphi(y) \rangle$. We would like to do PCA analysis on the set $\varphi(S) = \{\varphi(x_1), \ldots, \varphi(x_n)\}$.

There are two obstacles:

(1) We need to center the data and compute the inner products of pairs of centered data. More precisely, if the centroid of $\varphi(S)$ is

$$\mu = \frac{1}{n}(\varphi(x_1) + \cdots + \varphi(x_n)),$$

then we need to compute the inner products $\langle \varphi(x) - \mu, \varphi(y) - \mu \rangle$.

(2) Let us assume that $F = \mathbb{R}^d$ with the standard Euclidean inner product and that the data points $\varphi(x_i)$ are expressed as *row vectors* X_i of an $n \times d$ matrix X (as it is customary). Then the inner products $\kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ are given by the *kernel matrix* $\mathbf{K} = XX^\top$. Be aware that with this representation, in the expression $\langle \varphi(x_i), \varphi(x_j) \rangle$, $\varphi(x_i)$ is a d-dimensional column vector, while $\varphi(x_i) = X_i^\top$. However, the jth component $(Y_k)_j$ of the principal component Y_k (viewed as a n-dimensional column vector) is given by the projection of $\widehat{X}_j = X_j - \mu$ onto the direction u_k (viewing μ as a d-dimensional row vector), which is a unit eigenvector of the matrix $(X - \mu)^\top (X - \mu)$ (where $\widehat{X} = X - \mu$ is the matrix whose jth row is $\widehat{X}_j = X_j - \mu$), is given by the inner product

$$\langle X_j - \mu, u_k \rangle = (Y_k)_j;$$

see Definition 23.2 and Theorem 23.11. The problem is that we know what the matrix $(X - \mu)(X - \mu)^\top$ is from (1), because it can be expressed in terms of $\mathbf{K}$, but we don't know what $(X - \mu)^\top (X - \mu)$ is because we don't have access to $\widehat{X} = X - \mu$.

Both difficulties are easily overcome. For (1) we have

$$\begin{aligned} \langle \varphi(x) - \mu, \varphi(y) - \mu \rangle &= \left\langle \varphi(x) - \frac{1}{n}\sum_{k=1}^n \varphi(x_k), \varphi(y) - \frac{1}{n}\sum_{k=1}^n \varphi(x_k) \right\rangle \\ &= \kappa(x, y) - \frac{1}{n}\sum_{i=1}^n \kappa(x, x_i) - \frac{1}{n}\sum_{j=1}^n \kappa(x_j, y) + \frac{1}{n^2}\sum_{i,j=1}^n \kappa(x_i, x_j). \end{aligned}$$

For (2), if $\mathbf{K}$ is the kernel matrix $\mathbf{K} = (\kappa(x_i, x_j))$, then the kernel matrix $\widehat{\mathbf{K}}$ corresponding to the kernel function $\widehat{\kappa}$ given by

$$\widehat{\kappa}(x, y) = \langle \varphi(x) - \mu, \varphi(y) - \mu \rangle$$

can be expressed in terms of $\mathbf{K}$. Let $\mathbf{1}$ be the column vector (of dimension n) whose entries are all 1. Then $\mathbf{1}\mathbf{1}^\top$ is the $n \times n$ matrix whose entries are all 1. If A is an $n \times n$ matrix, then $\mathbf{1}^\top A$ is the row vector consisting of the sums of the columns of A, $A\mathbf{1}$ is the column vector consisting of the sums of the rows of A, and $\mathbf{1}^\top A \mathbf{1}$ is the sum of all the entries in A. Then it is easy to see that the kernel matrix corresponding to the kernel function $\widehat{\kappa}$ is given by

$$\widehat{\mathbf{K}} = \mathbf{K} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}(\mathbf{1}^\top \mathbf{K}\mathbf{1})\mathbf{1}\mathbf{1}^\top.$$

Suppose $\widehat{X} = X - \mu$ has rank r. To overcome the second problem, note that if

$$\widehat{X} = VDU^\top$$

is an SVD for $\widehat{X}$, then

$$\widehat{X}^\top = UD^\top V^\top$$

is an SVD for $\widehat{X}^\top$, and the $r \times r$ submatrix of $D^\top$ consisting of the first r rows and r columns of $D^\top$ (and D), is the diagonal Σ_r matrix consisting of the singular values $\sigma_1 \geq \cdots \geq \sigma_r$ of $\widehat{X}$, so we can express the matrix U_r consisting of the first r columns u_k of U in terms of the matrix V_r consisting of the first r columns v_k of V $(1 \leq k \leq r)$ as

$$U_r = \widehat{X}^\top V_r \Sigma_r^{-1}.$$

Furthermore, $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ are the nonzero eigenvalues of $\widehat{\mathbf{K}} = \widehat{X}\widehat{X}^\top$, and the columns of V_r are corresponding unit eigenvectors of $\widehat{\mathbf{K}}$. From

$$U_r = \widehat{X}^\top V_r \Sigma_r^{-1}$$

the kth column u_k of U_r (which is a unit eigenvector of $\widehat{X}^\top \widehat{X}$ associated with the eigenvalue σ_k^2) is given by

$$u_k = \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{X}_i^\top = \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{\varphi(x_i)}, \quad 1 \leq k \leq r,$$

so the projection of $\widehat{\varphi(x)}$ onto u_k is given by

$$\begin{aligned}
\langle \widehat{\varphi(x)}, u_k \rangle &= \left\langle \widehat{\varphi(x)}, \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{\varphi(x_i)} \right\rangle \\
&= \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \left\langle \widehat{\varphi(x)}, \widehat{\varphi(x_i)} \right\rangle = \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{\kappa}(x, x_i).
\end{aligned}$$

Therefore, the jth component of the principal component Y_k in the principal direction u_k is given by

$$(Y_k)_j = \langle X_j - \mu, u_k \rangle = \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{\kappa}(x_j, x_i) = \sum_{i=1}^n \sigma_k^{-1} (v_k)_i \widehat{\mathbf{K}}_{ij}.$$

The generalization of kernel PCA to a general embedding $\varphi\colon \mathcal{X} \to F$ of $\mathcal{X}$ in a (real) feature space $(F, \langle -, - \rangle)$ (where F is not restricted to be equal to $\mathbb{R}^d$) with the kernel matrix $\mathbf{K}$ given by

$$\mathbf{K}_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle,$$

goes as follows.

- Let r be the rank of $\widehat{\mathbf{K}}$, where

$$\widehat{\mathbf{K}} = \mathbf{K} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}(\mathbf{1}^\top\mathbf{K}\mathbf{1})\mathbf{1}\mathbf{1}^\top,$$

  let $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ be the nonzero eigenvalues of $\widehat{\mathbf{K}}$, and let $v_1, \ldots, v_r$ be corresponding unit eigenvectors. The notation

$$\alpha_k = \sigma_k^{-1} v_k$$

  is often used, where the α_k are called the *dual variables*.

- The column vector Y_k $(1 \leq k \leq r)$ defined by

$$Y_k = \left(\sum_{i=1}^n (\alpha_k)_i \widehat{\mathbf{K}}_{ij}\right)_{j=1}^n$$

  is called the *kth kernel principal component* (for short *kth kernel PCA*) of the data set $S = \{x_1, \ldots, x_n\}$ in the direction $u_k = \sum_{i=1}^n \sigma_k^{-1}(v_k)_i \widehat{X}_i^\top$ (even though the matrix $\widehat{X}$ is *not known*).

53.5 Summary

The main concepts and results of this chapter are listed below:

- Feature map, feature embedding, feature space.
- Kernel function.
- Positive definite kernel, real positive definite kernel.
- Gram matrix.
- Hadamard product, Schur product.
- Gaussian kernel.
- Reproducing kernel Hilbert space (RKHS).
- Reproducing property.
- Intersection kernel, union complement kernel, agreement kernel.
- Kernel PCA.
- k-th kernel PCA.

53.6 Problems

Problem 53.1. Referring back to Example 53.3, prove that if $\varphi_1 \colon X \to \mathbb{R}^{n_1}$ and $\varphi_2 \colon X \to \mathbb{R}^{n_2}$ are two feature maps and if $\kappa_1(x, y) = \langle \varphi_1(x), \varphi_1(y) \rangle$ and $\kappa_2(x, y) = \langle \varphi_2(x), \varphi_2(y) \rangle$ are the corresponding kernel functions, then the map defined by

$$\kappa(x, y) = \kappa_1(x, y)\kappa_2(x, y)$$

is a kernel function, for the feature space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and the feature map

$$\varphi(x)_{(i,j)} = (\varphi_1(x))_i(\varphi_2(x))_j, \qquad 1 \leq i \leq n_1,\ 1 \leq j \leq n_2.$$

Problem 53.2. Referring back to Example 53.3, prove that the feature embedding $\varphi \colon X \to \mathbb{R}^{\binom{n+m-1}{m}}$ given by

$$\varphi_{(i_1,\ldots,i_n)}(x) = \binom{m}{i_1 \cdots i_n}^{1/2} (\varphi_1(x))_1^{i_1} (\varphi_1(x))_1^{i_2} \cdots (\varphi_1(x))_1^{i_n}, \quad i_1 + i_2 + \cdots + i_n = m,\ i_j \in \mathbb{N},$$

where the n-tuples $(i_1, \ldots, i_n)$ are ordered lexicographically, defines the kernel function κ given by $\kappa(x, y) = (\kappa_1(x, y))^m$.

Problem 53.3. In Example 53.6, prove that for any two subsets A_1 and A_2 of D,

$$\langle \varphi(A_1), \varphi(A_2) \rangle = 2^{|A_1 \cap A_2|},$$

the number of common subsets of A_1 and A_2.

Problem 53.4. Prove that the pointwise limit of positive definite kernels is also a positive definite kernel.

Problem 53.5. Prove that if κ_1 is a positive definite kernel, then

$$\kappa(x, y) = e^{-\frac{\kappa_1(x,x) + \kappa_1(y,y) - 2\kappa_1(x,y)}{2\sigma^2}}$$

is a positive definite kernel.

Chapter 54

Soft Margin Support Vector Machines

In Sections 50.5 and 50.6 we considered the problem of separating two nonempty disjoint finite sets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$ in $\mathbb{R}^n$. The goal is to find a hyperplane H of equation $w^\top x - b = 0$ (where $w \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$), such that all the blue points u_i are in one of the two open half-spaces determined by H, and all the red points v_j are in the other open half-space determined by H; see Figure 54.1.

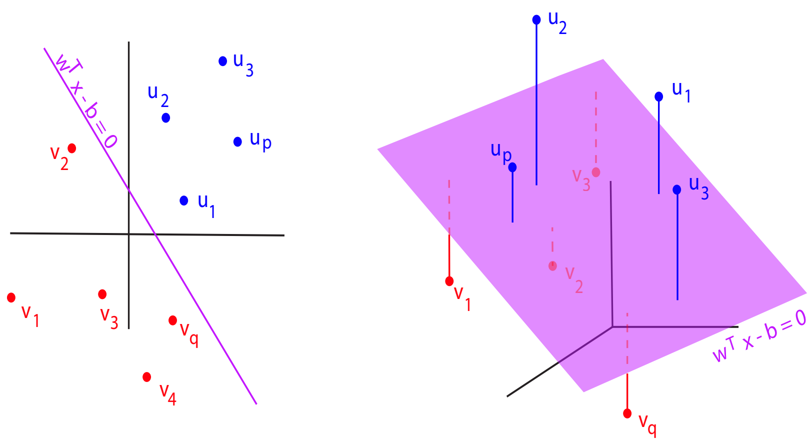


Figure 54.1: Two examples of the SVM separation problem. The left figure is SVM in $\mathbb{R}^2$, while the right figure is SVM in $\mathbb{R}^3$.

SVM picks a hyperplane which maximizes the minimum distance from these points to the hyperplane.

In this chapter we return to the problem of separating two disjoint sets of points, $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$, but this time we do not assume that these two sets are separable. To cope with nonseparability, we allow points to invade the safety zone around the separating hyperplane, and even points on the wrong side of the hyperplane. Such a method is called *soft margin support vector machine*. We discuss variations of this method, including ν-SV classification. In each case we present a careful derivation of the dual.

If the sets of points $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_q\}$ are not linearly separable (with $u_i, v_j \in \mathbb{R}^n$), we can use a trick from linear programming which is to introduce nonnegative "slack variables" $\epsilon = (\epsilon_1, \ldots, \epsilon_p) \in \mathbb{R}^p$ and $\xi = (\xi_1, \ldots, \xi_q) \in \mathbb{R}^q$ to relax the "hard" constraints

$$\begin{aligned} w^\top u_i - b &\geq \delta & \qquad i &= 1, \ldots, p \\ -w^\top v_j + b &\geq \delta & \qquad j &= 1, \ldots, q \end{aligned}$$

of Problem (SVM_{h1}) from Section 50.5 to the "soft" constraints

$$\begin{aligned} w^\top u_i - b &\geq \delta - \epsilon_i, & \epsilon_i &\geq 0 & i &= 1, \ldots, p \\ -w^\top v_j + b &\geq \delta - \xi_j, & \xi_j &\geq 0 & j &= 1, \ldots, q. \end{aligned}$$

Recall that $w \in \mathbb{R}^n$ and $b, \delta \in \mathbb{R}$.

If $\epsilon_i > 0$, the point u_i may be misclassified, in the sense that it can belong to the margin (the slab), or even to the wrong half-space classifying the negative (red) points. See Figures 54.5 (2) and (3). Similarly, if $\xi_j > 0$, the point v_j may be misclassified, in the sense that it can belong to the margin (the slab), or even to the wrong half-space classifying the positive (blue) points. We can think of ϵ_i as a measure of how much the constraint $w^\top u_i - b \geq \delta$ is violated, and similarly of ξ_j as a measure of how much the constraint $-w^\top v_j + b \geq \delta$ is violated. If $\epsilon = 0$ and $\xi = 0$, then we recover the original constraints. By making ϵ and ξ large enough, these constraints can always be satisfied. We add the constraint $w^\top w \leq 1$ and we minimize $-\delta$.

If instead of the constraints of Problem (SVM_{h1}) we use the hard constraints

$$\begin{aligned} w^\top u_i - b &\geq 1 & \qquad i &= 1, \ldots, p \\ -w^\top v_j + b &\geq 1 & \qquad j &= 1, \ldots, q \end{aligned}$$

of Problem (SVM_{h2}) (see Example 50.6), then we relax to the soft constraints

$$\begin{aligned} w^\top u_i - b &\geq 1 - \epsilon_i, & \epsilon_i &\geq 0 & i &= 1, \ldots, p \\ -w^\top v_j + b &\geq 1 - \xi_j, & \xi_j &\geq 0 & j &= 1, \ldots, q. \end{aligned}$$

In this case there is no constraint on w, but we minimize $(1/2)w^\top w$.

Ideally we would like to find a separating hyperplane that *minimizes the number of misclassified points*, which means that the variables ϵ_i and ξ_j should be as small as possible, but there is a trade-off in maximizing the margin (the thickness of the slab), and minimizing the number of misclassified points. This is reflected in the choice of the objective function, and there are several options, depending on whether we minimize a linear function of the variables ϵ_i and ξ_j, or a quadratic functions of these variables, or whether we include the term $(1/2)b^2$ in the objective function. These methods are known as *support vector classification* algorithms (for short *SVC* algorithms).

SVC algorithms seek an "optimal" separating hyperplane H of equation $w^\top x - b = 0$. If some new data $x \in \mathbb{R}^n$ comes in, we can classify it by determining in which of the two half spaces determined by the hyperplane H they belong by computing the sign of the quantity $w^\top x - b$. The function $\mathrm{sgn}\colon \mathbb{R} \to \{-1, 1\}$ is given by

$$\mathrm{sgn}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Then we define the *(binary) classification function* associated with the hyperplane H of equation $w^\top x - b = 0$ as

$$f(x) = \mathrm{sgn}(w^\top x - b).$$

Remarkably, all the known optimization problems for finding this hyperplane share the property that the weight vector w and the constant b are given by expressions that *only involves inner products of the input data points u_i and v_j*, and so does the classification function

$$f(x) = \mathrm{sgn}(w^\top x - b).$$

This is a key fact that allows a far reaching generalization of the support vector machine using the method of *kernels*.

The method of kernels consists in assuming that the input space $\mathbb{R}^n$ is embedded in a larger (possibly infinite dimensional) Euclidean space F (with an inner product $\langle -, - \rangle$) usually called a *feature space*, using a function

$$\varphi\colon \mathbb{R}^n \to F$$

called a *feature map*. The function $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle$$

is the kernel function associated with the embedding φ; see Chapter 53. The idea is that the feature map φ "unwinds" the input data, making it somehow more linear in the higher dimensional space F. Now even if we don't know what the feature space F is and what the embedding map φ is, we can pretend to solve our separation problem in F for the embedded data points $\varphi(u_i)$ and $\varphi(v_j)$. Thus we seek a hyperplane H of equation

$$\langle w, \zeta \rangle - b = 0, \quad \zeta \in F,$$

in the feature space F, to attempt to separate the points $\varphi(u_i)$ and the points $\varphi(v_j)$. As we said, it turns out that w and b are given by expression involving only the inner products $\kappa(u_i, u_j) = \langle \varphi(u_i), \varphi(u_j) \rangle, \kappa(u_i, v_j) = \langle \varphi(u_i), \varphi(v_j) \rangle$, and $\kappa(v_i, v_j) = \langle \varphi(v_i), \varphi(v_j) \rangle$, which form the symmetric $(p + q) \times (p + q)$ matrix $\mathbf{K}$ (a kernel matrix) given by

$$\mathbf{K}_{ij} = \begin{cases} \kappa(u_i, u_j) & 1 \leq i \leq p,\ 1 \leq j \leq q \\ -\kappa(u_i, v_{j-p}) & 1 \leq i \leq p,\ p + 1 \leq j \leq p + q \\ -\kappa(v_{i-p}, u_j) & p + 1 \leq i \leq p + q,\ 1 \leq j \leq p \\ \kappa(v_{i-p}, v_{j-q}) & p + 1 \leq i \leq p + q,\ p + 1 \leq j \leq p + q. \end{cases}$$

For example, if $p = 2$ and $q = 3$, we have the matrix

$$\mathbf{K} = \begin{pmatrix} \kappa(u_1, u_1) & \kappa(u_1, u_2) & -\kappa(u_1, v_1) & -\kappa(u_1, v_2) & -\kappa(u_1, v_3) \\ \kappa(u_2, u_1) & \kappa(u_2, u_2) & -\kappa(u_2, v_1) & -\kappa(u_2, v_2) & -\kappa(u_2, v_3) \\ -\kappa(v_1, u_1) & -\kappa(v_1, u_2) & \kappa(v_1, v_1) & \kappa(v_1, v_2) & \kappa(v_1, v_3) \\ -\kappa(v_2, u_1) & -\kappa(v_2, u_2) & \kappa(v_2, v_1) & \kappa(v_2, v_2) & \kappa(v_2, v_3) \\ -\kappa(v_3, u_1) & -\kappa(v_3, u_2) & \kappa(v_3, v_1) & \kappa(v_3, v_2) & \kappa(v_3, v_3) \end{pmatrix}.$$

Then the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

for points in the original data space $\mathbb{R}^n$ is also expressed solely in terms of the matrix $\mathbf{K}$ and the inner products $\kappa(u_i, x) = \langle \varphi(u_i), \varphi(x)\rangle$ and $\kappa(v_j, x) = \langle \varphi(v_j), \varphi(x)\rangle$. As a consequence, in the original data space $\mathbb{R}^n$, the hypersurface

$$\mathcal{S} = \{x \in \mathbb{R}^n \mid \langle w, \varphi(x)\rangle - b = 0\}$$

separates the data points u_i and v_j, but it is not an affine subspace of $\mathbb{R}^n$. The classification function f tells us on which "side" of $\mathcal{S}$ is a new data point $x \in \mathbb{R}^n$. Thus, we managed to separate the data points u_i and v_j that are not separable by an affine hyperplane, by a *nonaffine hypersurface* $\mathcal{S}$, by assuming that an embdedding $\varphi\colon \mathbb{R}^n \to F$ exists, even though we don't know what it is, but having access to F through the kernel function $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by the inner products $\kappa(x, y) = \langle \varphi(x), \varphi(y)\rangle$.

In practice the art of using the kernel method is to choose the right kernel (as the knight says in Indiana Jones, to "choose wisely.").

The method of kernels is very flexible. It also applies to the soft margin versions of SVM, but also to regression problems, to principal component analysis (PCA), and to other problems arising in machine learning.

We discussed the method of kernels in Chapter 53. Other comprehensive presentations of the method of kernels are found in Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159]. See also Bishop [23].

We first consider the soft margin SVM arising from Problem (SVM_{h1}).

54.1 Soft Margin Support Vector Machines; (SVM_{s1})

In this section we derive the dual function G associated with the following version of the soft margin SVM coming from Problem (SVM_{h1}), where the maximization of the margin δ has been replaced by the minimization of $-\delta$, and where we added a "regularizing term" $K\Big(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\Big)$ whose purpose is to make $\epsilon \in \mathbb{R}^p$ and $\xi \in \mathbb{R}^q$ *sparse* (that is, try to make ϵ_i and ξ_j have as many zeros as possible), where $K > 0$ is a fixed constant that can be adjusted to determine the influence of this regularizing term. If the primal problem (SVM_{s1}) has an optimal solution $(w, \delta, b, \epsilon, \xi)$, we attempt to use the dual function G to obtain it, but we will see that with this particular formulation of the problem, the constraint $w^\top w \leq 1$ causes troubles even though it is convex.

Soft margin SVM (SVM_{s1}):

$$
\begin{aligned}
&\text{minimize} \quad -\delta + K\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q\\
&\qquad w^\top w \leq 1.
\end{aligned}
$$

It is customary to write $\ell = p + q$. Figure 54.2 illustrates the correct margin half space associated with $w^\top x - b - \delta = 0$ while Figure 54.3 illustrates the correct margin half space associated with $w^\top x - b + \delta = 0$. Ideally, all the points should be contained in one of the two correct shifted margin regions described by affine constraints $w^\top u_i - b \geq \delta - \epsilon_i$, or $-w^\top v_j + b \geq \delta - \xi_j$.

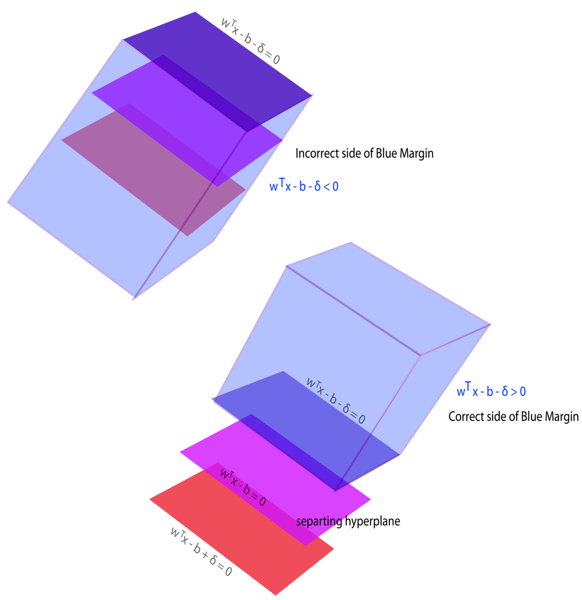


Figure 54.2: The blue margin half space associated with $w^\top x - b - \delta = 0$.

For this problem, the primal problem may have an optimal solution $(w, \delta, b, \epsilon, \xi)$ with $\|w\| = 1$ and $\delta > 0$, but if the sets of points are not linearly separable then an optimal solution of the dual may not yield w.

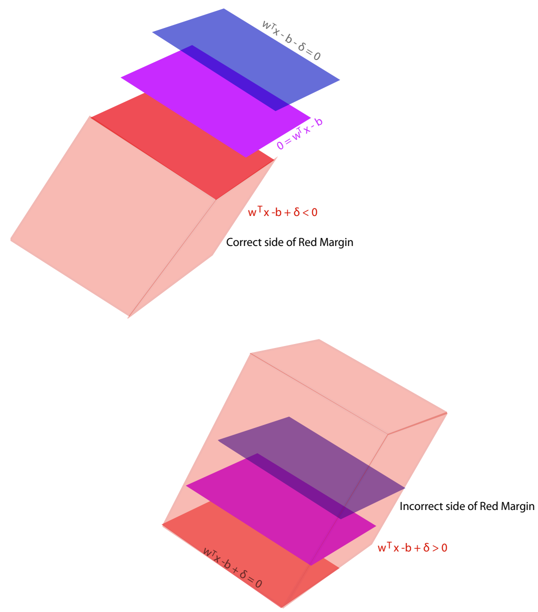


Figure 54.3: The red margin half space associated with $w^\top x - b + \delta = 0$.

The objective function of our problem is affine and the only nonaffine constraint $w^\top w \leq 1$ is convex. This constraint is qualified because for any $w \neq 0$ such that $w^\top w < 1$ and for any $\delta > 0$ and any b we can pick ϵ and ξ large enough so that the constraints are satisfied. Consequently, by Theorem 50.17(2) *if* the primal problem (SVM_{s1}) has an optimal solution, *then* the dual problem has a solution too, and the duality gap is zero.

Unfortunately this does not imply that an optimal solution of the dual yields an optimal solution of the primal because the hypotheses of Theorem 50.17(1) fail to hold. In general, there may not be a unique vector $(w, \epsilon, \xi, b, \delta)$ such that

$$\inf_{w,\epsilon,\xi,b,\delta} L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) = G(\lambda, \mu, \alpha, \beta, \gamma).$$

If the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable, then the dual problem may have a solution for which $\gamma = 0$,

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2},$$

and

$$\sum_{i=1}^{p} \lambda_i u_i = \sum_{j=1}^{q} \mu_j v_j,$$

so that the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$, which is a *partial function*, is defined and has the value $G(\lambda, \mu, \alpha, \beta, 0) = 0$. Such a pair (λ, μ) corresponds to the coefficients of two convex combinations

$$\sum_{i=1}^{p} 2\lambda_i u_i = \sum_{j=1}^{q} 2\mu_j v_j$$

which correspond to the *same point* in the (nonempty) intersection of the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$. It turns out that the only connection between w and the dual function is the equation

$$2\gamma w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

and when $\gamma = 0$ this is equation is $0 = 0$, *so the dual problem is useless to determine* w. This point seems to have been missed in the literature (for example, in Shawe–Taylor and Christianini [159], Section 7.2). What the dual problem does show is that $\delta \geq 0$. However, if $\gamma \neq 0$, then w is determined by any solution (λ, μ) of the dual.

It still remains to compute δ and b, which can be done under a mild hypothesis that we call the **Standard Margin Hypothesis**.

Let $\lambda \in \mathbb{R}^p_+$ be the Lagrange multipliers associated with the inequalities $w^\top u_i - b \geq \delta - \epsilon_i$, let $\mu \in \mathbb{R}^q_+$ be the Lagrange multipliers are associated with the inequalities $-w^\top v_j + b \geq \delta - \xi_j$, let $\alpha \in \mathbb{R}^p_+$ be the Lagrange multipliers associated with the inequalities $\epsilon_i \geq 0$, $\beta \in \mathbb{R}^q_+$ be the Lagrange multipliers associated with the inequalities $\xi_j \geq 0$, and let $\gamma \in \mathbb{R}^+$ be the Lagrange multiplier associated with the inequality $w^\top w \leq 1$.

The linear constraints are given by the $2(p+q) \times (n+p+q+2)$ matrix given in block form by

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} & \mathbf{1}_{p+q} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \end{pmatrix},$$

where X is the $n \times (p+q)$ matrix

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and the linear constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} & \mathbf{1}_{p+q} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \\ \delta \end{pmatrix} \leq \begin{pmatrix} 0_{p+q} \\ 0_{p+q} \end{pmatrix}.$$

More explicitly, C is the following matrix:

$$C = \begin{pmatrix} -u_1^\top & -1 & \cdots & 0 & 0 & \cdots & 0 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ -u_p^\top & 0 & \cdots & -1 & 0 & \cdots & 0 & 1 & 1 \\ v_1^\top & 0 & \cdots & 0 & -1 & \cdots & 0 & -1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ v_q^\top & 0 & \cdots & 0 & 0 & \cdots & -1 & -1 & 1 \\ 0 & -1 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & -1 & 0 & 0 \end{pmatrix}.$$

The objective function is given by

$$J(w, \epsilon, \xi, b, \delta) = -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$, and $\gamma \in \mathbb{R}^+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) &= -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\quad + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \delta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} + \gamma(w^\top w - 1). \end{aligned}$$

Since

$$\begin{aligned} \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \delta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} &= w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) \\ &\quad + \delta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu), \end{aligned}$$

the Lagrangian can be written as

$$\begin{aligned} L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) &= -\delta + K(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \gamma(w^\top w - 1) \\ &\quad - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \delta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) \\ &= (\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - 1)\delta + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \gamma(w^\top w - 1) \\ &\quad + \epsilon^\top(K\mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K\mathbf{1}_q - (\mu + \beta)) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$ we minimize $L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma)$ with respect to w, ϵ, ξ, b, and δ. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \delta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \delta)$ iff $\nabla L_{w,\epsilon,\xi,b,\delta} = 0$, so we compute the gradient with respect to $w, \epsilon, \xi, b, \delta$, and we get

$$\nabla L_{w,\epsilon,\xi,b,\delta} = \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + 2\gamma w \\ K\mathbf{1}_p - (\lambda + \alpha) \\ K\mathbf{1}_q - (\mu + \beta) \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - 1 \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\delta} = 0$ we get the equations

$$\begin{aligned} 2\gamma w &= -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ \lambda + \alpha &= K\mathbf{1}_p \\ \mu + \beta &= K\mathbf{1}_q \\ \mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= 1. \end{aligned} \tag{$*_w$}$$

The second and third equations are equivalent to the inequalities

$$0 \leq \lambda_i, \mu_j \leq K, \quad i = 1, \ldots, p,\ j = 1, \ldots, q,$$

often called *box constraints*, and the fourth and fifth equations yield

$$\mathbf{1}_p^\top \lambda = \mathbf{1}_q^\top \mu = \frac{1}{2}.$$

First let us consider the singular case $\gamma = 0$. In this case, $(*_w)$ implies that

$$X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0,$$

and the term $\gamma(w^\top w - 1)$ is missing from the Lagrangian, which in view of the other four equations above reduces to

$$L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, 0) = w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0.$$

In summary, we proved that if $\gamma = 0$, then

$$G(\lambda, \mu, \alpha, \beta, 0) = \begin{cases} 0 & \text{if } \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ 0 \leq \lambda_i \leq K,\ i = 1, \ldots, p \\ 0 \leq \mu_j \leq K,\ j = 1, \ldots, q \end{cases} \\ -\infty & \text{otherwise} \\ & \text{and } \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j = 0. \end{cases}$$

Geometrically, (λ, μ) corresponds to the coefficients of two convex combinations

$$\sum_{i=1}^{p} 2\lambda_i u_i = \sum_{j=1}^{q} 2\mu_j v_j$$

which correspond to the *same point* in the intersection of the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$ iff the sets $\{u_i\}$ and $\{v_j\}$ are *not linearly separable*. If the sets $\{u_i\}$ and $\{v_j\}$ are *linearly separable*, then the convex hulls $\mathrm{conv}(u_1, \ldots, u_p)$ and $\mathrm{conv}(v_1, \ldots, v_q)$ are disjoint, which implies that $\gamma > 0$.

Let us now assume that $\gamma > 0$. Plugging back w from equation $(*_w)$ into the Lagrangian, after simplifications we get

$$\begin{aligned} G(\lambda, \mu, \alpha, \beta, \gamma) &= -\frac{1}{2\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \frac{\gamma}{4\gamma^2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma \\ &= -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma, \end{aligned}$$

so if $\gamma > 0$ the dual function is independent of α, β and is given by

$$G(\lambda, \mu, \alpha, \beta, \gamma) = \begin{cases} -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma & \text{if} \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j = \frac{1}{2} \\ 0 \leq \lambda_i \leq K, \ i = 1, \ldots, p \\ 0 \leq \mu_j \leq K, \ j = 1, \ldots, q \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

Since $X^\top X$ is symmetric positive semidefinite and $\gamma \geq 0$, obviously

$$G(\lambda, \mu, \alpha, \beta, \gamma) \leq 0$$

for all $\gamma > 0$.

The dual program is given by

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{4\gamma} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \gamma \quad \text{if } \gamma > 0 \\ & 0 \quad \text{if} \quad \gamma = 0 \\ \text{subject to} & \\ & \sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0 \\ & \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = 1 \\ & 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ & 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

Also, if $\gamma = 0$, then $X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0$.

Maximizing with respect to $\gamma > 0$ by setting $\frac{\partial}{\partial \gamma} G(\lambda, \mu, \alpha, \beta, \gamma) = 0$ yields

$$\gamma^2 = \frac{1}{4} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

so we obtain

$$G(\lambda, \mu) = -\left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2}.$$

Finally, since $G(\lambda, \mu) = 0$ and $X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = 0$ if $\gamma = 0$, the dual program is equivalent to the following minimization program:

Dual of Soft margin SVM (SVM$_{s1}$):

$$\begin{aligned}
&\text{minimize} \quad \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = 1 \\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q.
\end{aligned}$$

Observe that the constraints imply that K must be chosen so that

$$K \geq \max\left\{ \frac{1}{2p}, \frac{1}{2q} \right\}.$$

If $(w, \delta, b, \epsilon, \xi)$ is an optimal solution of Problem (SVM$_{s1}$), then the complementary slackness conditions yield a classification of the points u_i and v_j in terms of the values of λ and μ. Indeed, we have $\epsilon_i \alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j \beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K$, it follows that $\epsilon_i \alpha_i = 0$ iff $\epsilon_i(K - \lambda_i) = 0$, and since $\mu_j + \beta_j = K$, we have $\xi_j \beta_j = 0$ iff $\xi_j(K - \mu_j) = 0$. Thus if $\epsilon_i > 0$, then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$, then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$, then $\xi_j = 0$ and v_j is correctly classified. We have the following classification:

(1) If $0 < \lambda_i < K$, then $\epsilon_i = 0$ and the i-th inequality is active, so

$$w^\top u_i - b - \delta = 0.$$

This means that u_i is on the blue margin (the hyperplane $H_{w,b+\delta}$ of equation $w^\top x = b + \delta$) and is classified correctly.

Similarly, if $0 < \mu_j < K$, then $\xi_j = 0$ and

$$w^\top v_j - b + \delta = 0,$$

so v_j is on the red margin (the hyperplane $H_{w,b-\delta}$ of equation $w^\top x = b - \delta$) and is classified correctly. See Figure 54.4.

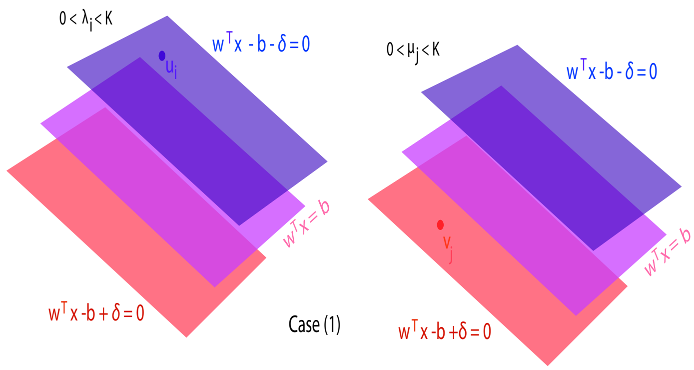


Figure 54.4: When $0 < \lambda_i < K$, u_i is contained within the blue margin hyperplane. When $0 < \mu_j < K$, v_j is contained within the red margin hyperplane.

(2) If $\lambda_i = K$, then the i-th inequality is active, so

$$w^\top u_i - b - \delta = -\epsilon_i.$$

If $\epsilon_i = 0$, then the point u_i is on the blue margin. If $\epsilon_i > 0$, then u_i is within the open half space bounded by the blue margin hyperplane $H_{w,b+\delta}$ and containing the separating hyperplane $H_{w,b}$; if $\epsilon_i \leq \delta$, then u_i is classified correctly, and if $\epsilon_i > \delta$, then u_i is misclassified (u_i lies on the wrong side of the separating hyperplane, the red side). See Figure 54.5.

Similarly, if $\mu_j = K$, then

$$w^\top v_j - b + \delta = \xi_j.$$

If $\xi_j = 0$, then the point v_j is on the red margin. If $\xi_j > 0$, then v_j is within the open half space bounded by the red margin hyperplane $H_{w,b-\delta}$ and containing the

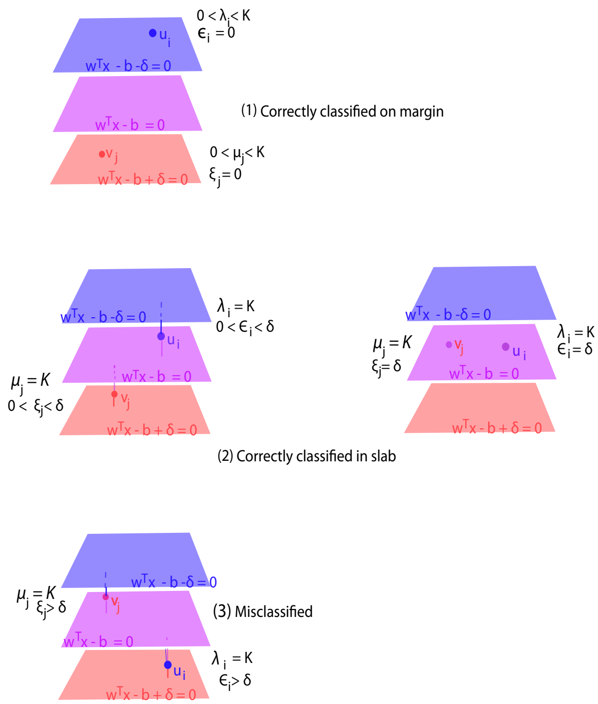


Figure 54.5: Figure (1) illustrates the case of u_i contained in the margin and occurs when $\epsilon_i = 0$. Figure (1) also illustrates the case of v_j contained in the margin when $\xi_j = 0$. The left illustration of Figure (2) is when u_i is inside the margin yet still on the correct side of the separating hyperplane $w^\top x - b = 0$. Similarly, v_j is inside the margin on the correct side of the separating hyperplane. The right illustration depicts u_i and v_j on the separating hyperplane. Figure (3) illustrations a misclassification of u_i and v_j.

separating hyperplane $H_{w,b}$; if $\xi_j \leq \delta$, then v_j is classified correctly, and if $\xi_j > \delta$, then v_j is misclassified (v_j lies on the wrong side of the separating hyperplane, the blue side). See Figure 54.5.

(3) If $\lambda_i = 0$, then $\epsilon_i = 0$ and the i-th inequality may or may not be active, so

$$w^\top u_i - b - \delta \geq 0.$$

Thus u_i is in the closed half space on the blue side bounded by the blue margin hyperplane $H_{w,b+\delta}$ (of course, classified correctly).

Similarly, if $\mu_j = 0$, then

$$w^\top v_j - b + \delta \leq 0$$

and v_j is in the closed half space on the red side bounded by the red margin hyperplane $H_{w,b-\delta}$ (of course, classified correctly). See Figure 54.6.

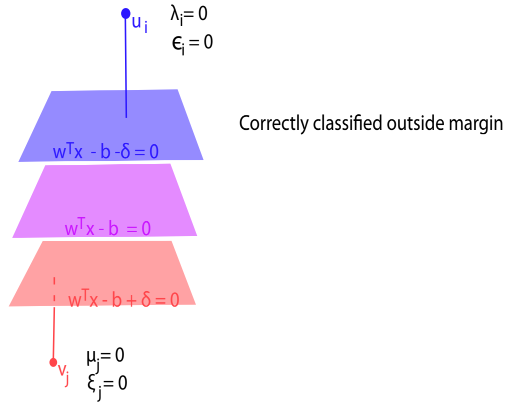


Figure 54.6: When $\lambda_i = 0$, u_i is correctly classified outside the blue margin. When $\mu_j = 0$, v_j is correctly classified outside the red margin.

Definition 54.1. The vectors u_i on the blue margin $H_{w,b+\delta}$ and the vectors v_j on the red margin $H_{w,b-\delta}$ are called *support vectors*. Support vectors correspond to vectors u_i for which $w^\top u_i - b - \delta = 0$ (which implies $\epsilon_i = 0$), and vectors v_j for which $w^\top v_j - b + \delta = 0$ (which implies $\xi_j = 0$). Support vectors u_i such that $0 < \lambda_i < K$ and support vectors v_j such that $0 < \mu_j < K$ are *support vectors of type 1*. Support vectors of type 1 play a special role so we denote the sets of indices associated with them by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K\}. \end{aligned}$$

We denote their cardinalities by $numsvl_1 = |I_\lambda|$ and $numsvm_1 = |I_\mu|$. Support vectors u_i such that $\lambda_i = K$ and support vectors v_j such that $\mu_j = K$ are *support vectors of type 2*. Those support vectors u_i such that $\lambda_i = 0$ and those support vectors v_j such that $\mu_j = 0$ are called *exceptional support vectors*.

The vectors u_i for which $\lambda_i = K$ and the vectors v_j for which $\mu_j = K$ are said to *fail the margin*. The sets of indices associated with the vectors failing the margin are denoted by

$$\begin{aligned} K_\lambda &= \{i \in \{1, \ldots, p\} \mid \lambda_i = K\} \\ K_\mu &= \{j \in \{1, \ldots, q\} \mid \mu_j = K\}. \end{aligned}$$

We denote their cardinalities by $p_f = |K_\lambda|$ and $q_f = |K_\mu|$.

Vectors u_i such that $\lambda_i > 0$ and vectors v_j such that $\mu_j > 0$ are said to *have margin at most δ*. The sets of indices associated with these vectors are denoted by

$$\begin{aligned} I_{\lambda>0} &= \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \\ I_{\mu>0} &= \{j \in \{1, \ldots, q\} \mid \mu_j > 0\}. \end{aligned}$$

We denote their cardinalities by $p_m = |I_{\lambda>0}|$ and $q_m = |I_{\mu>0}|$.

Obviously, $I_{\lambda>0} = I_\lambda \cup K_\lambda$ and $I_{\mu>0} = I_\mu \cup K_\mu$, so $p_f \leq p_m$ and $q_f \leq q_m$. *Intuitively a blue point that fails the margin is on the wrong side of the blue margin and a red point that fails the margin is on the wrong side of the red margin.* The points in $I_{\lambda>0}$ not in K_λ are on the blue margin and the points in $I_{\mu>0}$ not in K_μ are on the red margin. There are $p - p_m$ points u_i classified correctly on the blue side and outside the δ-slab and there are $q - q_m$ points v_j classified correctly on the red side and outside the δ-slab.

It is easy to show that we have the following bounds on K:

$$\max\left\{\frac{1}{2p_m}, \frac{1}{2q_m}\right\} \leq K \leq \min\left\{\frac{1}{2p_f}, \frac{1}{2q_f}\right\}.$$

These inequalities restrict the choice of K quite heavily.

It will also be useful to understand how points are classified in terms of ϵ_i (or ξ_j).

(1) If $\epsilon_i > 0$, then by complementary slackness $\lambda_i = K$, so the ith equation is active and by (2) above,

$$w^\top u_i - b - \delta = -\epsilon_i.$$

Since $\epsilon_i > 0$, the point u_i is within the open half space bounded by the blue margin hyperplane $H_{w,b+\delta}$ and containing the separating hyperplane $H_{w,b}$; if $\epsilon_i \leq \delta$, then u_i is classified correctly, and if $\epsilon_i > \delta$, then u_i is misclassified.

Similarly, if $\xi_j > 0$, then v_j is within the open half space bounded by the red margin hyperplane $H_{w,b-\delta}$ and containing the separating hyperplane $H_{w,b}$; if $\xi_j \leq \delta$, then v_j is classified correctly, and if $\xi_j > \delta$, then v_j is misclassified.

(2) If $\epsilon_i = 0$, then the point u_i is correctly classified. If $\lambda_i = 0$, then by (3) above, u_i is in the closed half space on the blue side bounded by the blue margin hyperplane $H_{w,b+\delta}$. If $\lambda_i > 0$, then by (1) and (2) above, the point u_i is on the blue margin.

Similarly, if $\xi_j = 0$, then the point v_j is correctly classified. If $\mu_j = 0$, then v_j is in the closed half space on the red side bounded by the red margin hyperplane $H_{w,b-\delta}$, and if $\mu_j > 0$, then the point v_j is on the red margin.

It shown in Section 54.2 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ.

If the optimal value is 0, then $\gamma = 0$ and $X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = 0$, so in this case it is not possible to determine w. However, if the optimal value is > 0, then once a solution for λ and μ is obtained, by $(*_w)$, we have

$$\begin{aligned}\gamma &= \frac{1}{2}\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}\right)^{1/2} \\ w &= \frac{1}{2\gamma}\left(\sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j\right),\end{aligned}$$

so we get

$$w = \frac{\displaystyle\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j}{\left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2}},$$

which is the result of making $\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j$ a unit vector, since

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix}.$$

It remains to find b and δ, which are not given by the dual program and for this we use the complementary slackness conditions.

The equations

$$\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2}$$

imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, but a priori, nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. If the following mild hypothesis holds, then b and δ can be found.

Standard Margin Hypothesis for (SVM$_{s1}$). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is a support vector of type 1 on the blue margin, and some v_{j_0} is a support of type 1 on the red margin.

If the **Standard Margin Hypothesis** for (SVM$_{s1}$) holds, then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = \delta \quad \text{and} \quad -w^\top v_{j_0} + b = \delta,$$

and we obtain the values of b and δ as

$$\begin{aligned} b &= \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0}) \\ \delta &= \frac{1}{2}(w^\top u_{i_0} - w^\top v_{j_0}). \end{aligned}$$

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K\}. \end{aligned}$$

Then it is easy to see that we can compute b and δ using the following averaging formulae:

$$\begin{aligned} b &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2 \\ \delta &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| - \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2. \end{aligned}$$

As we said earlier, the hypotheses of Theorem 50.17(2) hold, so *if* the primal problem (SVM$_{s1}$) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \delta, \lambda, \mu, \alpha, \beta, \gamma) = G(\lambda, \mu, \alpha, \beta, \gamma),$$

which means that

$$-\delta + K\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j \right) = - \left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2},$$

so we get

$$\delta = K\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j \right) + \left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2}.$$

Therefore, we confirm that $\delta \geq 0$.

It is important to note that the objective function of the dual program

$$-G(\lambda, \mu) = \left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2}$$

only involves the inner products of the u_i and the v_j through the matrix $X^\top X$, and similarly, the equation of the optimal hyperplane can be written as

$$\sum_{i=1}^p \lambda_i u_i^\top x - \sum_{j=1}^q \mu_j v_j^\top x - \left(\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right)^{1/2} b = 0,$$

an expression that only involves inner products of x with the u_i and the v_j and inner products of the u_i and the v_j.

As explained at the beginning of this chapter, this is a key fact that allows a generalization of the support vector machine using the method of *kernels*. We can define the following "kernelized" version of Problem (SVM$_{s1}$):

Soft margin kernel SVM (SVM$_{s1}$):

$$
\begin{aligned}
&\text{minimize} \quad -\delta + K\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right)\\
&\text{subject to}\\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q\\
&\qquad \langle w, w\rangle \leq 1.
\end{aligned}
$$

Tracing through the computation that led us to the dual program with u_i replaced by $\varphi(u_i)$ and v_j replaced by $\varphi(v_j)$, we find the following version of the dual program:

Dual of Soft margin kernel SVM (SVM$_{s1}$):

$$
\begin{aligned}
&\text{minimize} \quad \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \mathbf{K} \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0\\
&\qquad \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = 1\\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q,
\end{aligned}
$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given by

$$
\mathbf{K}_{ij} = \begin{cases}
\kappa(u_i, u_j) & 1 \leq i \leq p,\ 1 \leq j \leq q\\
-\kappa(u_i, v_{j-p}) & 1 \leq i \leq p,\ p+1 \leq j \leq p+q\\
-\kappa(v_{i-p}, u_j) & p+1 \leq i \leq p+q,\ 1 \leq j \leq p\\
\kappa(v_{i-p}, v_{j-q}) & p+1 \leq i \leq p+q,\ p+1 \leq j \leq p+q.
\end{cases}
$$

We also find that

$$
w = \frac{\displaystyle\sum_{i=1}^{p} \lambda_i \varphi(u_i) - \sum_{j=1}^{q} \mu_j \varphi(v_j)}{\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} K \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\right)^{1/2}}.
$$

Under the Standard Margin Hypothesis, there is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$, and we obtain the value of b and δ as

$$\begin{aligned} b &= \frac{1}{2}(\langle w, \varphi(u_{i_0}\rangle + \langle w, \varphi(v_{j_0})\rangle) \\ \delta &= \frac{1}{2}(\langle w, \varphi(u_{i_0})\rangle - \langle w, \varphi(v_{j_0})\rangle). \end{aligned}$$

Using the above value for w, we obtain

$$b = \frac{\sum_{i=1}^p \lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^q \mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))}{2\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} K \begin{pmatrix}\lambda \\ \mu\end{pmatrix}\right)^{1/2}}.$$

It follows that the classification function

$$f(x) = \operatorname{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$\begin{aligned} f(x) = \operatorname{sgn}\Bigg(&\sum_{i=1}^p \lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0})) \\ &- \sum_{j=1}^q \mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\Bigg), \end{aligned}$$

which is solely expressed in terms of the kernel κ.

Kernel methods for SVM are discussed in Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159].

54.2 Solving SVM (SVM$_{s1}$) Using ADMM

In order to solve (SVM$_{s1}$) using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$\begin{aligned} \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0 \\ \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j &= 1 \\ \lambda_i + \alpha_i &= K, \quad i = 1, \ldots, p \\ \mu_j + \beta_j &= K, \quad j = 1, \ldots, q. \end{aligned}$$

This is the $(p+q+2) \times 2(p+q)$ matrix A given by

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}.$$

We leave it as an exercise to prove that A has rank $p+q+2$. The right-hand side is

$$c = \begin{pmatrix} 0 \\ 1 \\ K\mathbf{1}_{p+q} \end{pmatrix}.$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = 2X^\top X, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $2(p+q)$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$, the $(p+q) \times (p+q)$ matrix $X^\top X$ must be augmented with zero's to make it a $2(p+q) \times 2(p+q)$ matrix P_a given by

$$P_a = \begin{pmatrix} X^\top X & 0_{p+q,p+q} \\ 0_{p+q,p+q} & 0_{p+q,p+q} \end{pmatrix},$$

and similarly q is augmented with zeros as the vector $q_a = 0_{2(p+q)}$.

Since the constraint $w^\top w \leq 1$ causes troubles, we trade it for a different objective function in which $-\delta$ is replaced by $(1/2)\left\|w\right\|_2^2$. This way we are left with purely affine constraints. In the next section we discuss a generalization of Problem (SVM$_{h2}$) obtained by adding a linear regularizing term.

54.3 Soft Margin Support Vector Machines; (SVM$_{s2}$)

In this section we consider the generalization of Problem (SVM$_{h2}$) where we minimize $(1/2)w^\top w$ by adding the "regularizing term" $K\Big(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j,\Big)$ for some $K > 0$. Recall that the margin δ is given by $\delta = 1/\left\|w\right\|$.

Soft margin SVM (SVM$_{s2}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^\top w + K\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

This is the classical problem discussed in all books on machine learning or pattern analysis, for instance Vapnik [182], Bishop [23], and Shawe–Taylor and Christianini [159]. The trivial solution where all variables are 0 is ruled out because of the presence of the 1 in the inequalities, but it is not clear that if (w, b, ϵ, ξ) is an optimal solution, then $w \neq 0$.

We prove that if the primal problem has an optimal solution (w, ϵ, ξ, b) with $w \neq 0$, then w is determined by any optimal solution (λ, μ) of the dual. We also prove that there is some i for which $\lambda_i > 0$ and some j for which $\mu_j > 0$. Under a mild hypothesis that we call the **Standard Margin Hypothesis**, b can be found.

Note that this framework is still somewhat sensitive to outliers because the penalty for misclassification is linear in ϵ and ξ.

First we write the constraints in matrix form. The $2(p+q) \times (n+p+q+1)$ matrix C is written in block form as

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} \end{pmatrix},$$

where X is the $n \times (p+q)$ matrix

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and the constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \end{pmatrix} \leq \begin{pmatrix} -\mathbf{1}_{p+q} \\ 0_{p+q} \end{pmatrix}.$$

The objective function $J(w, \epsilon, \xi, b)$ is given by

$$J(w, \epsilon, \xi, b) = \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \lambda, \mu, \alpha, \beta)$ with $\lambda, \alpha \in \mathbb{R}^p_+$ and with $\mu, \beta \in \mathbb{R}^q_+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \lambda, \mu, \alpha, \beta) &= \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\quad + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \mathbf{1}_{p+q}^\top & 0_{p+q}^\top \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \end{pmatrix}. \end{aligned}$$

Since

$$\begin{pmatrix} w^\top & \begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix} & b\end{pmatrix} C^\top \begin{pmatrix}\lambda \\ \mu \\ \alpha \\ \beta\end{pmatrix} = \begin{pmatrix} w^\top & \begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix} & b\end{pmatrix} \begin{pmatrix} X & & 0_{n,p+q} \\ -I_{p+q} & & -I_{p+q} \\ \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_{p+q}^\top \end{pmatrix} \begin{pmatrix}\lambda \\ \mu \\ \alpha \\ \beta\end{pmatrix},$$

we get

$$\begin{aligned}\begin{pmatrix} w^\top & \begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix} & b\end{pmatrix} C^\top \begin{pmatrix}\lambda \\ \mu \\ \alpha \\ \beta\end{pmatrix} &= \begin{pmatrix} w^\top & \begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix} & b\end{pmatrix} \begin{pmatrix} X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ -\begin{pmatrix}\lambda+\alpha \\ \mu+\beta\end{pmatrix} \\ \mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu \end{pmatrix} \\ &= w^\top X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \epsilon^\top(\lambda+\alpha) - \xi^\top(\mu+\beta) + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu),\end{aligned}$$

and since

$$\begin{pmatrix}\mathbf{1}_{p+q}^\top & 0_{p+q}^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu \\ \alpha \\ \beta\end{pmatrix} = \mathbf{1}_{p+q}^\top\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q},$$

the Lagrangian can be rewritten as

$$\begin{aligned}L(w,\epsilon,\xi,b,\lambda,\mu,\alpha,\beta) &= \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} + \epsilon^\top(K\mathbf{1}_p - (\lambda+\alpha)) + \xi^\top(K\mathbf{1}_q - (\mu+\beta)) \\ &\quad + b(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q}.\end{aligned}$$

To find the dual function $G(\lambda,\mu,\alpha,\beta)$ we minimize $L(w,\epsilon,\xi,b,\lambda,\mu,\alpha,\beta)$ with respect to w,ϵ,ξ and b. Since the Lagrangian is convex and $(w,\epsilon,\xi,b)\in\mathbb{R}^n\times\mathbb{R}^p\times\mathbb{R}^q\times\mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in (w,ϵ,ξ,b) iff $\nabla L_{w,\epsilon,\xi,b}=0$, so we compute its gradient with respect to w,ϵ,ξ and b, and we get

$$\nabla L_{w,\epsilon,\xi,b} = \begin{pmatrix} w + X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ K\mathbf{1}_p - (\lambda+\alpha) \\ K\mathbf{1}_q - (\mu+\beta) \\ \mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b}=0$ we get the equations

$$\begin{aligned} w &= -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \qquad (*_w) \\ \lambda+\alpha &= K\mathbf{1}_p \\ \mu+\beta &= K\mathbf{1}_q \\ \mathbf{1}_p^\top\lambda &= \mathbf{1}_q^\top\mu. \end{aligned}$$

The first and the fourth equation are identical to the Equations $(*_1)$ and $(*_2)$ that we obtained in Example 50.10. Since $\lambda, \mu, \alpha, \beta \geq 0$, the second and the third equation are equivalent to the box constraints

$$0 \leq \lambda_i, \mu_j \leq K, \quad i = 1, \ldots, p,\ j = 1, \ldots, q.$$

Using the equations that we just derived, after simplifications we get

$$G(\lambda, \mu, \alpha, \beta) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q},$$

which is independent of α and β and is identical to the dual function obtained in $(*_4)$ of Example 50.10. To be perfectly rigorous,

$$G(\lambda, \mu) = \begin{cases} -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} & \text{if } \begin{cases} \sum_{i=1}^p \lambda_i = \sum_{j=1}^q \mu_j \\ 0 \leq \lambda_i \leq K,\ i = 1, \ldots, p \\ 0 \leq \mu_j \leq K,\ j = 1, \ldots, q \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

As in Example 50.10, the the dual program can be formulated as

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\ &\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ &\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q, \end{aligned}$$

or equivalently

Dual of Soft margin SVM (SVM_{s2}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\ &\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\ &\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q. \end{aligned}$$

If (w, ϵ, ξ, b) is an optimal solution of Problem (SVM_{s2}), then the complementary slackness conditions yield a classification of the points u_i and v_j in terms of the values of λ and μ.

Indeed, we have $\epsilon_i\alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j\beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K$, it follows that $\epsilon_i\alpha_i = 0$ iff $\epsilon_i(K - \lambda_i) = 0$, and since $\mu_j + \beta_j = K$, we have $\xi_j\beta_j = 0$ iff $\xi_j(K - \mu_j) = 0$. Thus if $\epsilon_i > 0$, then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$, then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$, then $\xi_j = 0$ and v_j is correctly classified.

We have a classification of the points u_i and v_j in terms of λ and μ obtained from the classification given in Section 54.1 by replacing δ with 1. Since it is so similar, it is omitted. Let us simply recall that the vectors u_i on the blue margin and the vectors v_j on the red margin are called *support vectors*; these are the vectors u_i for which $w^\top u_i - b - 1 = 0$ (which implies $\epsilon_i = 0$), and the vectors v_j for which $w^\top v_j - b + 1 = 0$ (which implies $\xi_j = 0$). Those support vectors u_i such that $\lambda_i = 0$ and those support vectors such that $\mu_j = 0$ are called *exceptional support vectors*.

We also have the following classification of the points u_i and v_j terms of ϵ_i (or ξ_j) obtained by replacing δ with 1.

(1) If $\epsilon_i > 0$, then by complementary slackness $\lambda_i = K$, so the ith equation is active and by (2) above,
$$w^\top u_i - b - 1 = -\epsilon_i.$$
Since $\epsilon_i > 0$, the point u_i is within the open half space bounded by the blue margin hyperplane $H_{w,b+1}$ and containing the separating hyperplane $H_{w,b}$; if $\epsilon_i \leq 1$, then u_i is classified correctly, and if $\epsilon_i > 1$, then u_i is misclassified.

Similarly, if $\xi_j > 0$, then v_j is within the open half space bounded by the red margin hyperplane $H_{w,b-1}$ and containing the separating hyperplane $H_{w,b}$; if $\xi_j \leq 1$, then v_j is classified correctly, and if $\xi_j > 1$, then v_j is misclassified.

(2) If $\epsilon_i = 0$, then the point u_i is correctly classified. If $\lambda_i = 0$, then by (3) above, u_i is in the closed half space on the blue side bounded by the blue margin hyperplane $H_{w,b+\eta}$. If $\lambda_i > 0$, then by (1) and (2) above, the point u_i is on the blue margin.

Similarly, if $\xi_j = 0$, then the point v_j is correctly classified. If $\mu_j = 0$, then v_j is in the closed half space on the red side bounded by the red margin hyperplane $H_{w,b-\eta}$, and if $\mu_j > 0$, then the point v_j is on the red margin. See Figure 54.5 (3).

Vectors u_i for which $\lambda_i = K$ and vectors v_j such that $\xi_j = K$ are said to *fail the margin*.

It is shown in Section 54.4 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ.

Remark: The hard margin Problem (SVM$_{h2}$) corresponds to the special case of Problem (SVM$_{s2}$) in which $\epsilon = 0$, $\xi = 0$, and $K = +\infty$. Indeed, in Problem (SVM$_{h2}$) the terms involving ϵ and ξ are missing from the Lagrangian and the effect is that the box constraints are missing; we simply have $\lambda_i \geq 0$ and $\mu_j \geq 0$.

We can use the dual program to solve the primal. Once $\lambda \geq 0, \mu \geq 0$ have been found, w is given by

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j.$$

To find b we use the complementary slackness conditions.

If the primal has a solution $w \neq 0$, then the equation

$$w = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j$$

implies that either there is some index i_0 such that $\lambda_{i_0} > 0$ or there is some index j_0 such that $\mu_{j_0} > 0$. The constraint

$$\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0$$

implies that there is some index i_0 such that $\lambda_{i_0} > 0$ and there is some index j_0 such that $\mu_{j_0} > 0$. However, a priori, nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. Observe that the equation

$$\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0$$

implies that if there is some index i_0 such that $0 < \lambda_{i_0} < K$, then there is some index j_0 such that $0 < \mu_{j_0} < K$, and vice-versa. If the following mild hypothesis holds, then b can be found.

Standard Margin Hypothesis for (SVM$_{s2}$). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is a support vector of type 1 on the blue margin, and some v_{j_0} is a support vector of type 1 on the red margin.

If the **Standard Margin Hypothesis** for (SVM$_{s2}$) holds, then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = 1 \quad \text{and} \quad -w^\top v_{j_0} + b = 1,$$

and we obtain

$$b = \frac{1}{2}(w^\top u_{i_0} + w^\top v_{j_0}).$$

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K\}. \end{aligned}$$

Then it is easy to see that we can compute b using the following averaging formula

$$b = w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| + \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu| \right) /2.$$

Recall that $\delta = 1/\left\| w \right\|$.

Remark: There is a cheap version of Problem (SVM$_{s2}$) which consists in dropping the term $(1/2)w^\top w$ from the objective function:

Soft margin classifier (SVM$_{s2l}$):

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad - w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

The above program is a linear program that minimizes the number of misclassified points but does not care about enforcing a minimum margin. An example of its use is given in Boyd and Vandenberghe; see [29], Section 8.6.1.

The "kernelized" version of Problem (SVM$_{s2}$) is the following:

Soft margin kernel SVM (SVM$_{s2}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \langle w, \varphi(u_i)\rangle - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad - \langle w, \varphi(v_j)\rangle + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q. \end{aligned}$$

Redoing the computation of the dual function, we find that the dual program is given by

Dual of Soft margin kernel SVM (SVM$_{s2}$):

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{K} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\
\text{subject to} \quad & \\
& \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
& 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
& 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q,
\end{aligned}
$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given at the end of Section 54.1. We also find that

$$w = \sum_{i=1}^p \lambda_i \varphi(u_i) - \sum_{j=1}^q \mu_j \varphi(v_j),$$

so

$$b = \frac{1}{2}\Bigg(\sum_{i=1}^p \lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^q \mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))\Bigg),$$

and the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x) \rangle - b)$$

is given by

$$
\begin{aligned}
f(x) = \mathrm{sgn}\Bigg(&\sum_{i=1}^p \lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0})) \\
&- \sum_{j=1}^q \mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\Bigg).
\end{aligned}
$$

54.4 Solving SVM (SVM$_{s2}$) Using ADMM

In order to solve (SVM$_{s2}$) using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$
\begin{aligned}
\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0 \\
\lambda_i + \alpha_i &= K, \quad i = 1, \ldots, p \\
\mu_j + \beta_j &= K, \quad j = 1, \ldots, q.
\end{aligned}
$$

This is the $(p+q+1) \times 2(p+q)$ matrix A given by

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}.$$

We leave it as an exercise to prove that A has rank $p+q+1$. The right-hand side is

$$c = \begin{pmatrix} 0 \\ K\mathbf{1}_{p+q} \end{pmatrix}.$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = -\mathbf{1}_{p+q}.$$

Since there are $2(p+q)$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$, the $(p+q)\times(p+q)$ matrix $X^\top X$ must be augmented with zero's to make it a $2(p+q)\times 2(p+q)$ matrix P_a given by

$$P_a = \begin{pmatrix} X^\top X & 0_{p+q,p+q} \\ 0_{p+q,p+q} & 0_{p+q,p+q} \end{pmatrix},$$

and similarly q is augmented with zeros as the vector

$$q_a = \begin{pmatrix} -\mathbf{1}_{p+q} \\ 0_{p+q}. \end{pmatrix}$$

54.5 Soft Margin Support Vector Machines; $(\mathrm{SVM}_{s2'})$

In this section we consider a generalization of Problem (SVM_{s2}) for a version of the soft margin SVM coming from Problem (SVM_{h2}) by adding an extra degree of freedom, namely instead of the margin $\delta = 1/\|w\|$, we use the margin $\delta = \eta/\|w\|$ where η is some positive constant that we wish to maximize. To do so, we add a term $-K_m\eta$ to the objective function $(1/2)w^\top w$ as well as the "regularizing term" $K_s\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\right)$ whose purpose is to make ϵ and ξ sparse, where $K_m > 0$ (m refers to margin) and $K_s > 0$ (s refers to sparse) are fixed constants that can be adjusted to determine the influence of η and the regularizing term.

Soft margin SVM $(\mathrm{SVM}_{s2'})$:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w - K_m\eta + K_s\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1,\ldots,p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1,\ldots,q\\
&\qquad \eta \geq 0.
\end{aligned}$$

This version of the SVM problem was first discussed in Schölkopf, Smola, Williamson, and Bartlett [147] under the name of *ν-SVC* (or *ν-SVM*), and also used in Schölkopf, Platt, Shawe–Taylor, and Smola [146]. The ν-SVC method is also presented in Schölkopf and Smola [145] (which contains much more). The difference between the ν-SVC method and the method presented in Section 54.3, sometimes called the C-SVM method, was thoroughly investigated by Chan and Lin [36].

For this problem it is no longer clear that if $(w, \eta, b, \epsilon, \xi)$ is an optimal solution, then $w \neq 0$ and $\eta > 0$. In fact, if the sets of points are not linearly separable and if K_s is chosen too big, Problem $(\mathrm{SVM}_{s2'})$ may fail to have an optimal solution.

We show that in order for the problem to have a solution we must pick K_m and K_s so that

$$K_m \leq \min\{2pK_s, 2qK_s\}.$$

If we define ν by

$$\nu = \frac{K_m}{(p+q)K_s},$$

then $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{\frac{2p}{p+q}, \frac{2q}{p+q}\right\} \leq 1.$$

The reason for introducing ν is that $\nu(p+q)/2$ can be interpreted as the maximum number of points failing to achieve the margin $\delta = \eta/\|w\|$. We will show later that if the points u_i and v_j are not separable, then we must pick ν so that $\nu \geq 2/(p+q)$ for the method to have a solution for which $w \neq 0$ and $\eta > 0$.

The objective function of our problem is convex and the constraints are affine. Consequently, by Theorem 50.17(2) *if* the Primal Problem $(\mathrm{SVM}_{s2'})$ has an optimal solution, *then* the dual problem has a solution too, and the duality gap is zero. This does not immediately imply that an optimal solution of the dual yields an optimal solution of the primal because the hypotheses of Theorem 50.17(1) fail to hold.

We show that if the primal problem has an optimal solution $(w, \eta, \epsilon, \xi, b)$ with $w \neq 0$, then any optimal solution of the dual problem determines λ and μ, which in turn determine

w *via* the equation

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j, \qquad (*_w)$$

and $\eta \geq 0$.

It remains to determine b, η, ϵ and ξ. The solution of the dual does not determine b, η, ϵ, ξ directly, and we are not aware of necessary and sufficient conditions that ensure that they can be determined. The best we can do is to use the KKT conditions.

The simplest sufficient condition is what we call the

Standard Margin Hypothesis for $(\mathrm{SVM}_{s2'})$: There is some i_0 such that $0 < \lambda_{i_0} < K_s$, and there is some μ_{j_0} such that $0 < \mu_{j_0} < K_s$. This means that there is some support vector u_{i_0} of type 1 *and* there is some support vector v_{j_0} of type 1.

In this case, then by complementary slackness, it can be shown that $\epsilon_{i_0} = 0$, $\xi_{i_0} = 0$, and the corresponding inequalities are active, that is we have the equations

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

so we can solve for b and η. Then since by complementary slackness, if $\epsilon_i > 0$, then $\lambda_i = K_s$ and if $\xi_j > 0$, then $\mu_j = K_s$, all inequalities corresponding to such $\epsilon_i > 0$ and $\mu_j > 0$ are active, and we can solve for ϵ_i and ξ_j.

The linear constraints are given by the $(2(p+q)+1) \times (n+p+q+2)$ matrix given in block form by

$$C = \begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} & \mathbf{1}_{p+q} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \\ 0_n^\top & 0_{p+q}^\top & 0 & -1 \end{pmatrix},$$

where X is the $n \times (p+q)$ matrix

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and the linear constraints are expressed by

$$\begin{pmatrix} X^\top & -I_{p+q} & \begin{matrix} \mathbf{1}_p \\ -\mathbf{1}_q \end{matrix} & \mathbf{1}_{p+q} \\ 0_{p+q,n} & -I_{p+q} & 0_{p+q} & 0_{p+q} \\ 0_n^\top & 0_{p+q}^\top & 0 & -1 \end{pmatrix} \begin{pmatrix} w \\ \epsilon \\ \xi \\ b \\ \eta \end{pmatrix} \leq \begin{pmatrix} 0_{p+q} \\ 0_{p+q} \\ 0 \end{pmatrix}.$$

The objective function is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}.$$

The Lagrangian $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$, and $\gamma \in \mathbb{R}_+$ is given by

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) &= \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\quad + \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \eta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix}. \end{aligned}$$

Since

$$\begin{aligned} \begin{pmatrix} w^\top & \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} & b & \eta \end{pmatrix} C^\top \begin{pmatrix} \lambda \\ \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix} &= w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) \\ &\quad + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) - \gamma\eta, \end{aligned}$$

the Lagrangian can be written as

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) &= \frac{1}{2} w^\top w - K_m \eta + K_s(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \epsilon^\top(\lambda + \alpha) \\ &\quad - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) - \gamma\eta \\ &= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + (\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - K_m - \gamma)\eta \\ &\quad + \epsilon^\top(K_s \mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K_s \mathbf{1}_q - (\mu + \beta)) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$ we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute its gradient with respect to $w, \epsilon, \xi, b, \eta$, and we get

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + w \\ K_s \mathbf{1}_p - (\lambda + \alpha) \\ K_s \mathbf{1}_q - (\mu + \beta) \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - K_m - \gamma \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \tag{$*_w$}$$

$$\begin{aligned}\lambda + \alpha &= K_s \mathbf{1}_p \\ \mu + \beta &= K_s \mathbf{1}_q \\ \mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu,\end{aligned}$$

and

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu = K_m + \gamma. \tag{$*_\gamma$}$$

The second and third equations are equivalent to the box constraints

$$0 \leq \lambda_i, \mu_j \leq K_s, \quad i = 1, \ldots, p, \ j = 1, \ldots, q,$$

and since $\gamma \geq 0$ equation $(*_\gamma)$ is equivalent to

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu \geq K_m.$$

Plugging back w from $(*_w)$ into the Lagrangian, after simplifications we get

$$\begin{aligned}G(\lambda, \mu, \alpha, \beta) &= \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},\end{aligned}$$

so the dual function is independent of α, β and is given by

$$G(\lambda, \mu) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

The dual program is given by

$$\begin{aligned}&\text{maximize} \quad -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\ &\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m \\ &\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\ &\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.\end{aligned}$$

Finally, the dual program is equivalent to the following minimization program:

Dual of Soft margin SVM $(\mathrm{SVM}_{s2'})$:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0\\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

If $(w, \eta, \epsilon, \xi, b)$ is an optimal solution of Problem $(\mathrm{SVM}_{s2'})$ with $w \neq 0$ and $\eta \neq 0$, then the complementary slackness conditions yield a classification of the points u_i and v_j in terms of the values of λ and μ. Indeed, we have $\epsilon_i \alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j \beta_j = 0$ for $j = 1, \ldots, q$. Also, if $\lambda_i > 0$, then the corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K_s$, it follows that $\epsilon_i \alpha_i = 0$ iff $\epsilon_i(K_s - \lambda_i) = 0$, and since $\mu_j + \beta_j = K_s$, we have $\xi_j \beta_j = 0$ iff $\xi_j(K_s - \mu_j) = 0$. Thus if $\epsilon_i > 0$, then $\lambda_i = K_s$, and if $\xi_j > 0$, then $\mu_j = K_s$. Consequently, if $\lambda_i < K_s$, then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K_s$, then $\xi_j = 0$ and v_j is correctly classified.

We have the following classification which is basically the classification given in Section 54.1 obtained by replacing δ with η (recall that $\eta > 0$ and $\delta = \eta / \|w\|$) .

(1) If $0 < \lambda_i < K_s$, then $\epsilon_i = 0$ and the i-th inequality is active, so

$$w^\top u_i - b - \eta = 0.$$

This means that u_i is on the blue margin (the hyperplane $H_{w,b+\eta}$ of equation $w^\top x = b + \eta$) and is classified correctly.

Similarly, if $0 < \mu_j < K_s$, then $\xi_j = 0$ and

$$w^\top v_j - b + \eta = 0,$$

so v_j is on the red margin (the hyperplane $H_{w,b-\eta}$ of equation $w^\top x = b - \eta$) and is classified correctly.

(2) If $\lambda_i = K_s$, then the i-th inequality is active, so

$$w^\top u_i - b - \eta = -\epsilon_i.$$

If $\epsilon_i = 0$, then the point u_i is on the blue margin. If $\epsilon_i > 0$, then u_i is within the open half space bounded by the blue margin hyperplane $H_{w,b+\eta}$ and containing the separating hyperplane $H_{w,b}$; if $\epsilon_i \leq \eta$, then u_i is classified correctly, and if $\epsilon_i > \eta$, then u_i is misclassified (u_i lies on the wrong side of the separating hyperplane, the red side).

Similarly, if $\mu_j = K_s$, then

$$w^\top v_j - b + \eta = \xi_j.$$

If $\xi_j = 0$, then the point v_j is on the red margin. If $\xi_j > 0$, then v_j is within the open half space bounded by the red margin hyperplane $H_{w,b-\eta}$ and containing the separating hyperplane $H_{w,b}$; if $\xi_j \leq \eta$, then v_j is classified correctly, and if $\xi_j > \eta$, then v_j is misclassified (v_j lies on the wrong side of the separating hyperplane, the blue side).

(3) If $\lambda_i = 0$, then $\epsilon_i = 0$ and the i-th inequality may or may not be active, so

$$w^\top u_i - b - \eta \geq 0.$$

Thus u_i is in the closed half space on the blue side bounded by the blue margin hyperplane $H_{w,b+\eta}$ (of course, classified correctly).

Similarly, if $\mu_j = 0$, then

$$w^\top v_j - b + \eta \leq 0$$

and v_j is in the closed half space on the red side bounded by the red margin hyperplane $H_{w,b-\eta}$ (of course, classified correctly).

Definition 54.2. The vectors u_i on the blue margin $H_{w,b+\eta}$ and the vectors v_j on the red margin $H_{w,b-\eta}$ are called *support vectors*. Support vectors correspond to vectors u_i for which $w^\top u_i - b - \eta = 0$ (which implies $\epsilon_i = 0$), and vectors v_j for which $w^\top v_j - b + \eta = 0$ (which implies $\xi_j = 0$). Support vectors u_i such that $0 < \lambda_i < K_s$ and support vectors v_j such that $0 < \mu_j < K_s$ are *support vectors of type 1*. Support vectors of type 1 play a special role so we denote the sets of indices associated with them by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}. \end{aligned}$$

We denote their cardinalities by $numsvl_1 = |I_\lambda|$ and $numsvm_1 = |I_\mu|$. Support vectors u_i such that $\lambda_i = K_s$ and support vectors v_j such that $\mu_j = K_s$ are *support vectors of type 2*. Those support vectors u_i such that $\lambda_i = 0$ and those support vectors v_j such that $\mu_j = 0$ are called *exceptional support vectors*.

The vectors u_i for which $\lambda_i = K_s$ and the vectors v_j for which $\mu_j = K_s$ are said to *fail the margin*. The sets of indices associated with the vectors failing the margin are denoted by

$$\begin{aligned} K_\lambda &= \{i \in \{1, \ldots, p\} \mid \lambda_i = K_s\} \\ K_\mu &= \{j \in \{1, \ldots, q\} \mid \mu_j = K_s\}. \end{aligned}$$

We denote their cardinalities by $p_f = |K_\lambda|$ and $q_f = |K_\mu|$.

It will also be useful to understand how points are classified in terms of ϵ_i (or ξ_j).

(1) If $\epsilon_i > 0$, then by complementary slackness $\lambda_i = K_s$, so the ith equation is active and by (2) above,

$$w^\top u_i - b - \eta = -\epsilon_i.$$

Since $\epsilon_i > 0$, the point u_i is strictly within the open half space bounded by the blue margin hyperplane $H_{w,b+\eta}$ and containing the separating hyperplane $H_{w,b}$ (excluding the blue hyperplane $H_{w,b+\eta}$); if $\epsilon_i \leq \eta$, then u_i is classified correctly, and if $\epsilon_i > \eta$, then u_i is misclassified.

Similarly, if $\xi_j > 0$, then v_j is strictly within the open half space bounded by the red margin hyperplane $H_{w,b-\eta}$ and containing the separating hyperplane $H_{w,b}$ (excluding the red hyperplane $H_{w,b-\eta}$); if $\xi_j \leq \eta$, then v_j is classified correctly, and if $\xi_j > \eta$, then v_j is misclassified.

(2) If $\epsilon_i = 0$, then the point u_i is correctly classified. If $\lambda_i = 0$, then by (3) above, u_i is in the closed half space on the blue side bounded by the blue margin hyperplane $H_{w,b+\eta}$. If $\lambda_i > 0$, then by (1) and (2) above, the point u_i is on the blue margin.

Similarly, if $\xi_j = 0$, then the point v_j is correctly classified. If $\mu_j = 0$, then v_j is in the closed half space on the red side bounded by the red margin hyperplane $H_{w,b-\eta}$, and if $\mu_j > 0$, then the point v_j is on the red margin.

Also observe that if $\lambda_i > 0$, then u_i is in the closed half space bounded by the blue hyperplane $H_{w,b+\eta}$ and containing the separating hyperplane $H_{w,b}$ (including the blue hyperplane $H_{w,b+\eta}$).

Similarly, if $\mu_j > 0$, then v_j is in the closed half space bounded by the red hyperplane $H_{w,b+\eta}$ and containing the separating hyperplane $H_{w,b}$ (including the red hyperplane $H_{w,b+\eta}$).

Definition 54.3. Vectors u_i such that $\lambda_i > 0$ and vectors v_j such that $\mu_j > 0$ are said to *have margin at most* δ. The sets of indices associated with these vectors are denoted by

$$\begin{aligned} I_{\lambda>0} &= \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \\ I_{\mu>0} &= \{j \in \{1, \ldots, q\} \mid \mu_j > 0\}. \end{aligned}$$

We denote their cardinalities by $p_m = |I_{\lambda>0}|$ and $q_m = |I_{\mu>0}|$.

Vectors u_i such that $\epsilon_i > 0$ and vectors v_j such that $\xi_j > 0$ are said to *strictly fail the margin*. The corresponding sets of indices are denoted by

$$\begin{aligned} E_\lambda &= \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\} \\ E_\mu &= \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}. \end{aligned}$$

We write $p_{sf} = |E_\lambda|$ and $q_{sf} = |E_\mu|$.

We have the inclusions $E_\lambda \subseteq K_\lambda$ and $E_\mu \subseteq K_\mu$. The difference between the first sets and the second sets is that the second sets may contain support vectors of type 2 such that $\lambda_i = K_s$ and $\epsilon_i = 0$, or $\mu_j = K_s$ and $\xi_j = 0$. We also have the equations $I_\lambda \cup (K_\lambda - E_\lambda) \cup E_\lambda = I_{\lambda>0}$ and $I_\mu \cup (K_\mu - E_\mu) \cup E_\mu = I_{\mu>0}$, and the inequalities $p_{sf} \leq p_f \leq p_m$ and $q_{sf} \leq q_f \leq q_m$.

The blue points u_i of index $i \in I_{\lambda>0}$ are classified as follows:

(1) If $i \in I_\lambda$, then u_i is a support vector of type 1 ($\lambda_i < K_s$).

(2) If $i \in K_\lambda - E_\lambda$, then u_i is a support vector of type 2 ($\lambda_i = K_s$).

(3) If $i \in E_\lambda$, then u_i strictly fails the margin, that is $\epsilon_i > 0$.

Similarly the red points v_j of index $j \in I_{\mu>0}$ are classified as follows:

(1) If $j \in I_\mu$, then v_j is a support vector of type 1 ($\mu_j < K_s$).

(2) If $j \in K_\mu - E_\mu$, then v_j is a support vector of type 2 ($\mu_j = K_s$).

(3) If $j \in E_\mu$, then v_j strictly fails the margin, that is $\xi_j > 0$.

Note that $p_m - p_f$ is the number of blue support vectors of type 1 and $q_m - q_f$ is the number of red support vectors of type 1. The remaining blue points u_i for which $\lambda_i = 0$ are either exceptional support vectors or they are (strictly) in the open half-space corresponding to the blue side. Similarly, the remaining red points v_j for which $\mu_j = 0$ are either exceptional support vectors or they are (strictly) in the open half-space corresponding to the red side.

It is shown in Section 54.8 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ. Once a solution for λ and μ is obtained, we have

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j.$$

As we said earlier, the hypotheses of Theorem 50.17(2) hold, so *if* the primal problem ($\mathrm{SVM}_{s2'}$) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) = G(\lambda, \mu, \alpha, \beta, \gamma),$$

which means that

$$\frac{1}{2} w^\top w - K_m \eta + K_s \left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j \right) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

and since

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

we get

$$\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - K_m\eta + K_s\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\right) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

which yields

$$\eta = \frac{K_s}{K_m}\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\right) + \frac{1}{K_m}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}. \tag{$*$}$$

Therefore, we confirm that $\eta \geq 0$.

Remarks: Since we proved that if the Primal Problem (SVM$_{s2'}$) has an optimal solution with $w \neq 0$, then $\eta \geq 0$, one might wonder why the constraint $\eta \geq 0$ was included. If we delete this constraint, it is easy to see that the only difference is that instead of the equation

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu = K_m + \gamma \tag{$*_1$}$$

we obtain the equation

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu = K_m. \tag{$*_2$}$$

If $\eta > 0$, then by complementary slackness $\gamma = 0$, in which case $(*_1)$ and $(*_2)$ are equivalent. But if $\eta = 0$, then γ could be strictly positive.

The option to omit the constraint $\eta \geq 0$ in the primal is slightly advantageous because then the dual involves $2(p+q)$ instead of $2(p+q)+1$ Lagrange multipliers, so the constraint matrix is a $(p+q+2) \times 2(p+q)$ matrix instead of a $(p+q+2) \times (2(p+q)+1)$ matrix and the matrix defining the quadratic functional is a $2(p+q) \times 2(p+q)$ matrix instead of a $(2(p+q)+1) \times (2(p+q)+1)$ matrix; see Section 54.8.

Under the **Standard Margin Hypothesis** for (SVM$_{s2'}$), there is some i_0 such that $0 < \lambda_{i_0} < K_s$ and some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ and $\xi_{j_0} = 0$, so we have the two active constraints

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for b and η and we get

$$\begin{aligned} b &= \frac{w^\top u_{i_0} + w^\top v_{j_0}}{2} \\ \eta &= \frac{w^\top u_{i_0} - w^\top v_{j_0}}{2} \\ \delta &= \frac{\eta}{\|w\|}. \end{aligned}$$

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}. \end{aligned}$$

Then it is easy to see that we can compute b and η using the following averaging formulae:

$$b = w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| + \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu| \right) /2$$
$$\eta = w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| - \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu| \right) /2.$$

The "kernelized" version of Problem (SVM$_{s2'}$) is the following:

Soft margin kernel SVM (SVM$_{s2'}$):

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2}\langle w, w\rangle - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ & \text{subject to} \\ & \qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ & \qquad - \langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\ & \qquad \eta \geq 0. \end{aligned}$$

Tracing through the derivation of the dual program we obtain

Dual of the Soft margin kernel SVM (SVM$_{s2'}$):

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{K} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ & \text{subject to} \\ & \qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\ & \qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m \\ & \qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\ & \qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q, \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

As in Section 54.3, we obtain

$$w = \sum_{i=1}^p \lambda_i \varphi(u_i) - \sum_{j=1}^q \mu_j \varphi(v_j),$$

so

$$b = \frac{1}{2}\bigg(\sum_{i=1}^{p} \lambda_i(\kappa(u_i, u_{i_0}) + \kappa(u_i, v_{j_0})) - \sum_{j=1}^{q} \mu_j(\kappa(v_j, u_{i_0}) + \kappa(v_j, v_{j_0}))\bigg),$$

and the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$\begin{aligned} f(x) = \mathrm{sgn}\bigg(&\sum_{i=1}^{p} \lambda_i(2\kappa(u_i, x) - \kappa(u_i, u_{i_0}) - \kappa(u_i, v_{j_0})) \\ &- \sum_{j=1}^{q} \mu_j(2\kappa(v_j, x) - \kappa(v_j, u_{i_0}) - \kappa(v_j, v_{j_0}))\bigg). \end{aligned}$$

54.6 Classification of the Data Points in Terms of ν (SVM$_{s2'}$)

For a finer classification of the points it turns out to be convenient to consider the ratio

$$\nu = \frac{K_m}{(p+q)K_s}.$$

First note that in order for the constraints to be satisfied, some relationship between K_s and K_m must hold. In addition to the constraints

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s,$$

we also have the constraints

$$\begin{aligned} \sum_{i=1}^{p} \lambda_i &= \sum_{j=1}^{q} \mu_j \\ \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j &\geq K_m \end{aligned}$$

which imply that

$$\sum_{i=1}^{p} \lambda_i \geq \frac{K_m}{2} \quad \text{and} \quad \sum_{j=1}^{q} \mu_j \geq \frac{K_m}{2}. \tag{$\dagger$}$$

Since λ, μ are all nonnegative, if $\lambda_i = K_s$ for all i and if $\mu_j = K_s$ for all j, then

$$\frac{K_m}{2} \leq \sum_{i=1}^{p} \lambda_i \leq pK_s \quad \text{and} \quad \frac{K_m}{2} \leq \sum_{j=1}^{q} \mu_j \leq qK_s,$$

so these constraints are not satisfied unless $K_m \leq \min\{2pK_s, 2qK_s\}$, so we assume that $K_m \leq \min\{2pK_s, 2qK_s\}$. The equations in (†) also imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, and so $p_m \geq 1$ and $q_m \geq 1$.

For a finer classification of the points we find it convenient to define $\nu > 0$ such that

$$\nu = \frac{K_m}{(p+q)K_s},$$

so that the objective function $J(w, \epsilon, \xi, b, \eta)$ is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2}w^\top w + (p+q)K_s \left(-\nu\eta + \frac{1}{p+q} \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \right).$$

Observe that the condition $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{\frac{2p}{p+q}, \frac{2q}{p+q}\right\} \leq 1.$$

Since we obtain an equivalent problem by rescaling by a common positive factor, theoretically it is convenient to normalize K_s as

$$K_s = \frac{1}{p+q},$$

in which case $K_m = \nu$. This method is called the *ν-support vector machine*. Actually, to program the method, it may be more convenient assume that K_s is arbitrary. This helps in avoiding λ_i and μ_j to become to small when $p+q$ is relatively large.

The equations (†) and the box inequalities

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s$$

also imply the following facts:

Proposition 54.1. *If Problem* (SVM$_{s2'}$) *has an optimal solution with $w \neq 0$ and $\eta > 0$, then the following facts hold:*

(1) Let p_f be the number of points u_i such that $\lambda_i = K_s$, and let q_f the number of points v_j such that $\mu_j = K_s$. Then $p_f, q_f \leq \nu(p+q)/2$.

(2) Let p_m be the number of points u_i such that $\lambda_i > 0$, and let q_m the number of points v_j such that $\mu_j > 0$. Then $p_m, q_m \geq \nu(p+q)/2$. We have $p_m \geq 1$ and $q_m \geq 1$.

(3) If $p_f \geq 1$ or $q_f \geq 1$, then $\nu \geq 2/(p+q)$.

Proof. (1) Recall that for an optimal solution with $w \neq 0$ and $\eta > 0$, we have $\gamma = 0$, so by $(*_\gamma)$ we have the equations

$$\sum_{i=1}^{p} \lambda_i = \frac{K_m}{2} \quad \text{and} \quad \sum_{j=1}^{q} \mu_j = \frac{K_m}{2}.$$

The point u_i fails to achieve the margin iff $\lambda_i = K_s = K_m/(\nu(p+q))$, so if there are p_f such points then

$$\frac{K_m}{2} = \sum_{i=1}^{p} \lambda_i \geq \frac{K_m p_f}{\nu(p+q)},$$

so

$$p_f \leq \frac{\nu(p+q)}{2}.$$

A similar reasoning applies if v_j fails to achieve the margin δ with $\sum_{i=1}^{p} \lambda_i$ replaced by $\sum_{j=1}^{q} \mu_j$.

(2) A point u_i has margin at most δ iff $\lambda_i > 0$. If

$$I_{\lambda>0} = \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \quad \text{and} \quad p_m = |I_{\lambda>0}|,$$

then

$$\frac{K_m}{2} = \sum_{i=1}^{p} \lambda_i = \sum_{i \in I_{\lambda>0}} \lambda_i,$$

and since $\lambda_i \leq K_s = K_m/(\nu(p+q))$, we have

$$\frac{K_m}{2} = \sum_{i \in I_{\lambda>0}} \lambda_i \leq \frac{K_m p_m}{\nu(p+q)},$$

which yields

$$p_m \geq \frac{\nu(p+q)}{2}.$$

A similar reasoning applies if a point v_j has margin at most δ. We already observed that (†) implies that $p_m \geq 1$ and $q_m \geq 1$.

(3) This follows immediately from (1). $\square$

Observe that $p_f = q_f = 0$ means that there are no points in the open slab containing the separating hyperplane, namely, the points u_i and the points v_j are separable. So if the points u_i and the points v_j are not separable, then we must pick ν such that $2/(p+q) \leq \nu \leq \min\{2p/(p+q), 2q/(p+q)\}$ for the method to succeed. Otherwise, the method is trying to produce a solution where $w = 0$ and $\eta = 0$, and it does not converge (γ is nonzero). Actually, Proposition 54.1 yields more accurate bounds on ν for the method to converge, namely

$$\max\left\{\frac{2p_f}{p+q}, \frac{2q_f}{p+q}\right\} \leq \nu \leq \min\left\{\frac{2p_m}{p+q}, \frac{2q_m}{p+q}\right\}.$$

By a previous remark, $p_f \leq p_m$ and $q_f \leq q_m$, the first inequality being strict if there is some i such that $0 < \lambda_i < K$, and the second inequality being strict if there is some j such that $0 < \mu_j < K$. This will be the case under the **Standard Margin Hypothesis**.

Observe that a small value of ν keeps p_f and q_f small, which is achieved if the δ-slab is narrow (to avoid having points on the wrong sides of the margin hyperplanes). A large value of ν allows p_m and q_m to be fairly large, which is achieved if the δ-slab is wide. Thus the smaller ν is, the narrower the δ-slab is, and the larger ν is, the wider the δ-slab is. This is the opposite of the behavior that we witnessed in ν-regression (see Section 56.1).

54.7 Existence of Support Vectors for $(\mathrm{SVM}_{s2'})$

We now consider the issue of the existence of support vectors. We will show that in the "generic case" there is always some blue support vector and some red support vector. The term generic has to do with the choice of ν and will be explained below.

Given any real numbers u, v, x, y, if $\max\{u, v\} < \min\{x, y\}$, then $u < x$ and $v < y$. This is because $u, v \leq \max\{u, v\} < \min\{x, y\} \leq x, y$. Consequently, since by Proposition 54.1, $\max\{2p_f/(p+q), 2q_f/(p+q)\} \leq \nu$, if $\nu < \min\{2p/(p+q), 2q/(p+q)\}$, then $p_f < p$ and $q_f < q$, and since $p_{sf} \leq p_f$ and $q_{sf} \leq q_f$, we also have $p_{sf} < p$ and $q_{sf} < q$. This implies that there are constraints corresponding to some $i \notin E_\lambda$ (in which case $\epsilon_i = 0$) and to some $j \notin E_\mu$ (in which case $\xi_j = 0$), of the form

$$\begin{aligned} w^\top u_i - b &\geq \eta & \qquad & i \notin E_\lambda \\ -w^\top v_j + b &\geq \eta & & j \notin E_\mu. \end{aligned}$$

If $w^\top u_i - b = \eta$ for some $i \notin E_\lambda$ and $-w^\top v_j + b = \eta$ for some $j \notin E_\mu$, then we have a blue support vector and a red support vector. Otherwise, we show how to modify b and η to obtain an optimal solution with a blue support vector and a red support vector.

Proposition 54.2. *For every optimal solution $(w, b, \eta, \epsilon, \xi)$ of Problem $(\mathrm{SVM}_{s2'})$ with $w \neq 0$ and $\eta > 0$, if*

$$\nu < \min\{2p/(p+q), 2q/(p+q)\}$$

and if either no u_i is a support vector or no v_j is a support vector, then there is another optimal solution (for the same w) with some i_0 such that $\epsilon_{i_0} = 0$ and $w^\top u_{i_0} - b = \eta$, and there is some j_0 such that $\xi_{j_0} = 0$ and $-w^\top v_{j_0} + b = \eta$; in other words, some u_{i_0} and some v_{j_0} is a support vector; in particular, $p_{sf} < p$ and $q_{sf} < q$.

Proof. We just explained that $p_{sf} < p$ and $q_{sf} < q$, so the following constraints hold:

$$\begin{aligned} w^\top u_i - b &= \eta - \epsilon_i & \qquad \epsilon_i &> 0 & \qquad & i \in E_\lambda \\ -w^\top v_j + b &= \eta - \xi_j & \xi_j &> 0 & & j \in E_\mu \\ w^\top u_i - b &\geq \eta & & & & i \notin E_\lambda \\ -w^\top v_j + b &\geq \eta & & & & j \notin E_\mu, \end{aligned}$$

where there is some $i \notin E_\lambda$ and some $j \notin E_\mu$.

If our optimal solution does not have a blue support vector and a red support vector, then either $w^\top u_i - b > \eta$ for all $i \notin E_\lambda$ or $-w^\top v_j + b > \eta$ for all $j \notin E_\mu$.

Case 1. We have

$$\begin{aligned} w^\top u_i - b &> \eta & \quad & i \notin E_\lambda \\ -w^\top v_j + b &\geq \eta & & j \notin E_\mu. \end{aligned}$$

There are two subcases.

Case 1a. Assume that there is some $j \notin E_\mu$ such that $-w^\top v_j + b = \eta$. Our strategy is to increase η and b by a small amount θ in such a way that some inequality becomes an equation for some $i \notin E_\lambda$. Geometrically, this amounts to raising the separating hyperplane $H_{w,b}$ and increasing the width of the slab, keeping the red margin hyperplane unchanged. See Figure 54.7.

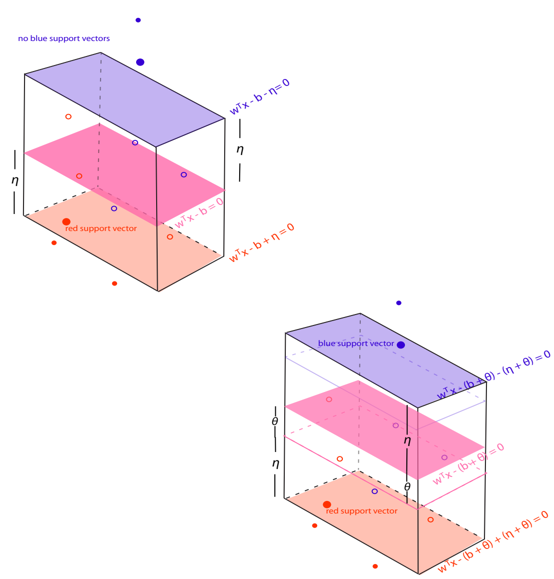


Figure 54.7: In this illustration points with errors are denoted by open circles. In the original, upper left configuration, there is no blue support vector. By raising the pink separating hyperplane and increasing the margin, we end up with a blue support vector.

Let us pick θ such that

$$\theta = (1/2) \min\{w^\top u_i - b - \eta \mid i \notin E_\lambda\}.$$

Our hypotheses imply that $\theta > 0$. We can write

$$\begin{aligned}
w^\top u_i - (b+\theta) &= \eta + \theta - (\epsilon_i + 2\theta) & \epsilon_i > 0 & \quad i \in E_\lambda \\
-w^\top v_j + b + \theta &= \eta + \theta - \xi_j & \xi_j > 0 & \quad j \in E_\mu \\
w^\top u_i - (b+\theta) &\geq \eta + \theta & & \quad i \notin E_\lambda \\
-w^\top v_j + b + \theta &\geq \eta + \theta & & \quad j \notin E_\mu.
\end{aligned}$$

By hypothesis

$$-w^\top v_j + b + \theta = \eta + \theta \quad \text{for some } j \notin E_\mu,$$

and by the choice of θ,

$$w^\top u_i - (b+\theta) = \eta + \theta \quad \text{for some } i \notin E_\lambda.$$

The new value of the objective function is

$$\begin{aligned}
\omega(\theta) &= \frac{1}{2} w^\top w - \nu(\eta + \theta) + \frac{1}{p+q}\left(\sum_{i \in E_\lambda} (\epsilon_i + 2\theta) + \sum_{j \in E_\mu} \xi_j\right) \\
&= \frac{1}{2} w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i \in E_\lambda} \epsilon_i + \sum_{j \in E_\mu} \xi_j\right) - \left(\nu - \frac{2p_{sf}}{p+q}\right)\theta.
\end{aligned}$$

By Proposition 54.1 we have

$$\max\left\{\frac{2p_f}{p+q}, \frac{2q_f}{p+q}\right\} \leq \nu$$

and $p_{sf} \leq p_f$ and $q_{sf} \leq q_f$, which implies that

$$\nu - \frac{2p_{sf}}{p+q} \geq 0, \tag{$*_1$}$$

and so $\omega(\theta) \leq \omega(0)$. If inequality $(*_1)$ is strict, then this contradicts the optimality of the original solution. Therefore, $\nu = 2p_{sf}/(p+q)$, $\omega(\theta) = \omega(0)$, and $(w, b+\theta, \eta+\theta, \epsilon+2\theta, \xi)$ is an optimal solution such that

$$\begin{aligned}
w^\top u_i - (b+\theta) &= \eta + \theta \\
-w^\top v_j + b + \theta &= \eta + \theta
\end{aligned}$$

for some $i \notin E_\lambda$ and some $j \notin E_\mu$.

Case 1b. We have $-w^\top v_j + b > \eta$ for all $j \notin E_\mu$. Our strategy is to increase η and the errors by a small θ in such a way that some inequality becomes an equation for some $i \notin E_\lambda$ or for some $j \notin E_\mu$. Geometrically, this corresponds to increasing the width of the slab, keeping the separating hyperplane unchanged. See Figures 54.8 and 54.9. Then we are reduced to *Case 1a* or *Case 2a*.

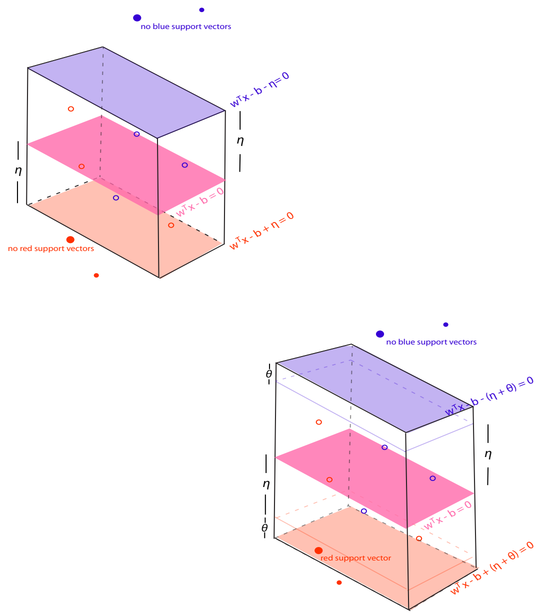


Figure 54.8: In this illustration points with errors are denoted by open circles. In the original, upper left configuration, there is no blue support vector and no red support vector. By increasing the margin, we end up with a red support vector and reduce to Case 1a.

We have

$$\begin{aligned}
w^\top u_i - b &= \eta - \epsilon_i & \epsilon_i &> 0 & i &\in E_\lambda\\
-w^\top v_j + b &= \eta - \xi_j & \xi_j &> 0 & j &\in E_\mu\\
w^\top u_i - b &> \eta & & & i &\notin E_\lambda\\
-w^\top v_j + b &> \eta & & & j &\notin E_\mu.
\end{aligned}$$

Let us pick θ such that

$$\theta = \min\{w^\top u_i - b - \eta,\ -w^\top v_j + b - \eta \mid i \notin E_\lambda, j \notin E_\mu\}.$$

Our hypotheses imply that $\theta > 0$. We can write

$$\begin{aligned}
w^\top u_i - b &= \eta + \theta - (\epsilon_i + \theta) & \epsilon_i &> 0 & i &\in E_\lambda\\
-w^\top v_j + b &= \eta + \theta - (\xi_j + \theta) & \xi_j &> 0 & j &\in E_\mu\\
w^\top u_i - b &\geq \eta + \theta & & & i &\notin E_\lambda\\
-w^\top v_j + b &\geq \eta + \theta & & & j &\notin E_\mu,
\end{aligned}$$

and by the choice of θ, either

$$w^\top u_i - b = \eta + \theta \quad \text{for some } i \notin E_\lambda$$

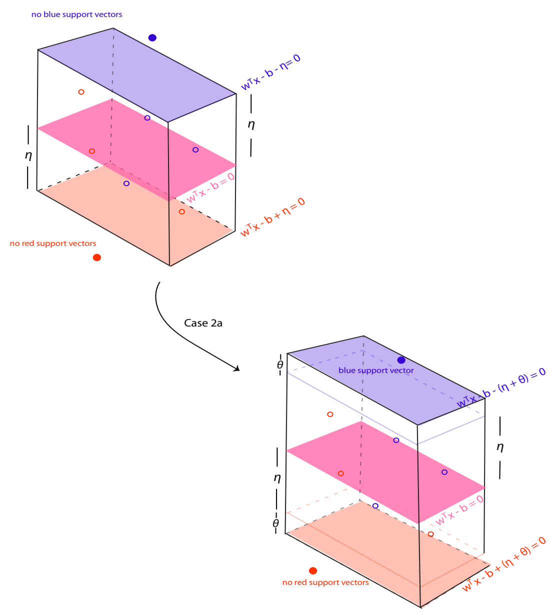


Figure 54.9: In this illustration points with errors are denoted by open circles. In the original, upper left configuration, there is no blue support vector and no red support vector. By increasing the margin, we end up with a blue support vector and reduce to Case 2a.

or

$$-w^\top v_j + b = \eta + \theta \quad \text{for some } j \notin E_\mu.$$

The new value of the objective function is

$$\begin{aligned}\omega(\theta) &= \frac{1}{2}w^\top w - \nu(\eta+\theta) + \frac{1}{p+q}\left(\sum_{i\in E_\lambda}(\epsilon_i+\theta) + \sum_{j\in E_\mu}(\xi_j+\theta)\right)\\ &= \frac{1}{2}w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i\in E_\lambda}\epsilon_i + \sum_{j\in E_\mu}\xi_j\right) - \left(\nu - \frac{p_{sf}+q_{sf}}{p+q}\right)\theta.\end{aligned}$$

Since $\max\{2p_f/(p+q), 2q_f/(p+q)\} \leq \nu$ implies that $(p_f+q_f)/(p+q) \leq \nu$ and $p_{sf} \leq p_f$, $q_{sf} \leq q_f$, we have

$$\nu - \frac{p_{sf}+q_{sf}}{p+q} \geq 0, \tag{$*_2$}$$

and so $\omega(\theta) \leq \omega(0)$. If inequality $(*_2)$ is strict, then this contradicts the optimality of the original solution. Therefore, $\nu = (p_{sf}+q_{sf})/(p+q)$, $\omega(\theta) = \omega(0)$ and $(w, b, \eta+\theta, \epsilon+\theta, \xi+\theta)$ is an optimal solution such that either

$$w^\top u_i - b = \eta + \theta \quad \text{for some } i \notin E_\lambda$$

or

$$-w^\top v_j + b = \eta + \theta \quad \text{for some } j \notin E_\mu.$$

We are now reduced to *Case 1a* or *Case 2a*.

Case 2. We have

$$\begin{aligned} w^\top u_i - b &\geq \eta & & i \notin E_\lambda \\ -w^\top v_j + b &> \eta & & j \notin E_\mu. \end{aligned}$$

There are two subcases.

Case 2a. Assume that there is some $i \notin E_\lambda$ such that $w^\top u_i - b = \eta$. Our strategy is to increase η and decrease b by a small amount θ in such a way that some inequality becomes an equation for some $j \notin E_\mu$. Geometrically, this amounts to lowering the separating hyperplane $H_{w,b}$ and increasing the width of the slab, keeping the blue margin hyperplane unchanged. See Figure 54.10.

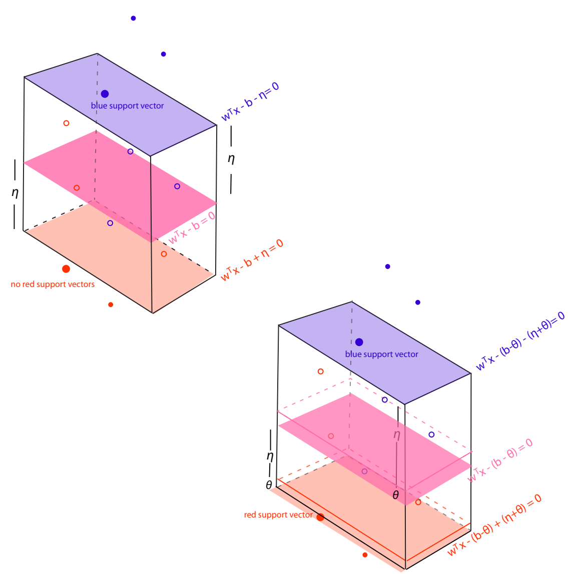


Figure 54.10: In this illustration points with errors are denoted by open circles. In the original, upper left configuration, there is no red support vector. By lowering the pink separating hyperplane and increasing the margin, we end up with a red support vector.

Let us pick θ such that

$$\theta = (1/2) \min\{-w^\top v_j + b - \eta \mid j \notin E_\mu\}.$$

Our hypotheses imply that $\theta > 0$. We can write

$$\begin{aligned} w^\top u_i - (b - \theta) &= \eta + \theta - \epsilon_i & \epsilon_i &> 0 & & i \in E_\lambda \\ -w^\top v_j + b - \theta &= \eta + \theta - (\xi_j + 2\theta) & \xi_j &> 0 & & j \in E_\mu \\ w^\top u_i - (b - \theta) &\geq \eta + \theta & & & & i \notin E_\lambda \\ -w^\top v_j + b - \theta &\geq \eta + \theta & & & & j \notin E_\mu. \end{aligned}$$

By hypothesis

$$w^\top u_i - (b - \theta) = \eta + \theta \quad \text{for some } i \notin E_\lambda,$$

and by the choice of θ,

$$-w^\top v_j + b - \theta = \eta + \theta \quad \text{for some } j \notin E_\mu.$$

The new value of the objective function is

$$\begin{aligned}\omega(\theta) &= \frac{1}{2}w^\top w - \nu(\eta + \theta) + \frac{1}{p+q}\left(\sum_{i \in E_\lambda} \epsilon_i + \sum_{j \in E_\mu}(\xi_j + 2\theta)\right) \\ &= \frac{1}{2}w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i \in E_\lambda} \epsilon_i + \sum_{j \in E_\mu} \xi_j\right) - \left(\nu - \frac{2q_{sf}}{p+q}\right)\theta.\end{aligned}$$

The rest of the proof is similar to Case 1a with p_{sf} replaced by q_{sf}.

Case 2b. We have $w^\top u_i - b > \eta$ for all $i \notin E_\lambda$. Since by hypothesis $-w^\top v_j + b > \eta$ for all $j \notin E_\mu$, Case 2b is identical to Case 1b, and we are done. ☐

A subtle point here is that Proposition 54.2 shows that if there is an optimal solution, then there is one with a blue and a red support vector, but it does not guarantee that these are support vectors of type 1. Since the dual program does not determine b and η unless these support vectors are of type 1, from a practical point of view this proposition is not helpful.

The proof of Proposition 54.2 reveals that there are three critical values for ν:

$$\frac{2p_{sf}}{p+q}, \quad \frac{2q_{sf}}{p+q}, \quad \frac{p_{sf} + q_{sf}}{p+q}.$$

These values can be avoided by requiring the strict inequality

$$\max\left\{\frac{2p_{sf}}{p+q}, \frac{2q_{sf}}{p+q}\right\} < \nu.$$

Then the following corollary holds.

Theorem 54.3. *For every optimal solution $(w, b, \eta, \epsilon, \xi)$ of Problem* (SVM$_{s2'}$) *with $w \neq 0$ and $\eta > 0$, if*

$$\max\{2p_f/(p+q), 2q_f/(p+q)\} < \nu < \min\{2p/(p+q), 2q/(p+q)\},$$

then some u_{i_0} and some v_{j_0} is a support vector.

Proof. We proceed by contradiction. Suppose that for every optimal solution with $w \neq 0$ and $\eta > 0$ no u_i is a blue support vector or no v_j is a red support vector. Since $\nu < \min\{2p/(p+q), 2q/(p+q)\}$, Proposition 54.2 holds, so there is another optimal solution. But since the critical values of ν are avoided, the proof of Proposition 54.2 shows that the value of the objective function for this new optimal solution is strictly smaller than the original optimal value, a contradiction. ☐

We also have the following proposition that gives a sufficient condition implying that η and b can be found in terms of an optimal solution (λ, μ) of the dual.

Proposition 54.4. *If $(w, b, \eta, \epsilon, \xi)$ is an optimal solution of Problem $(\mathrm{SVM}_{s2'})$ with $w \neq 0$ and $\eta > 0$, if*

$$\max\{2p_f/(p+q), 2q_f/(p+q)\} < \nu < \min\{2p/(p+q), 2q/(p+q)\},$$

then η and b can always be determined from an optimal solution (λ, μ) of the dual in terms of a single support vector.

Proof. By Theorem 54.3 some u_{i_0} and some v_{j_0} is a support vector. As we already explained, Problem $(\mathrm{SVM}_{s2'})$ satisfies the conditions for having a zero duality gap. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) = G(\lambda, \mu, \alpha, \beta),$$

which means that

$$\frac{1}{2} w^\top w - \nu\eta + \frac{1}{p+q}\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

and since

$$w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

we get

$$\frac{1}{p+q}\left(\sum_{i=1}^{p} \epsilon_i + \sum_{j=1}^{q} \xi_j\right) = \nu\eta - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}. \tag{$*$}$$

Let $K_\lambda = \{i \in \{1, \ldots, p\} \mid \lambda_i = K_s\}$ and $K_\mu = \{j \in \{1, \ldots, q\} \mid \mu_j = K_s\}$. By definition, $p_f = |K_\lambda|$ and $q_f = |K_\mu|$ (here we assuming that $K_s = 1/(p+q)$). By complementary slackness the following equations are active:

$$\begin{aligned} w^\top u_i - b &= \eta - \epsilon_i & & i \in K_\lambda \\ -w^\top v_j + b &= \eta - \xi_j & & j \in K_\mu. \end{aligned}$$

But $(*)$ can be written as

$$\frac{1}{p+q}\left(\sum_{i \in K_\lambda} \epsilon_i + \sum_{j \in K_\mu} \xi_j\right) = \nu\eta - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}, \tag{$**$}$$

and since

$$\begin{aligned} \epsilon_i &= \eta - w^\top u_i + b & & i \in K_\lambda \\ \xi_j &= \eta + w^\top v_j - b & & j \in K_\mu, \end{aligned}$$

by substituting in the Equation $(**)$ we get

$$\left(\nu - \frac{p_f + q_f}{p+q}\right)\eta = \frac{p_f - q_f}{p+q} b + \frac{1}{p+q} w^\top \left(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i\right) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}.$$

We also know that $w^\top u_{i_0} - b = \eta$ and $-w^\top v_{j_0} + b = \eta$ for some i_0 and some j_0. In the first case $b = -\eta + w^\top u_{i_0}$, and by substituting b in the above equation we get the equation

$$\begin{aligned}\left(\nu - \frac{p_f + q_f}{p+q}\right)\eta &= -\frac{p_f - q_f}{p+q}\eta + \frac{p_f - q_f}{p+q} w^\top u_{i_0} + \frac{1}{p+q} w^\top \left(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i\right) \\ &\quad + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix},\end{aligned}$$

that is,

$$\begin{aligned}\left(\nu - \frac{2q_f}{p+q}\right)\eta &= \frac{p_f - q_f}{p+q} w^\top u_{i_0} + \frac{1}{p+q} w^\top \left(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i\right) \\ &\quad + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}.\end{aligned}$$

In the second case $b = \eta + w^\top v_{j_0}$, and we get the equation

$$\begin{aligned}\left(\nu - \frac{p_f + q_f}{p+q}\right)\eta &= \frac{p_f - q_f}{p+q}\eta + \frac{p_f - q_f}{p+q} w^\top v_{j_0} + \frac{1}{p+q} w^\top \left(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i\right) \\ &\quad + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix},\end{aligned}$$

that is,

$$\begin{aligned}\left(\nu - \frac{2p_f}{p+q}\right)\eta &= \frac{p_f - q_f}{p+q} w^\top v_{j_0} + \frac{1}{p+q} w^\top \left(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i\right) \\ &\quad + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}.\end{aligned}$$

We need to choose ν such that $2p_f/(p+q) - \nu \neq 0$ and $2q_f/(p+q) - \nu \neq 0$. Since by Proposition 54.1, we have $\max\{2p_f/(p+q), 2q_f/(p+q)\} \leq \nu$, it suffices to pick ν such that $\max\{2p_f/(p+q), 2q_f/(p+q)\} < \nu$. If this condition is satisfied we can solve for η, and then we find b from either $b = -\eta + w^\top u_{i_0}$ or $b = \eta + w^\top v_{j_0}$. $\square$

Remark: Of course the hypotheses of the proposition imply that $w^\top u_{i_0} - b = \eta$ and $-w^\top v_{j_0} + b = \eta$ for some i_0 and some j_0. Thus we can also compute b and η using the formulae

$$\begin{aligned}b &= \frac{w^\top(u_{i_0} + v_{j_0})}{2} \\ \eta &= \frac{w^\top(u_{i_0} - v_{j_0})}{2}.\end{aligned}$$

The interest of Proposition 54.4 lies in the fact that it allows us to compute b and η knowing only a *single* support vector.

In practice we can only find support vectors of type 1 so Proposition 54.4 is useful if we can only find some blue support vector of type 1 *or* some red support vector of type 1.

As earlier, if we define I_λ and I_μ as

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}, \end{aligned}$$

then we have the following cases to compute η and b.

(1) If $I_\lambda \neq \emptyset$ and $I_\mu \neq \emptyset$, then

$$\begin{aligned} b &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2 \\ \eta &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| - \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2. \end{aligned}$$

(2) If $I_\lambda \neq \emptyset$ and $I_\mu = \emptyset$, then

$$\begin{aligned} b &= -\eta + w^\top \Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| \\ ((p+q)\nu - 2q_f)\eta &= (p_f - q_f) w^\top \Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| + w^\top \Big(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \Big) \\ &\quad + (p+q) \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

(3) If $I_\lambda = \emptyset$ and $I_\mu \neq \emptyset$, then

$$\begin{aligned} b &= \eta + w^\top \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \\ ((p+q)\nu - 2p_f)\eta &= (p_f - q_f) w^\top \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| + w^\top \Big(\sum_{i \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \Big) \\ &\quad + (p+q) \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

The above formulae correspond to $K_s = 1/(p+q)$. In general we have to replace the rightmost $(p+q)$ by $1/K_s$.

We have examples where there is a single support vector of type 1 and $\nu = 2q_f/(p+q)$, so the above method fails. Curiously, perturbing ν slightly yields a solution with some blue support vector of type 1 and some red support vector of type 1, and so we have not yet found an example where the above method succeeds with a single support vector of type 1. This suggests to conduct some perturbation analysis but it appears to be nontrivial.

Among its advantages, the support vector machinery is conducive to finding interesting statistical bounds in terms of the *VC dimension*, a notion invented by Vapnik and Chernovenkis. We will not go into this here and instead refer the reader to Vapnik [182] (especially, Chapter 4 and Chapters 9-13).

54.8 Solving SVM $(\mathrm{SVM}_{s2'})$ Using ADMM

In order to solve $(\mathrm{SVM}_{s2'})$ using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$\begin{aligned}
\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0\\
\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j - \gamma &= K_m\\
\lambda_i + \alpha_i &= K_s, \quad i = 1, \ldots, p\\
\mu_j + \beta_j &= K_s, \quad j = 1, \ldots, q,
\end{aligned}$$

with $K_m = (p+q)K_s\nu$. This is the $(p+q+2)\times(2(p+q)+1)$ matrix A given by

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top & 0\\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top & -1\\ I_p & 0_{p,q} & I_p & 0_{p,q} & 0_p\\ 0_{q,p} & I_q & 0_{q,p} & I_q & 0_q \end{pmatrix}.$$

Observe the remarkable analogy with the matrix arising in ν-regression in Section 56.3, except that $p = q = m$ and that -1 is replaced by $+1$. We leave it as an exercise to prove that A has rank $p+q+2$. The right-hand side is

$$c = \begin{pmatrix} 0\\ K_m\\ K_s\mathbf{1}_{p+q}\end{pmatrix}.$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q\end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $2(p+q)+1$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta, \gamma)$, the $(p+q) \times (p+q)$ matrix $X^\top X$ must be augmented with zero's to make it a $(2(p+q)+1) \times (2(p+q)+1)$ matrix P_a given by

$$P_a = \begin{pmatrix} X^\top X & 0_{p+q,p+q+1} \\ 0_{p+q+1,p+q} & 0_{p+q+1,p+q+1} \end{pmatrix},$$

and similarly q is augmented with zeros as the vector $q_a = 0_{2(p+q)+1}$.

As we mentioned in Section 54.5, since $\eta \geq 0$ for an optimal solution, we can drop the constraint $\eta \geq 0$ from the primal problem. In this case there are $2(p+q)$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$. It is easy to see that the objective function of the dual is unchanged and the set of constraints is

$$\begin{aligned} \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0 \\ \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j &= K_m \\ \lambda_i + \alpha_i &= K_s, \quad i = 1, \ldots, p \\ \mu_j + \beta_j &= K_s, \quad j = 1, \ldots, q, \end{aligned}$$

with $K_m = (p+q)K_s\nu$. The constraint matrix corresponding to this system of equations is the $(p+q+2) \times 2(p+q)$ matrix A_2 given by

$$A_2 = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}.$$

We leave it as an exercise to prove that A_2 has rank $p+q+2$. The right-hand side is

$$c_2 = \begin{pmatrix} 0 \\ K_m \\ K_s \mathbf{1}_{p+q} \end{pmatrix}.$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $2(p+q)$ Lagrange multipliers the $(p+q) \times (p+q)$ matrix $X^\top X$ must be augmented with zero's to make it a $2(p+q) \times 2(p+q)$ matrix P_{2a} given by

$$P_{2a} = \begin{pmatrix} X^\top X & 0_{p+q,p+q} \\ 0_{p+q,p+q} & 0_{p+q,p+q} \end{pmatrix},$$

and similarly q is augmented with zeros as the vector $q_{2a} = 0_{2(p+q)}$.

The `Matlab` programs implementing the above method are given in Appendix B, Section B.2. We ran our program on two sets of 30 points each generated at random using the following code which calls the function `runSVMs2pbv3`:

```
rho = 10;
u16 = 10.1*randn(2,30)+7 ;
v16 = -10.1*randn(2,30)-7;
[~,~,~,~,~,~,w3] = runSVMs2pbv3(0.37,rho,u16,v16,1/60)
```

We picked $K = 1/60$ and various values of ν starting with $\nu = 0.37$, which appears to be the smallest value for which the method converges; see Figure 54.11.

In this example, $p_f = 10, q_f = 11, p_m = 12, q_m = 12$. The quadratic solver converged after 8121 steps to reach primal and dual residuals smaller than 10^{-10}.

Reducing ν below $\nu = 0.37$ has the effect that p_f, q_f, p_m, q_m decrease but the following situation arises. Shrinking η a little bit has the effect that $p_f = 9, q_f = 10, p_m = 10, q_m = 11$. Then $\max\{p_f, q_f\} = \min\{p_m, q_m\} = 10$, so the only possible value for ν is $\nu = 20/60 = 1/3 = 0.3333333\cdots$. When we run our program with $\nu = 1/3$, it returns a value of η less than 10^{-13} and a value of w whose components are also less than 10^{-13}. This is probably due to numerical precision. Values of ν less than $1/3$ cause the same problem. It appears that the geometry of the problem constrains the values of p_f, q_f, p_m, q_m in such a way that it has no solution other than $w = 0$ and $\eta = 0$.

Figure 54.12 shows the result of running the program with $\nu = 0.51$. We have $p_f = 15, q_f = 16, p_m = 16, q_m = 16$. Interestingly, for $\nu = 0.5$, we run into the singular situation where there is only one support vector and $\nu = 2p_f/(p+q)$.

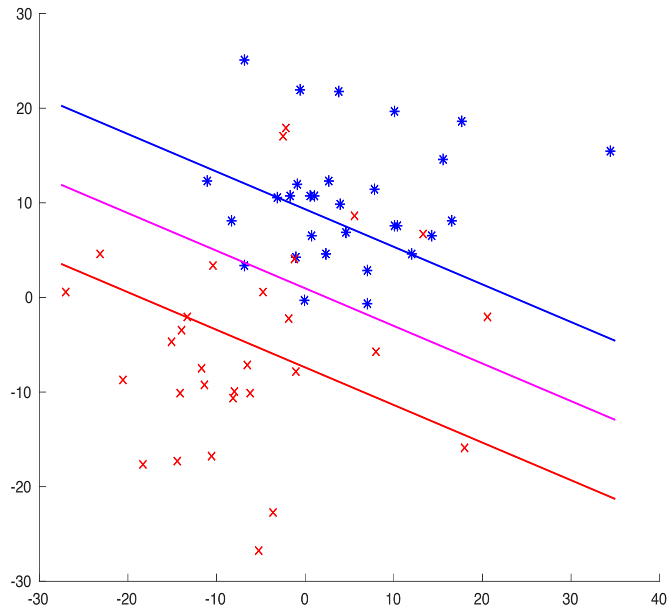


Figure 54.11: Running (SVM$_{s2'}$) on two sets of 30 points; $\nu = 0.37$.

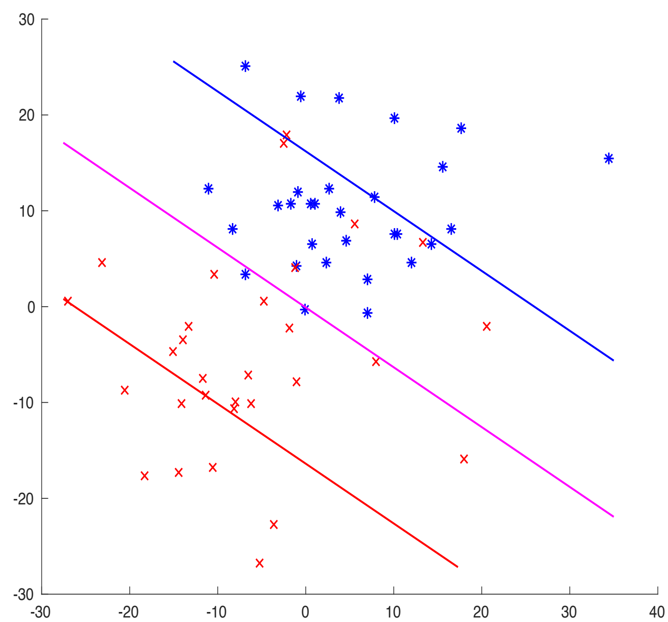


Figure 54.13: Running (SVM$_{s2'}$) on two sets of 30 points; $\nu = 0.71$.

Next Figure 54.13 shows the result of running the program with $\nu = 0.71$. We have

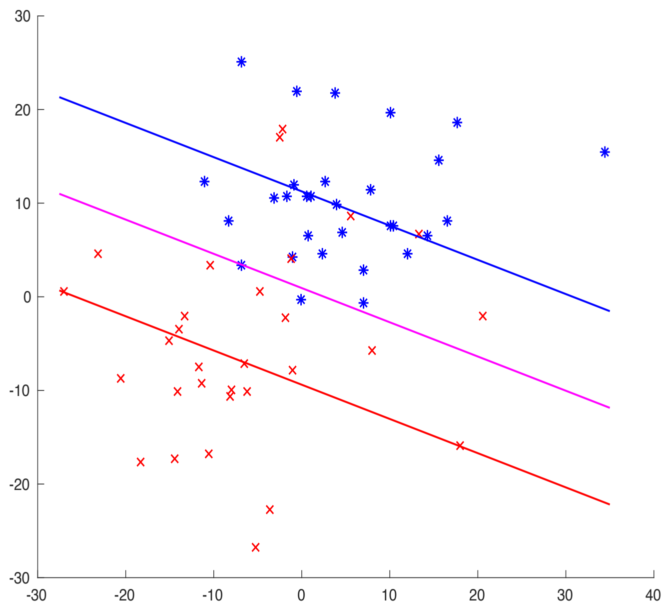


Figure 54.12: Running ($\text{SVM}_{s2'}$) on two sets of 30 points; $\nu = 0.51$.

$p_f = 21, q_f = 21, p_m = 22, q_m = 23$. Interestingly, for $\nu = 0.7$, we run into the singular situation where there are no support vectors.

For our next to the last run, Figure 54.14 shows the result of running the program with $\nu = 0.95$. We have $p_f = 28, q_f = 28, p_m = 29, q_m = 29$.

Figure 54.15 shows the result of running the program with $\nu = 0.97$. We have $p_f = 29, q_f = 29, p_m = 30, q_m = 30$, which shows that the largest margin has been achieved. However, after 80000 iterations the dual residual is less than 10^{-12} but the primal residual is approximately 10^{-4} (our tolerance for convergence is 10^{-10}, which is quite high). Nevertheless the result is visually very good.

54.9 Soft Margin Support Vector Machines; (SVM_{s3})

In this section we consider a variation of Problem ($\text{SVM}_{s2'}$) by adding the term $(1/2)b^2$ to the objective function. The result is that in minimizing the Lagrangian to find the dual function G, not just w but also b is determined and η is determined under a mild condition on ν. We also suppress the constraint $\eta \geq 0$ which turns out to be redundant.

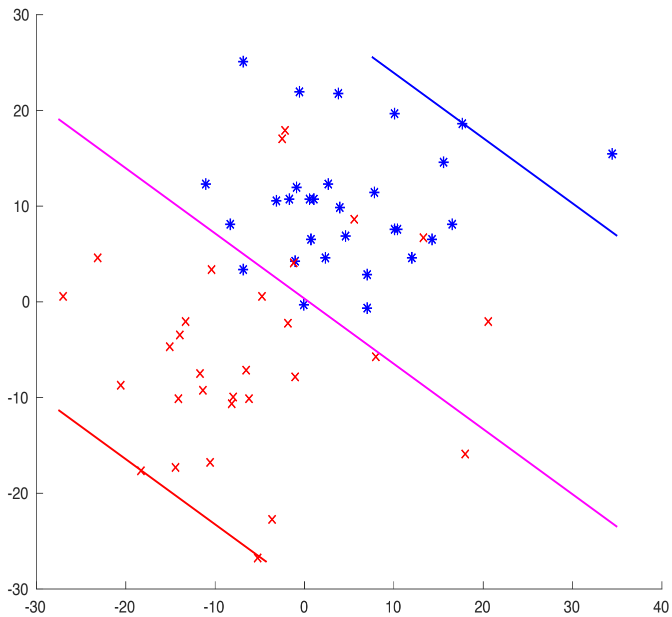


Figure 54.14: Running (SVM$_{s2'}$) on two sets of 30 points; $\nu = 0.95$.

Soft margin SVM (SVM$_{s3}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + (p+q)K_s\left(-\nu\eta + \frac{1}{p+q}\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\right)\\
&\text{subject to}\\
&\qquad\quad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad - w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

To simplify the presentation we assume that $K_s = 1/(p+q)$. When writing a computer program it is more convenient to assume that K_s is arbitrary. In this case, ν needs to be replaced by $(p+q)K_s\nu$ in all the formulae.

The Lagrangian $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta)$ with $\lambda, \alpha \in \mathbb{R}^p_+$, $\mu, \beta \in \mathbb{R}^q_+$ is given by

$$\begin{aligned}
L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) &= \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \frac{b^2}{2} - \nu\eta + K_s(\epsilon^\top \mathbf{1}_p + \xi^\top \mathbf{1}_q) - \epsilon^\top(\lambda + \alpha)\\
&\quad - \xi^\top(\mu + \beta) + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu)\\
&= \frac{1}{2}w^\top w + w^\top X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + \frac{b^2}{2} + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu)\\
&\quad + \epsilon^\top(K_s\mathbf{1}_p - (\lambda + \alpha)) + \xi^\top(K_s\mathbf{1}_q - (\mu + \beta)).
\end{aligned}$$

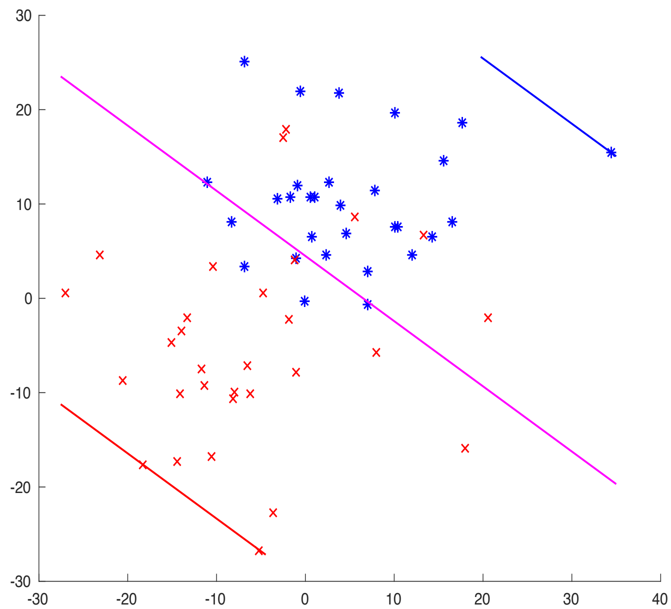


Figure 54.15: Running ($\mathrm{SVM}_{s2'}$) on two sets of 30 points; $\nu = 0.97$.

To find the dual function $G(\lambda, \mu, \alpha, \beta)$, we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute its gradient with respect to $w, \epsilon, \xi, b, \eta$, and we get

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + w \\ K_s \mathbf{1}_p - (\lambda + \alpha) \\ K_s \mathbf{1}_q - (\mu + \beta) \\ b + \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \tag{$*_w$}$$

$$\begin{aligned} \lambda + \alpha &= K_s \mathbf{1}_p \\ \mu + \beta &= K_s \mathbf{1}_q \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= \nu, \end{aligned}$$

and

$$b = -(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \tag{$*_b$}$$

The second and third equations are equivalent to the box constraints

$$0 \leq \lambda_i, \mu_j \leq K_s, \quad i = 1, \ldots, p, \; j = 1, \ldots, q.$$

Since we assumed that the primal problem has an optimal solution with $w \neq 0$, we have

$$X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \neq 0.$$

Plugging back w from $(*_w)$ and b from $(*_b)$ into the Lagrangian, we get

$$\begin{aligned}
G(\lambda, \mu, \alpha, \beta) &= \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \frac{1}{2} b^2 - b^2 \\
&= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \frac{1}{2} b^2 \\
&= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix},
\end{aligned}$$

so the dual function is independent of α, β and is given by

$$G(\lambda, \mu) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

The dual program is given by

$$\begin{aligned}
&\text{maximize} \quad -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

Finally, the dual program is equivalent to the following minimization program:

Dual of the Soft margin SVM (SVM$_{s3}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

The classification of the points u_i and v_j in terms of the values of λ and μ and Definition 54.2 and Definition 54.3 are unchanged.

It is shown in Section 54.12 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ. Once a solution for λ and μ is obtained, we have

$$\begin{aligned} w &= -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j \\ b &= -(\mathbf{1}_p^\top\lambda - \mathbf{1}_q^\top\mu) = -\sum_{i=1}^{p}\lambda_i + \sum_{j=1}^{q}\mu_j. \end{aligned}$$

We can compute η using duality. As we said earlier, the hypotheses of Theorem 50.17(2) hold, so *if* the primal problem (SVM_{s3}) has an optimal solution with $w \neq 0$, *then* the dual problem has a solution too, and the duality gap is zero. Therefore, for optimal solutions we have

$$L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta) = G(\lambda, \mu, \alpha, \beta),$$

which means that

$$\begin{aligned} &\frac{1}{2}w^\top w + \frac{b^2}{2} - (p+q)K_s\nu\eta + K_s\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right) \\ &\qquad = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \end{aligned}$$

We can use the above equation to determine η.

Since

$$\frac{1}{2}w^\top w + \frac{b^2}{2} = \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix},$$

we get

$$(p+q)K_s\nu\eta = K_s\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}. \quad (*)$$

Since

$$X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}$$

is positive semidefinite, we confirm that $\eta \geq 0$.

Since nonzero ϵ_i and ξ_j may only occur for vectors u_i and v_j that fail the margin, namely $\lambda_i = K_s$, $\mu_j = K_s$, the corresponding constraints are active and we can solve for ϵ_i and ξ_j in terms of b and η. Let K_λ and K_μ be the sets of indices corresponding to points failing the margin,

$$K_\lambda = \{i \in \{1, \dots, p\} \mid \lambda_i = K_s\}$$
$$K_\mu = \{j \in \{1, \dots, q\} \mid \mu_j = K_s\}.$$

By definition $p_f = |K_\lambda|, q_f = |K_\mu|$. Then for every $i \in K_\lambda$ we have

$$\epsilon_i = \eta + b - w^\top u_i$$

and for every $j \in K_\mu$ we have

$$\xi_j = \eta - b + w^\top v_j.$$

Using the above formulae we obtain

$$\begin{aligned}
\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j &= \sum_{i \in K_\lambda} \epsilon_i + \sum_{j \in K_\mu} \xi_j \\
&= \sum_{i \in K_\lambda} (\eta + b - w^\top u_i) + \sum_{j \in K_\mu} (\eta - b + w^\top v_j) \\
&= (p_f + q_f)\eta + (p_f - q_f)b + w^\top \left(\sum_{j \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \right)
\end{aligned}$$

Substituting this expression in $(*)$ we obtain

$$\begin{aligned}
(p+q)K_s\nu\eta &= K_s \left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j \right) + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&= K_s \left((p_f + q_f)\eta + (p_f - q_f)b + w^\top \left(\sum_{j \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \right) \right) \\
&\quad + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix},
\end{aligned}$$

which yields

$$\begin{aligned}
((p+q)\nu - p_f - q_f)\eta &= (p_f - q_f)b + w^\top \left(\sum_{j \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \right) \\
&\quad + \frac{1}{K_s} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.
\end{aligned}$$

We show in Proposition 54.5 that $p_f + q_f \leq (p+q)\nu$, so if $\nu > (p_f + q_f)/(p+q)$, we can solve for η in terms of b, w, and λ, μ. But b and w are expressed in terms of λ, μ as

$$\begin{aligned}
w &= -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
b &= -\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = -\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu
\end{aligned}$$

so η is also expressed in terms of λ, μ.

The condition $\nu > (p_f + q_f)/(p+q)$ cannot be satisfied if $p_f + q_f = p + q$, but in this case all points fail the margin, which indicates that δ is too big, so we reduce ν and try again.

Remark: The equation

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \nu$$

implies that either there is some i_0 such that $\lambda_{i_0} > 0$ *or* there is some j_0 such that $\mu_{j_0} > 0$, which implies that $p_m + q_m \geq 1$.

Another way to compute η is to assume the Standard Margin Hypothesis for (SVM_{s3}). Under the **Standard Margin Hypothesis** for (SVM_{s3}), either there is some i_0 such that $0 < \lambda_{i_0} < K_s$ *or* there is some j_0 such that $0 < \mu_{j_0} < K_s$, in other words, there is some support vector of type 1. By the complementary slackness conditions $\epsilon_{i_0} = 0$ or $\xi_{j_0} = 0$, so we have

$$w^\top u_{i_0} - b = \eta, \quad \text{or} \quad - w^\top v_{j_0} + b = \eta,$$

and we can solve for η.

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\} \\ I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}. \end{aligned}$$

Then it is easy to see that we can compute η using the following averaging formulae: If $I_\lambda \neq \emptyset$, then

$$\eta = w^\top \left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| - b,$$

and if $I_\mu \neq \emptyset$, then

$$\eta = b - w^\top \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu|.$$

Theoretically the condition $\nu > (p_f + q_f)/(p+q)$ is less restrictive that the **Standard Margin Hypothesis** but in practice we have never observed an example for which $\nu > (p_f + q_f)/(p+q)$ and yet the **Standard Margin Hypothesis** fails.

The "kernelized" version of Problem (SVM_{s3}) is the following:

Soft margin kernel SVM (SVM_{s3}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + \frac{1}{2}b^2 - \nu\eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\ &\text{subject to} \\ &\qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\ &\qquad - \langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q, \end{aligned}$$

with $K_s = 1/(p+q)$.

Tracing through the derivation of the dual program, we obtain

Dual of the Soft margin kernel SVM (SVM$_{s3}$):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q,
\end{aligned}
$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

We obtain

$$
\begin{aligned}
w &= \sum_{i=1}^p \lambda_i \varphi(u_i) - \sum_{j=1}^q \mu_j \varphi(v_j) \\
b &= -\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j.
\end{aligned}
$$

The classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

is given by

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^p \lambda_i(\kappa(u_i, x) + 1) - \sum_{j=1}^q \mu_j(\kappa(v_j, x) + 1)\right).$$

54.10 Classification of the Data Points in Terms of ν (SVM$_{s3}$)

The equations (†) and the box inequalities

$$0 \leq \lambda_i \leq K_s, \quad 0 \leq \mu_j \leq K_s$$

also imply the following facts (recall that $\delta = \eta / \|w\|$):

Proposition 54.5. *If Problem* (SVM$_{s3}$) *has an optimal solution with $w \neq 0$ and $\eta > 0$ then the following facts hold:*

(1) Let p_f be the number of points u_i such that $\lambda_i = K_s$, and let q_f the number of points v_j such that $\mu_j = K_s$. Then $p_f + q_f \leq (p+q)\nu$.

(2) Let p_m be the number of points u_i such that $\lambda_i > 0$, and let q_m the number of points v_j such that $\mu_j > 0$. Then $p_m + q_m \geq (p+q)\nu$. We have $p_m + q_m \geq 1$.

(3) If $p_f \geq 1$ or $q_f \geq 1$, then $\nu \geq 1/(p+q)$.

Proof. (1) Recall that for an optimal solution with $w \neq 0$ and $\eta > 0$ we have the equation

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \nu.$$

Since there are p_f points u_i such that $\lambda_i = K_s = 1/(p+q)$ and q_f points v_j such that $\mu_j = K_s = 1/(p+q)$, we have

$$\nu = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq \frac{p_f + q_f}{p+q},$$

so

$$p_f + q_f \leq \nu(p+q).$$

(2) If

$$I_{\lambda>0} = \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \quad \text{and} \quad p_m = |I_{\lambda>0}|$$

and

$$I_{\mu>0} = \{j \in \{1, \ldots, q\} \mid \mu_j > 0\} \quad \text{and} \quad q_m = |I_{\mu>0}|,$$

then

$$\nu = \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \sum_{i \in I_{\lambda>0}} \lambda_i + \sum_{j \in I_{\mu>0}} \mu_j,$$

and since $\lambda_i, \mu_j \leq K_s = 1/(p+q)$, we have

$$\nu = \sum_{i \in I_{\lambda>0}} \lambda_i + \sum_{j \in I_{\mu>0}} \mu_j \leq \frac{p_m + q_m}{p+q},$$

which yields

$$p_m + q_m \geq \nu(p+q).$$

We already noted earlier that $p_m + q_m \geq 1$.

(3) This follows immediately from (1). $\square$

Note that if ν is chosen so that $\nu < 1/(p+q)$, then $p_f = q_f = 0$, which means that none of the data points are misclassified; in other words, the u_is and v_js are linearly separable. Thus we see that if the u_is and v_js are not linearly separable we must pick ν such that $1/(p+q) \leq \nu \leq 1$ for the method to succeed. In fact, by Proposition 54.5, we must choose ν so that

$$\frac{p_f + q_f}{p+q} \leq \nu \leq \frac{p_m + q_m}{p+q}.$$

Furthermore, in order to be able to determine b, we must have the strict inequality

$$\frac{p_f + q_f}{p+q} < \nu.$$

54.11 Existence of Support Vectors for (SVM_{s3})

The following proposition is the version of Proposition 54.2 for Problem (SVM_{s3}).

Proposition 54.6. *For every optimal solution $(w, b, \eta, \epsilon, \xi)$ of Problem (SVM_{s3}) with $w \neq 0$ and $\eta > 0$, if $\nu < 1$ and if no u_i is a support vector and no v_j is a support vector, then there is another optimal solution such that some u_{i_0} or some v_{j_0} is a support vector.*

Proof. We may assume that $K_s = 1/(p+q)$ and we proceed by contradiction. Thus we assume that for all $i \in \{1, \ldots, p\}$, if $\epsilon_i = 0$, then the constraint $w^\top u_i - b \geq \eta$ is not active, namely $w^\top u_i - b > \eta$, *and* for all $j \in \{1, \ldots, q\}$, if $\xi_j = 0$, then the constraint $-w^\top v_j + b \geq \eta$ is not active, namely $-w^\top v_j + b > \eta$.

Let $E_\lambda = \{i \in \{1, \ldots, p\} \mid \epsilon_i > 0\}$ and let $E_\mu = \{j \in \{1, \ldots, q\} \mid \xi_j > 0\}$. By definition, $p_{sf} = |E_\lambda|$, $q_{sf} = |E_\mu|$, $p_{sf} \leq p_f$ and $q_{sf} \leq q_f$, so by Proposition 54.1,

$$\frac{p_{sf} + q_{sf}}{p+q} \leq \frac{p_f + q_f}{p+q} \leq \nu.$$

Therefore, if $\nu < 1$, then $p_{sf} + q_{sf} < p + q$, which implies that either there is some $i \notin E_\lambda$ such that $\epsilon_i = 0$ or there is some $j \notin E_\mu$ such that $\xi_j = 0$.

By complementary slackness all the constraints for which $i \in E_\lambda$ and $j \in E_\mu$ are active, so our hypotheses are

$$\begin{array}{lll}
w^\top u_i - b = \eta - \epsilon_i & \epsilon_i > 0 & i \in E_\lambda \\
-w^\top v_j + b = \eta - \xi_j & \xi_j > 0 & j \in E_\mu \\
w^\top u_i - b > \eta & & i \notin E_\lambda \\
-w^\top v_j + b > \eta & & j \notin E_\mu,
\end{array}$$

and either there is some $i \notin E_\lambda$ or there is some $j \notin E_\mu$. Our strategy, as illustrated in Figures 54.8 and 54.9, is to increase the width η of the slab keeping the separating hyperplane unchanged. Let us pick θ such that

$$\theta = \min\{w^\top u_i - b - \eta, \ -w^\top v_j + b - \eta \mid i \notin E_\lambda, j \notin E_\mu\}.$$

Our hypotheses imply that $\theta > 0$. We can write

$$\begin{aligned} w^\top u_i - b &= \eta + \theta - (\epsilon_i + \theta) & \epsilon_i + \theta > 0 \qquad & i \in E_\lambda \\ -w^\top v_j + b &= \eta + \theta - (\xi_j + \theta) & \xi_j + \theta > 0 \qquad & j \in E_\mu \\ w^\top u_i - b &\geq \eta + \theta & & i \notin E_\lambda \\ -w^\top v_j + b &\geq \eta + \theta & & j \notin E_\mu, \end{aligned}$$

and by the choice of θ, either

$$w^\top u_i - b = \eta + \theta \quad \text{for some } i \notin E_\lambda$$

or

$$-w^\top v_j + b = \eta + \theta \quad \text{for some } j \notin E_\mu.$$

The original value of the objective function is

$$\omega(0) = \frac{1}{2} w^\top w + \frac{1}{2} b^2 - \nu\eta + \frac{1}{p+q}\biggl(\sum_{i \in E_\lambda} \epsilon_i + \sum_{j \in E_\mu} \xi_j\biggr),$$

and the new value is

$$\begin{aligned} \omega(\theta) &= \frac{1}{2} w^\top w + \frac{1}{2} b^2 - \nu(\eta + \theta) + \frac{1}{p+q}\biggl(\sum_{i \in E_\lambda} (\epsilon_i + \theta) + \sum_{j \in E_\mu} (\xi_j + \theta)\biggr) \\ &= \frac{1}{2} w^\top w + \frac{1}{2} b^2 - \nu\eta + \frac{1}{p+q}\biggl(\sum_{i \in E_\lambda} \epsilon_i + \sum_{j \in E_\mu} \xi_j\biggr) - \biggl(\nu - \frac{p_{sf} + q_{sf}}{p+q}\biggr)\theta. \end{aligned}$$

By Proposition 54.1,

$$\frac{p_{sf} + q_{sf}}{p+q} \leq \frac{p_f + q_f}{p+q} \leq \nu,$$

so

$$\nu - \frac{p_{sf} + q_{sf}}{p+q} \geq 0,$$

and so $\omega(\theta) \leq \omega(0)$. If the inequality is strict, then this contradicts the optimality of the original solution. Therefore, $\omega(\theta) = \omega(0)$ and $(w, b, \eta + \theta, \epsilon + \theta, \xi + \theta)$ is an optimal solution such that either

$$w^\top u_i - b = \eta + \theta \quad \text{for some } i \notin E_\lambda$$

or

$$-w^\top v_j + b = \eta + \theta \quad \text{for some } j \notin E_\mu,$$

as desired. $\square$

Proposition 54.6 cannot be strengthened to claim that there is some support vector u_{i_0} *and* some support vector v_{j_0}. We found examples for which the above condition fails for ν large enough.

The proof of Proposition 54.6 reveals that $(p_{sf} + q_{sf})/(p + q)$ is a critical value for ν. if this value is avoided we have the following corollary.

Theorem 54.7. *For every optimal solution* $(w, b, \eta, \epsilon, \xi)$ *of Problem* (SVM$_{s3}$) *with* $w \neq 0$ *and* $\eta > 0$*, if*

$$(p_{sf} + q_{sf})/(p+q) < \nu < 1,$$

then some u_{i_0} *or some* v_{j_0} *is a support vector.*

The proof proceeds by contradiction using Proposition 54.6 (for a very similar proof, see the proof of Theorem 54.3).

54.12 Solving SVM (SVM$_{s3}$) Using ADMM

In order to solve (SVM$_{s3}$) using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$\begin{aligned} \sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j &= K_m \\ \lambda_i + \alpha_i &= K_s, \quad i = 1, \ldots, p \\ \mu_j + \beta_j &= K_s, \quad j = 1, \ldots, q \end{aligned}$$

with $K_m = (p+q)K_s\nu$. This is the $(p+q+1) \times 2(p+q)$ matrix A given by

$$A = \begin{pmatrix} \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}.$$

We leave it as an exercise to prove that A has rank $p+q+1$. The right-hand side is

$$c = \begin{pmatrix} K_m \\ K_s \mathbf{1}_{p+q} \end{pmatrix}.$$

The symmetric positive semidefinite $(p+q) \times (p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix}, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $2(p+q)$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$, the $(p+q) \times (p+q)$ matrix P must be augmented with zero's to make it a $2(p+q) \times 2(p+q)$ matrix P_a given by

$$P_a = \begin{pmatrix} P & 0_{p+q,p+q} \\ 0_{p+q,p+q} & 0_{p+q,p+q} \end{pmatrix},$$

and similarly q is augmented with zeros as the vector

$$q_a = 0_{2(p+q)}.$$

The Matlab programs implementing the above method are given in Appendix B, Section B.3. We ran our program on the same input data points used in Section 54.8, namely

```
u16 = 10.1*randn(2,30)+7 ;
v16 = -10.1*randn(2,30)-7;
[~,~,~,~,~,~,w3] = runSVMs3b(0.365,rho,u16,v16,1/60)
```

We picked $K = 1/60$ and various values of ν starting with $\nu = 0.365$, which appears to be the smallest value for which the method converges; see Figure 54.16.

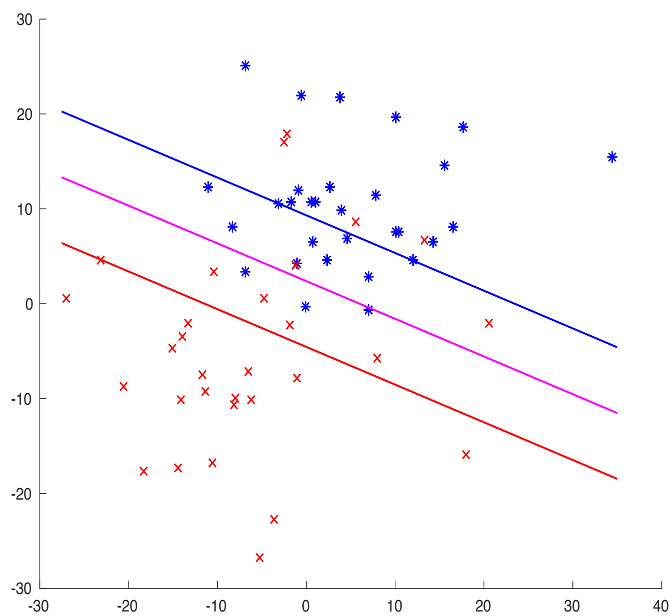


Figure 54.16: Running (SVM$_{s3}$) on two sets of 30 points; $\nu = 0.365$.

We have $p_f = 10, q_f = 10, p_m = 12$ and $q_m = 11$, as opposed to $p_f = 10, q_f = 11, p_m = 12, q_m = 12$, which was obtained by running (SVM$_{s2'}$); see Figure 54.11. A slightly narrower margin is achieved.

Next we ran our program with $\nu = 0.5$, see Figure 54.17. We have $p_f = 13, q_f = 16, p_m = 14$ and $q_m = 17$.

We also ran our program with $\nu = 0.71$, see Figure 54.18. We have $p_f = 21, q_f = 21, p_m = 22$ and $q_m = 22$. The value $\nu = 0.7$ is a singular value for which there are no support vectors and $\nu = (p_f + q_f)/(p + q)$.

Finally we ran our program with $\nu = 0.98$, see Figure 54.19. We have $p_f = 28, q_f = 30, p_m = 29$ and $q_m = 30$.

Because the term $(1/2)b^2$ is added to the objective function to be minimized, it turns out that (SVM_{s3}) yields values of b and η that are smaller than the values returned by ($\text{SVM}_{s2'}$). This is the reason why a smaller margin width could be obtained for $\nu = 0.365$. On the other hand, (SVM_{s3}) is unable to achieve as big a margin as ($\text{SVM}_{s2'}$) for values of $\nu \geq 0.97$, because the separating line produced by (SVM_{s3}) is lower than the the separating line produced by ($\text{SVM}_{s2'}$).

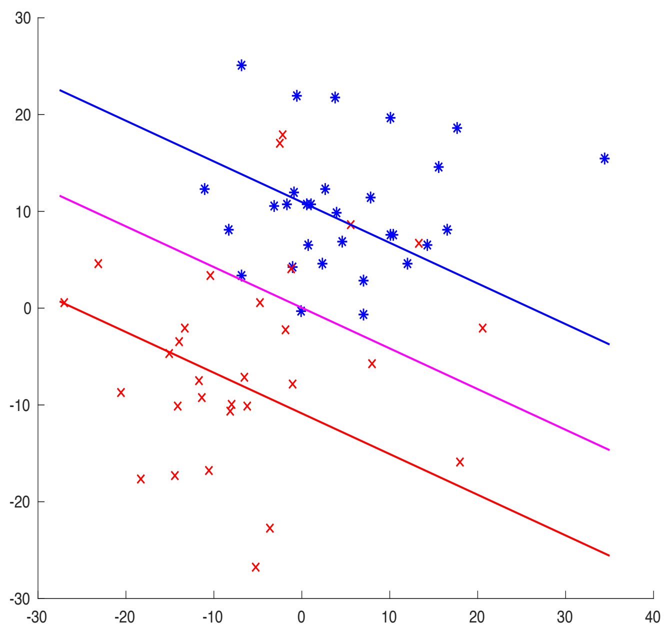


Figure 54.17: Running (SVM_{s3}) on two sets of 30 points; $\nu = 0.5$.

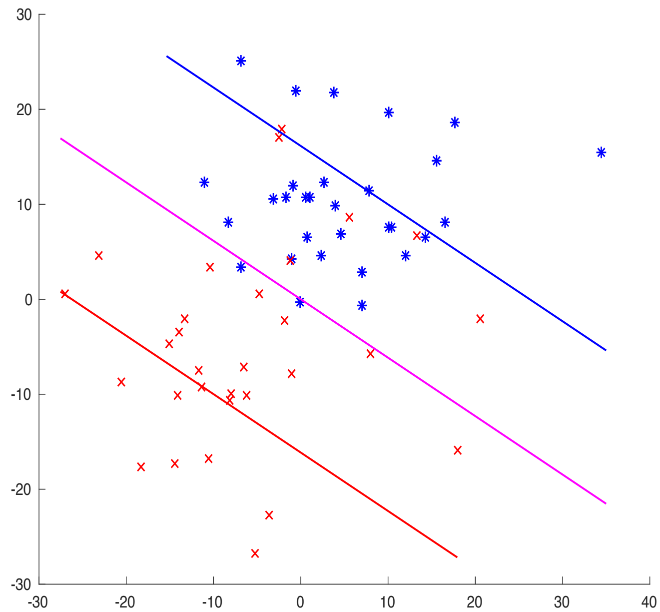


Figure 54.18: Running (SVM_{s3}) on two sets of 30 points; $\nu = 0.71$.

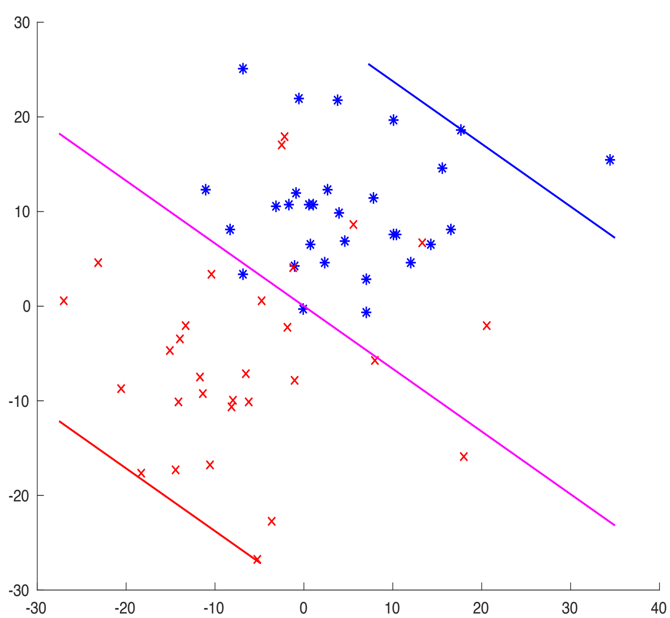


Figure 54.19: Running (SVM_{s3}) on two sets of 30 points; $\nu = 0.98$.

54.13 Soft Margin SVM; (SVM$_{s4}$)

In this section we consider the version of Problem (SVM$_{s2'}$) in which instead of using the function $K\left(\sum_{i=1}^{p}\epsilon_i + \sum_{j=1}^{q}\xi_j\right)$ as a regularizing function we use the quadratic function $K(\|\epsilon\|_2^2 + \|\xi\|_2^2)$.

Soft margin SVM (SVM$_{s4}$):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + (p+q)K_s\left(-\nu\eta + \frac{1}{p+q}(\epsilon^\top\epsilon + \xi^\top\xi)\right)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q\\
&\qquad \eta \geq 0,
\end{aligned}
$$

where ν and K_s are two given positive constants. As we saw earlier, theoretically, it is convenient to pick $K_s = 1/(p+q)$. When writing a computer program, it is preferable to assume that K_s is arbitrary. In this case ν needs to be replaced by $(p+q)K_s\nu$ in all the formulae obtained with $K_s = 1/(p+q)$.

The new twist with this formulation of the problem is that if $\epsilon_i < 0$, then the corresponding inequality $w^\top u_i - b \geq \eta - \epsilon_i$ implies the inequality $w^\top u_i - b \geq \eta$ obtained by setting ϵ_i to zero while reducing the value of $\|\epsilon\|^2$, and similarly if $\xi_j < 0$, then the corresponding inequality $-w^\top v_j + b \geq \eta - \xi_j$ implies the inequality $-w^\top v_j + b \geq \eta$ obtained by setting ξ_j to zero while reducing the value of $\|\xi\|^2$. Therefore, if (w, b, ϵ, ξ) is an optimal solution of Problem (SVM$_{s4}$), it is not necessary to restrict the slack variables ϵ_i and ξ_j to the nonnegative, which simplifies matters a bit. In fact, we will see that for an optimal solution, $\epsilon = \lambda/(2K_s)$ and $\xi = \mu/(2K_s)$. The variable η can also be determined by expressing that the duality gap is zero.

One of the advantages of this methods is that ϵ is determined by λ, ξ is determined by μ, and η and b are determined by λ and μ. This method *does not* require support vectors to compute b. We can omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$; see Section 54.14.

A drawback of Program (SVM$_{s4}$) is that for fixed K_s, the quantity $\delta = \eta/\|w\|$ and the hyperplanes $H_{w,b}, H_{w,b+\eta}$ and $H_{w,b-\eta}$ are independent of ν. This will be shown in Theorem 54.8. Thus this method is less flexible than (SVM$_{s2'}$) and (SVM$_{s3}$).

The Lagrangian is given by

$$
\begin{aligned}
L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \gamma) &= \frac{1}{2} w^\top w - \nu\eta + K_s(\epsilon^\top \epsilon + \xi^\top \xi) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&\quad - \epsilon^\top \lambda - \xi^\top \mu + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) - \gamma\eta \\
&= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu - \gamma) \\
&\quad + K_s(\epsilon^\top \epsilon + \xi^\top \xi) - \epsilon^\top \lambda - \xi^\top \mu + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu).
\end{aligned}
$$

To find the dual function $G(\lambda, \mu, \gamma)$ we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \gamma)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute $\nabla L_{w,\epsilon,\xi,b,\eta}$. The gradient $\nabla L_{w,\epsilon,\xi,b,\eta}$ is given by

$$
\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} w + X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ 2K_s \epsilon - \lambda \\ 2K_s \xi - \mu \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu - \gamma \end{pmatrix}.
$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$
\begin{aligned}
w &= -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \qquad (*_w) \\
2K_s \epsilon &= \lambda \\
2K_s \xi &= \mu \\
\mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu \\
\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= \nu + \gamma.
\end{aligned}
$$

The last two equations are identical to the last two equations obtained in Problem $(\mathrm{SVM}_{s2'})$. We can use the other equations to obtain the following expression for the dual function $G(\lambda, \mu, \gamma)$,

$$
\begin{aligned}
G(\lambda, \mu, \gamma) &= -\frac{1}{4K_s}(\lambda^\top \lambda + \mu^\top \mu) - \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.
\end{aligned}
$$

Consequently the dual program is equivalent to the minimization program

Dual of the Soft margin SVM (SVM_{s4}):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K_s} I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq \nu \\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q.
\end{aligned}
$$

The above program is similar to the program that was obtained for Problem ($\mathrm{SVM}_{s2'}$) but the matrix $X^\top X$ is replaced by the matrix $X^\top X + (1/2K_s)I_{p+q}$, which is positive definite since $K_s > 0$, and also the inequalities $\lambda_i \leq K_s$ and $\mu_j \leq K_s$ no longer hold.

It is shown in Section 54.14 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ. We obtain w from λ and μ, as in Problem ($\mathrm{SVM}_{s2'}$); namely,

$$
w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j.
$$

Since the variables ϵ_i and ξ_j are not restricted to be nonnegative we no longer have complementary slackness conditions involving them, but we know that

$$
\epsilon = \frac{\lambda}{2K_s}, \quad \xi = \frac{\mu}{2K_s}.
$$

Also since the constraints

$$
\sum_{i=1}^p \lambda_i \geq \frac{\nu}{2} \quad \text{and} \quad \sum_{j=1}^q \mu_j \geq \frac{\nu}{2}
$$

imply that there is some i_0 such that $\lambda_{i_0} > 0$ and some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ and $\xi_{j_0} > 0$, which means that at least two points are misclassified, so Problem (SVM_{s4}) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable.

Because $\epsilon_i = \lambda_i/(2K_s)$, $\xi_j = \mu_j/(2K_s)$, and there is no upper bound K_s on λ_i and μ_j, the classification of the points is simpler than in the previous cases.

(1) If $\lambda_i = 0$, then $\epsilon_i = 0$ and the inequality $w^\top u_i - b - \eta \geq 0$ holds. If equality holds then u_i is a support vector on the blue margin (the hyperplane $H_{w,b+\eta}$). Otherwise u_i is

in the blue open half-space bounded by the margin hyperplane $H_{w,b+\eta}$ (not containing the separating hyperplane $H_{w,b}$). See Figure 54.20.

Similarly, if $\mu_j = 0$, then $\xi_j = 0$ and the inequality $-w^\top v_j + b - \eta \geq$ holds. If equality holds then v_j is a support vector on the red margin (the hyperplane $H_{w,b-\eta}$). Otherwise v_j is in the red open half-space bounded by the margin hyperplane $H_{w,b-\eta}$ (not containing the separating hyperplane $H_{w,b}$). See Figure 54.20.

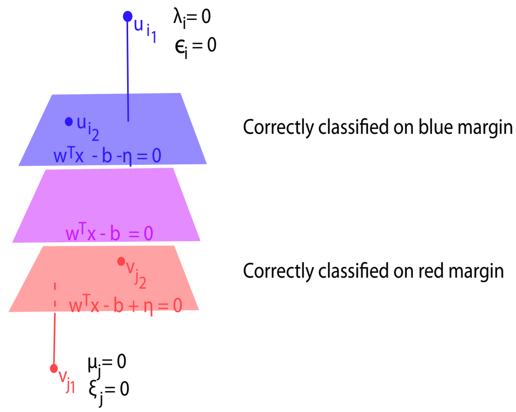


Figure 54.20: When $\lambda_i = 0$, u_i is correctly classified on or outside the blue margin. When $\mu_j = 0$, v_j is correctly classified on or outside outside the red margin.

(2) If $\lambda_i > 0$, then $\epsilon_i = \lambda_i/(2K_s) > 0$. The corresponding constraint is active, so we have

$$w^\top u_i - b = \eta - \epsilon_i.$$

If $\epsilon_i \leq \eta$, then the points u_i is inside the slab bounded by the blue margin hyperplane $H_{w,b+\eta}$ and the separating hyperplane $H_{w,b}$. If $\epsilon_i > \eta$, then the point u_i belongs to the open half-space bounded by the separating hyperplane and containing the red margin hyperplane (the red side); it is misclassified. See Figure 54.21.

Similarly, if $\mu_j > 0$, then $\xi_j = \mu_j/(2K_s) > 0$. The corresponding constraint is active, so we have

$$-w^\top v_j + b = \eta - \xi_j.$$

If $\xi_j \leq \eta$, then the points v_j is inside the slab bounded by the red margin hyperplane $H_{w,b-\eta}$ and the separating hyperplane $H_{w,b}$. If $\xi_j > \eta$, then the point v_j belongs to the open half-space bounded by the separating hyperplane and containing the blue margin hyperplane (the blue side); it is misclassified. See Figure 54.21.

We can use the fact that the duality gap is 0 to find η. We have

$$\frac{1}{2}w^\top w - \nu\eta + K_s(\epsilon^\top \epsilon + \xi^\top \xi) = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

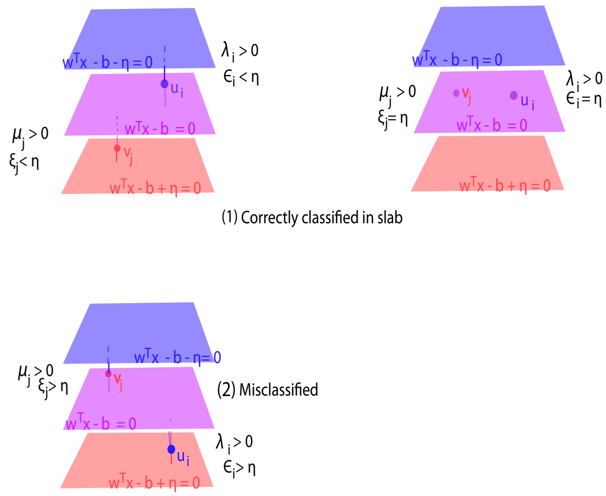


Figure 54.21: The classification of points for SVM$_{s4}$ when the Lagrange multipliers are positive. The left illustration of Figure (1) is when u_i is inside the margin yet still on the correct side of the separating hyperplane $w^\top x - b = 0$. Similarly, v_j is inside the margin on the correct side of the separating hyperplane. The right illustration depicts u_i and v_j on the separating hyperplane. Figure (2) illustrations a misclassification of u_i and v_j.

and since

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

we get

$$\begin{aligned} \nu\eta &= K_s(\epsilon^\top \epsilon + \xi^\top \xi) + \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{4K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &= \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

The above confirms that at optimality we have $\eta \geq 0$.

Remark: If we do not assume that $K_s = 1/(p+q)$, then the above formula must be replaced by

$$(p+q)K_s \nu\eta = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

Since η is determined independently of the existence of support vectors, the margin hyperplane $H_{w,b+\eta}$ may not contain any point u_i and the margin hyperplane $H_{w,b-\eta}$ may not contain any point v_j.

We can solve for b using some active constraint corresponding to any i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$ (by a previous remark, the constraints imply that such i_0 and

j_0 must exist). To improve numerical stability we average over the following sets of indices. Let I_λ and I_μ be the set of indices given by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \dots, p\} \mid \lambda_i > 0\} \\ I_\mu &= \{j \in \{1, \dots, q\} \mid \mu_j > 0\}, \end{aligned}$$

and let $p_m = |I_\lambda|$ and $q_m = |I_\mu|$. We obtain the formula

$$b = \left(w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / p_m + \left(\sum_{j \in I_\mu} v_j \right) / q_m \right) + \left(\sum_{i \in I_\lambda} \epsilon_i \right) / p_m - \left(\sum_{j \in I_\mu} \xi_j \right) / q_m \right) / 2.$$

We now prove that for a fixed K_s, the solution to Problem (SVM$_{s4}$) is unique and independent of the value of ν.

Theorem 54.8. *For K_s and ν fixed, if Problem* (SVM$_{s4}$) *succeeds, then it has a unique solution. If Problem* (SVM$_{s4}$) *succeeds and returns $(\lambda, \mu, \eta, w, b)$ for the value ν and $(\lambda^\kappa, \mu^\kappa, \eta^\kappa,$ w^κ, $b^\kappa)$ for the value $\kappa\nu$ with $\kappa > 0$, then*

$$\lambda^\kappa = \kappa\lambda, \;\; \mu^\kappa = \kappa\mu, \;\; \eta^\kappa = \kappa\eta, \;\; w^\kappa = \kappa w, \;\; b^\kappa = \kappa b.$$

As a consequence, $\delta = \eta / \|w\| = \eta^\kappa / \|w^\kappa\| = \delta^\kappa$, and the hyperplanes $H_{w,b}, H_{w,b+\eta}$ and $H_{w,b-\eta}$ are independent of ν.

Proof. We already observed that for an optimal solution with $\eta > 0$, we have $\gamma = 0$. This means that (λ, μ) is a solution of the problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\ &\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\ &\qquad \lambda_i \geq 0, \quad i = 1, \dots, p \\ &\qquad \mu_j \geq 0, \quad j = 1, \dots, q. \end{aligned}$$

Since $K_s > 0$ and $X^\top X$ is symmetric positive semidefinite, the matrix $P = X^\top X + \frac{1}{2K_s} I_{p+q}$ is symmetric positive definite. Let $\Omega = \mathbb{R}^{p+q}$ and let U be the convex set given by

$$U = \left\{ \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \in \mathbb{R}_+^{p+q} \;\middle|\; \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ (p+q) K_s \nu \end{pmatrix} \right\}.$$

Since the matrix P is symmetric positive definite, the functional

$$F(\lambda, \mu) = -G(\lambda, \mu) = \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda \\ \mu\end{pmatrix}$$

is strictly convex and U is convex, so by Theorem 40.13(2,4), if it has a minimum, then it is unique. Consider the convex set

$$U^\kappa = \left\{ \begin{pmatrix}\lambda \\ \mu\end{pmatrix} \in \mathbb{R}^{p+q}_+ \;\middle|\; \begin{pmatrix}\mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top\end{pmatrix} \begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \begin{pmatrix}0 \\ (p+q)K_s\kappa\nu\end{pmatrix} \right\}.$$

Observe that

$$\kappa U = \left\{ \begin{pmatrix}\kappa\lambda \\ \kappa\mu\end{pmatrix} \in \mathbb{R}^{p+q}_+ \;\middle|\; \begin{pmatrix}\mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top\end{pmatrix} \begin{pmatrix}\kappa\lambda \\ \kappa\mu\end{pmatrix} = \begin{pmatrix}0 \\ (p+q)K_s\kappa\nu\end{pmatrix} \right\} = U^\kappa.$$

By Theorem 40.13(3), $(\lambda, \mu) \in U$ is a minimum of F over U iff

$$dF_{\lambda,\mu}\begin{pmatrix}\lambda' - \lambda \\ \mu' - \mu\end{pmatrix} \geq 0 \quad \text{for all} \quad \begin{pmatrix}\lambda' \\ \mu'\end{pmatrix} \in U.$$

Since

$$dF_{\lambda,\mu}\begin{pmatrix}\lambda' - \lambda \\ \mu' - \mu\end{pmatrix} = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda' - \lambda \\ \mu' - \mu\end{pmatrix}$$

the above conditions are equivalent to

$$\begin{aligned}
\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda' - \lambda \\ \mu' - \mu\end{pmatrix} &\geq 0 \\
\begin{pmatrix}\mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top\end{pmatrix} \begin{pmatrix}\lambda \\ \mu\end{pmatrix} &= \begin{pmatrix}0 \\ (p+q)K_s\nu\end{pmatrix} \\
\lambda, \lambda' \in \mathbb{R}^p_+,\ \mu, \mu' &\in \mathbb{R}^q_+.
\end{aligned}$$

Since $\kappa > 0$, by multiplying the above inequality by κ^2 and the equations by κ, the following conditions hold:

$$\begin{aligned}
\begin{pmatrix}\kappa\lambda^\top & \kappa\mu^\top\end{pmatrix} P \begin{pmatrix}\kappa\lambda' - \kappa\lambda \\ \kappa\mu' - \kappa\mu\end{pmatrix} &\geq 0 \\
\begin{pmatrix}\mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top\end{pmatrix} \begin{pmatrix}\kappa\lambda \\ \kappa\mu\end{pmatrix} &= \begin{pmatrix}0 \\ (p+q)K_s\kappa\nu\end{pmatrix} \\
\kappa\lambda, \kappa\lambda' \in \mathbb{R}^p_+,\ \kappa\mu, \kappa\mu' &\in \mathbb{R}^q_+.
\end{aligned}$$

By Theorem 40.13(3), $(\kappa\lambda, \kappa\mu) \in U^\kappa$ is a minimum of F over U^κ, and because F is strictly convex and U^κ is convex, if F has a minimum over U^κ, then $(\kappa\lambda, \kappa\mu) \in U^\kappa$ is the unique minimum. Therefore, $\lambda^\kappa = \kappa\lambda$, $\mu^\kappa = \kappa\mu$.

Since w is given by the equation

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

and since we just showed that $\lambda^\kappa = \kappa\lambda$, $\mu^\kappa = \kappa\mu$, we deduce that $w^\kappa = \kappa w$.

We showed earlier that η is given by the equation

$$(p+q)K_s\nu\eta = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

If we replace ν by $\kappa\nu$, since λ is replaced by $\kappa\lambda$ and μ by $\kappa\nu$, we see that $\eta^\kappa = \kappa\eta$. Finally, b is given by the equation

$$b = \frac{w^\top(u_{i_0} + v_{j_0}) + \epsilon_{i_0} - \xi_{j_0}}{2}$$

for and i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$. If λ is replaced by $\kappa\lambda$ and μ by $\kappa\mu$, since $\epsilon = \lambda/(2K_s)$ and $\xi = \mu/(2K_s)$, we see that ϵ is replaced by $\kappa\epsilon$ and ξ by $\kappa\xi$, so $b^\kappa = \kappa b$.

Since $w^\kappa = \kappa w$ and $\eta^\kappa = \kappa\eta$ we obtain $\delta = \eta/\left\|w\right\| = \eta^\kappa/\left\|w^\kappa\right\| = \delta^\kappa$. Since $w^\kappa = \kappa w$, $\eta^\kappa = \kappa\eta$ and $b^\kappa = \kappa b$, the normalized equations of the hyperplanes $H_{w,b}$, $H_{w,b+\eta}$ and $H_{w,b-\eta}$ (obtained by dividing by $\|w\|$) are all identical, so the hyperplanes $H_{w,b}$, $H_{w,b+\eta}$ and $H_{w,b-\eta}$ are independent of ν. $\square$

The width of the slab is controlled by K. The larger K is the smaller is the width of the slab. Theoretically, since this method does not rely on support vectors to compute b, it cannot fail if a solution exists, but in practice the quadratic solver does not converge for values of K that are too large. However, the method handles very small values of K, which can yield slabs of excessive width.

The "kernelized" version of Problem (SVM$_{s4}$) is the following:

Soft margin kernel SVM (SVM$_{s4}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\langle w, w\rangle - \nu\eta + K_s(\epsilon^\top\epsilon + \xi^\top\xi) \\
&\text{subject to} \\
&\qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\
&\qquad -\langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q \\
&\qquad \eta \geq 0,
\end{aligned}$$

with $K_s = 1/(p+q)$.

By going over the derivation of the dual program, we obtain

Dual of the Soft margin kernel SVM (SVM_{s4}):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \frac{1}{2K_s} I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq \nu \\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q,
\end{aligned}
$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1. Then w, b, and $f(x)$ are obtained exactly as in Section 54.5.

54.14 Solving SVM (SVM_{s4}) Using ADMM

In order to solve (SVM_{s4}) using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$
\begin{aligned}
\sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0 \\
\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j - \gamma &= K_m,
\end{aligned}
$$

with $K_m = (p+q)K_s\nu$. This is the $2 \times (p+q+1)$ matrix A given by

$$
A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0 \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & -1 \end{pmatrix}.
$$

We leave it as an exercise to prove that A has rank 2. The right-hand side is

$$
c = \begin{pmatrix} 0 \\ K_m \end{pmatrix}.
$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$
P = X^\top X + \frac{1}{2K_s} I_{p+q}, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},
$$

and

$$
q = 0_{p+q}.
$$

Since there are $p+q+1$ Lagrange multipliers (λ, μ, γ), the $(p+q) \times (p+q)$ matrix P must be augmented with zero's to make it a $(p+q+1) \times (p+q+1)$ matrix P_a given by

$$P_a = \begin{pmatrix} X^\top X & 0_{p+q} \\ 0_{p+q}^\top & 0 \end{pmatrix},$$

and similarly q is augmented with zeros as the vector $q_a = 0_{p+q+1}$.

As in Section 54.8, since $\eta \geq 0$ for an optimal solution, we can drop the constraint $\eta \geq 0$ from the primal problem. In this case, there are $p+q$ Lagrange multipliers (λ, μ). It is easy to see that the objective function of the dual is unchanged and the set of constraints is

$$\begin{aligned} \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j &= 0 \\ \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j &= K_m, \end{aligned}$$

with $K_m = (p+q)K_s\nu$. The matrix corresponding to the above equations is the $2 \times (p+q)$ matrix A_2 given by

$$A_2 = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top \end{pmatrix}.$$

We leave it as an exercise to prove that A_2 has rank 2. The right-hand side is

$$c_2 = \begin{pmatrix} 0 \\ K_m \end{pmatrix}.$$

The symmetric positive semidefinite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X + \frac{1}{2K_s} I_{p+q}, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $p+q$ Lagrange multipliers (λ, μ), the $(p+q) \times (p+q)$ matrix P need not be augmented with zero's, so $P_{2a} = P$ and similarly $q_{2a} = 0_{p+q}$.

We ran our `Matlab` implementation of the above version of (SVM_{s4}) on the data set of Section 54.12. Since the value of ν is irrelevant, we picked $\nu = 1$. First we ran our program with $K = 190$; see Figure 54.22. We have $p_m = 23$ and $q_m = 18$. The program does not converge for $K \geq 200$.

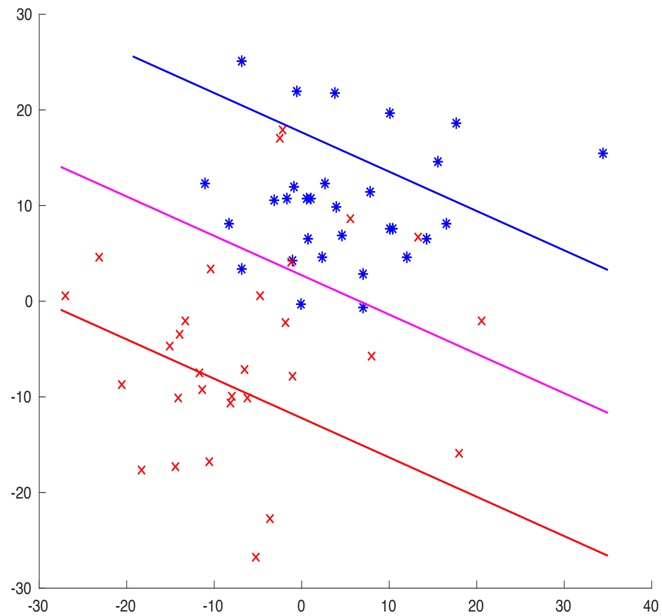


Figure 54.22: Running (SVM$_{s4}$) on two sets of 30 points; $K = 190$.

Our second run was made with $K = 1/12000$; see Figure 54.23. We have $p_m = 30$ and $q_m = 30$ and we see that the width of the slab is a bit excessive. This example demonstrates that the margin lines need not contain data points.

54.15 Soft Margin SVM; (SVM$_{s5}$)

In this section we consider the version of Problem (SVM$_{s4}$) in which we add the term $(1/2)b^2$ to the objective function. We also drop the constraint $\eta \geq 0$ which is redundant.

Soft margin SVM (SVM$_{s5}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + (p+q)K_s\left(-\nu\eta + \frac{1}{p+q}(\epsilon^\top\epsilon + \xi^\top\xi)\right) \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q, \end{aligned}$$

where ν and K_s are two given positive constants. As we saw earlier, it is convenient to pick $K_s = 1/(p+q)$. When writing a computer program, it is preferable to assume that K_s is arbitrary. In this case ν must be replaced by $(p+q)K_s\nu$ in all the formulae.

One of the advantages of this methods is that ϵ is determined by λ, ξ is determined by μ (as in (SVM$_{s4}$)), and both η and b determined by λ and μ. As the previous method, this

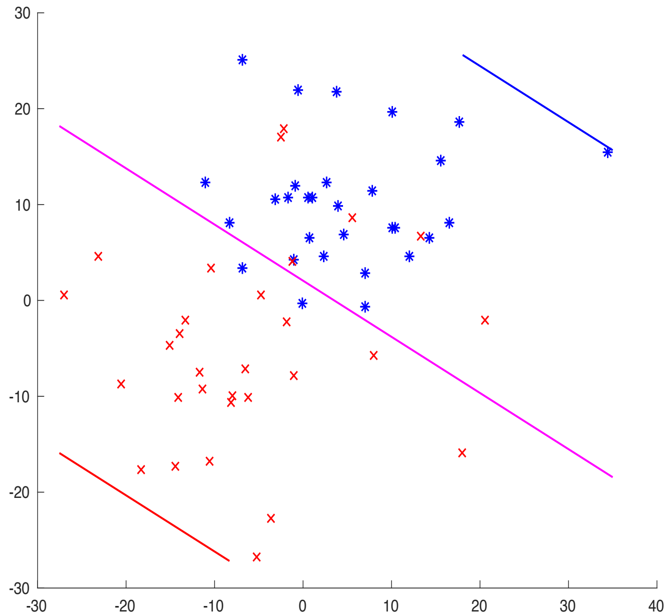


Figure 54.23: Running (SVM$_{s4}$) on two sets of 30 points; $K = 1/12000$.

method *does not* require support vectors to compute b. We can omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$.

A drawback of Program (SVM$_{s5}$) is that for fixed K_s, the quantity $\delta = \eta/\|w\|$ and the hyperplanes $H_{w,b}, H_{w,b+\eta}$ and $H_{w,b-\eta}$ are independent of ν. This will be shown in Theorem 54.9. Thus this method is less flexible than (SVM$_{s2'}$) and (SVM$_{s3}$).

The Lagrangian is given by

$$\begin{aligned}
L(w, \epsilon, \xi, b, \eta, \lambda, \mu) &= \frac{1}{2} w^\top w + \frac{1}{2} b^2 - \nu\eta + K_s(\epsilon^\top \epsilon + \xi^\top \xi) + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\
&\quad - \epsilon^\top \lambda - \xi^\top \mu + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu) \\
&= \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \eta(\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu) \\
&\quad + K_s(\epsilon^\top \epsilon + \xi^\top \xi) - \epsilon^\top \lambda - \xi^\top \mu + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) + \frac{1}{2} b^2.
\end{aligned}$$

To find the dual function $G(\lambda, \mu)$ we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu)$ with respect to w, ϵ, ξ, b, and η. Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$, so we compute $\nabla L_{w,\epsilon,\xi,b,\eta}$. The gradient $\nabla L_{w,\epsilon,\xi,b,\eta}$ is given by

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} w + X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ 2K_s\epsilon - \lambda \\ 2K_s\xi - \mu \\ b + \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - \nu \end{pmatrix}.$$

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$\begin{aligned} w &= -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, && (*_w) \\ 2K_s\epsilon &= \lambda \\ 2K_s\xi &= \mu \\ b &= -(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu) \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu &= \nu. \end{aligned}$$

As we said earlier, both w an b are determined by λ and μ. We can use the equations to obtain the following expression for the dual function $G(\lambda, \mu, \gamma)$,

$$\begin{aligned} G(\lambda, \mu, \gamma) &= -\frac{1}{4K_s}(\lambda^\top \lambda + \mu^\top \mu) - \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \frac{b^2}{2} \\ &= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

Consequently the dual program is equivalent to the minimization program

Dual of the Soft margin SVM (SVM$_{s5}$):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\ &\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\ &\qquad \mu_j \geq 0, \quad j = 1, \ldots, q. \end{aligned}$$

It is shown in Section 54.16 how the dual program is solved using ADMM from Section 52.6. If the primal problem is solvable, this yields solutions for λ and μ.

The constraint

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu$$

implies that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$, so we have $\epsilon_{i_0} > 0$ or $\xi_{j_0} > 0$, which means that at least one point is misclassified. Thus Problem (SVM$_{s5}$) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable.

We can use the fact that the duality gap is 0 to find η. We have

$$\begin{aligned}\frac{1}{2}w^\top w + \frac{b^2}{2} - \nu\eta + K_s(\epsilon^\top\epsilon + \xi^\top\xi) \\ = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix},\end{aligned}$$

so we get

$$\begin{aligned}\nu\eta &= K_s(\epsilon^\top\epsilon + \xi^\top\xi) + \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{4K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ &= \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}.\end{aligned}$$

The above confirms that at optimality we have $\eta \geq 0$.

Remark: If we do not assume that $K_s = 1/(p+q)$, then the above formula must be replaced by

$$(p+q)K_s\nu\eta = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}.$$

There is a version of Theorem 54.8 stating that for a fixed K_s, the solution to Problem (SVM$_{s5}$) is unique and independent of the value of ν.

Theorem 54.9. *For K_s and ν fixed, if Problem* (SVM$_{s5}$) *succeeds then it has a unique solution. If Problem* (SVM$_{s5}$) *succeeds and returns $(\lambda, \mu, \eta, w, b)$ for the value ν and $(\lambda^\kappa, \mu^\kappa, \eta^\kappa, w^\kappa, b^\kappa)$ for the value $\kappa\nu$ with $\kappa > 0$, then*

$$\lambda^\kappa = \kappa\lambda,\ \ \mu^\kappa = \kappa\mu,\ \ \eta^\kappa = \kappa\eta,\ \ w^\kappa = \kappa w,\ \ b^\kappa = \kappa b.$$

As a consequence, $\delta = \eta/\left\|w\right\| = \eta^\kappa/\left\|w^\kappa\right\| = \delta^\kappa$, and the hyperplanes $H_{w,b}$, $H_{w,b+\eta}$ and $H_{w,b-\eta}$ are independent of ν.

Proof. The proof is an easy adaptation of the proof of Theorem 54.8 so we only give a sketch. The two crucial points are that the matrix

$$P = X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K_s}I_{p+q}$$

is symmetric positive definite and that we have the single equational constraint

$$\mathbf{1}_p^\top\lambda + \mathbf{1}_q^\top\mu = (p+q)K_s\nu$$

defining the convex set

$$U = \left\{ \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \in \mathbb{R}_+^{p+q} \mid \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu = (p+q)K_s\nu \right\}.$$

The proof is essentially the proof of 54.8 using the above SPD matrix and convex set. $\square$

The "kernelized" version of Problem (SVM_{s5}) is the following:

Soft margin kernel SVM (SVM_{s5}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + \frac{1}{2}b^2 - \nu\eta + K_s(\epsilon^\top \epsilon + \xi^\top \xi) \\ &\text{subject to} \\ &\qquad \langle w, \varphi(u_i)\rangle - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\ &\qquad -\langle w, \varphi(v_j)\rangle + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q, \end{aligned}$$

with $K_s = 1/(p+q)$.

Tracing through the derivation of the dual program, we obtain

Dual of the Soft margin kernel SVM (SVM_{s5}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(\mathbf{K} + \begin{pmatrix} \mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top \end{pmatrix} + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\ &\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\ &\qquad \mu_j \geq 0, \quad j = 1, \ldots, q, \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1. Then w, b, and $f(x)$ are obtained exactly as in Section 54.13.

54.16 Solving SVM (SVM_{s5}) Using ADMM

In order to solve (SVM_5) using ADMM we need to write the matrix corresponding to the constraints in equational form,

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = K_m,$$

with $K_m = (p+q)K_s\nu$. This is the $1 \times (p+q)$ matrix A given by

$$A = \begin{pmatrix} \mathbf{1}_p^\top & \mathbf{1}_q^\top \end{pmatrix}.$$

Obviously, A has rank 1. The right-hand side is

$$c = K_m.$$

The symmetric positive definite $(p+q)\times(p+q)$ matrix P defining the quadratic functional is

$$P = X^\top X + \begin{pmatrix} \mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top \end{pmatrix} + \frac{1}{2K_s} I_{p+q}, \quad \text{with} \quad X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix},$$

and

$$q = 0_{p+q}.$$

Since there are $p+q$ Lagrange multipliers (λ, μ), the $(p+q)\times(p+q)$ matrix P does not have to be augmented with zero's.

We ran our `Matlab` implementation of the above version of (SVM$_{s5}$) on the data set of Section 54.14. Since the value of ν is irrelevant, we picked $\nu = 1$. First we ran our program with $K = 190$; see Figure 54.24. We have $p_m = 23$ and $q_m = 18$. The program does not converge for $K \geq 200$.

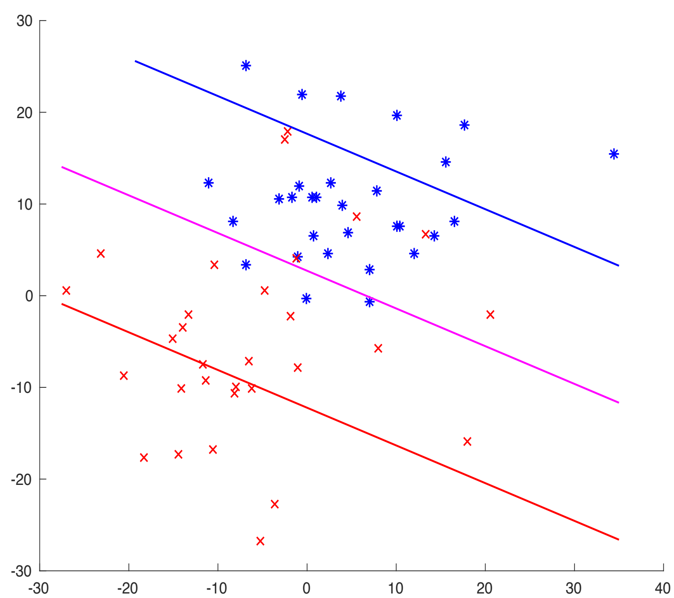


Figure 54.24: Running (SVM$_{s5}$) on two sets of 30 points; $K = 190$.

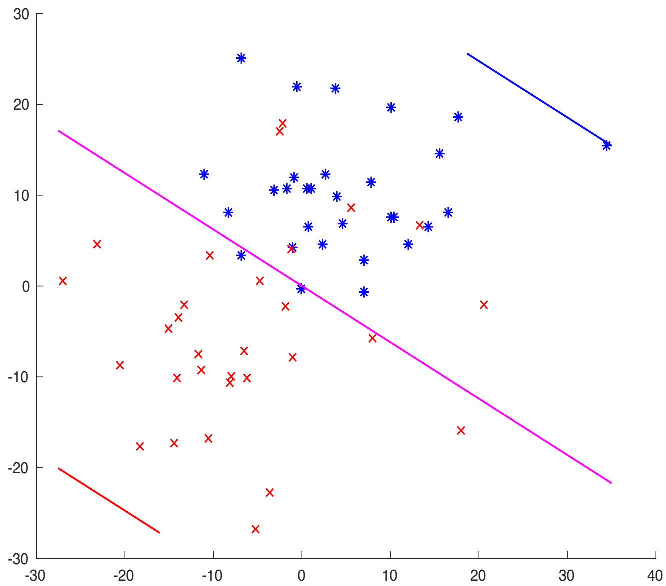


Figure 54.25: Running (SVM_{s5}) on two sets of 30 points; $K = 1/13000$.

Our second run was made with $K = 1/13000$; see Figure 54.25. We have $p_m = 30$ and $q_m = 30$ and we see that the width of the slab is a bit excessive. This example demonstrates that the margin lines need not contain data points.

Method (SVM_{s5}) always returns a value for b and η smaller than the value returned by (SVM_{s4}) (because of the term $(1/2)b^2$ added to the objective function) but in this example the difference is too small to be noticed.

54.17 Summary and Comparison of the SVM Methods

In this chapter we considered six variants for solving the soft margin binary classification problem for two sets of points $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$ using support vector classification methods. The objective is to find a separating hyperplane $H_{w,b}$ of equation $w^\top x - b = 0$. We also try to find two "margin hyperplanes" $H_{w,b+\delta}$ of equation $w^\top x - b - \delta = 0$ (the blue margin hyperplane) and $H_{w,b-\delta}$ of equation $w^\top x - b + \delta = 0$ (the red margin hyperplane) such that δ is as big as possible and yet the number of misclassified points is minimized, which is achieved by allowing an error $\epsilon_i \geq 0$ for every point u_i, in the sense that the constraint

$$w^\top u_i - b \geq \delta - \epsilon_i$$

should hold, and an error $\xi_j \geq 0$ for every point v_j, in the sense that the constraint

$$-w^\top v_j + b \geq \delta - \xi_j$$

should hold.

The goal is to design an objective function that minimizes ϵ and ξ and maximizes δ. The optimization problem should also solve for w and b, and for this some constraint has to be placed on w. Another goal is to try to use the dual program to solve the optimization problem, because the solutions involve inner products, and thus the problem is amenable to a generalization using kernel functions.

The first attempt, which is to use the objective function

$$J(w, \epsilon, \xi, b, \delta) = -\delta + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$$

and the constraint $w^\top w \leq 1$, does not work very well because this constraint needs to be guarded by a Lagrange multiplier $\gamma \geq 0$, and as a result, minimizing the Lagrangian L to find the dual function G gives an equation for solving w of the form

$$2\gamma w = -X^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

but if the sets $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$ are not linearly separable, then an optimal solution may occurs for $\gamma = 0$, in which case it is impossible to determine w. This is Problem (SVM$_{s1}$) considered in Section 54.1.

Soft margin SVM (SVM$_{s1}$):

$$\begin{aligned}
&\text{minimize} \quad -\delta + K\left(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\right)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \delta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \delta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q\\
&\qquad w^\top w \leq 1.
\end{aligned}$$

It is customary to write $\ell = p + q$.

It is shown in Section 54.1 that the dual program is equivalent to the following minimization program:

Dual of the Soft margin SVM (SVM_{s1}):

$$\begin{aligned}
&\text{minimize} \quad \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = 1 \\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q.
\end{aligned}$$

The points u_i and v_j are naturally classified in terms of the values of λ_i and μ_j. The numbers of points in each category have a direct influence on the choice of the parameter K. Let us summarize some of the keys items from Definition 54.1.

The vectors u_i on the blue margin $H_{w,b+\delta}$ and the vectors v_j on the red margin $H_{w,b-\delta}$ are called *support vectors*. Support vectors correspond to vectors u_i for which $w^\top u_i - b - \delta = 0$ (which implies $\epsilon_i = 0$), and vectors v_j for which $w^\top v_j - b + \delta = 0$ (which implies $\xi_j = 0$). Support vectors u_i such that $0 < \lambda_i < K$ and support vectors v_j such that $0 < \mu_j < K$ are *support vectors of type 1*. Support vectors of type 1 play a special role so we denote the sets of indices associated with them by

$$\begin{aligned}
I_\lambda &= \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K\} \\
I_\mu &= \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K\}.
\end{aligned}$$

We denote their cardinalities by $numsvl_1 = |I_\lambda|$ and $numsvm_1 = |I_\mu|$.

The vectors u_i for which $\lambda_i = K$ and the vectors v_j for which $\mu_j = K$ are said to *fail the margin*. The sets of indices associated with the vectors failing the margin are denoted by

$$\begin{aligned}
K_\lambda &= \{i \in \{1, \ldots, p\} \mid \lambda_i = K\} \\
K_\mu &= \{j \in \{1, \ldots, q\} \mid \mu_j = K\}.
\end{aligned}$$

We denote their cardinalities by $p_f = |K_\lambda|$ and $q_f = |K_\mu|$.

Vectors u_i such that $\lambda_i > 0$ and vectors v_j such that $\mu_j > 0$ are said to *have margin at most δ*. The sets of indices associated with these vectors are denoted by

$$\begin{aligned}
I_{\lambda>0} &= \{i \in \{1, \ldots, p\} \mid \lambda_i > 0\} \\
I_{\mu>0} &= \{j \in \{1, \ldots, q\} \mid \mu_j > 0\}.
\end{aligned}$$

We denote their cardinalities by $p_m = |I_{\lambda>0}|$ and $q_m = |I_{\mu>0}|$.

Obviously, $p_f \leq p_m$ and $q_f \leq q_m$. There are $p - p_m$ points u_i classified correctly on the blue side and outside the δ-slab and there are $q - q_m$ points v_j classified correctly on the red side and outside the δ-slab. Intuitively a blue point that fails the margin is on the wrong side of the blue margin and a red point that fails the margin is on the wrong side of the red margin.

It can be shown that that K must be chosen so that

$$\max\left\{\frac{1}{2p_m}, \frac{1}{2q_m}\right\} \leq K \leq \min\left\{\frac{1}{2p_f}, \frac{1}{2q_f}\right\}.$$

If the optimal value is 0, then $\gamma = 0$ and $X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = 0$, so in this case it is not possible to determine w. However, if the optimal value is > 0, then once a solution for λ and μ is obtained, we have

$$\begin{aligned}
\gamma &= \frac{1}{2}\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}\right)^{1/2} \\
w &= \frac{1}{2\gamma}\left(\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j\right),
\end{aligned}$$

so we get

$$w = \frac{\displaystyle\sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j}{\left(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}\right)^{1/2}},$$

If the following mild hypothesis holds, then b and δ can be found.

Standard Margin Hypothesis for (SVM$_{s1}$). There is some index i_0 such that $0 < \lambda_{i_0} < K$ and there is some index j_0 such that $0 < \mu_{j_0} < K$. This means that some u_{i_0} is a support vector of type 1 on the blue margin, and some v_{j_0} is a support vector of type 1 on the red margin.

If the **Standard Margin Hypothesis** for (SVM$_{s1}$) holds, then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and we have the active equations

$$w^\top u_{i_0} - b = \delta \quad \text{and} \quad -w^\top v_{j_0} + b = \delta,$$

and we obtain the value of b and δ as

$$\begin{aligned}
b &= \frac{1}{2} w^\top (u_{i_0} + v_{j_0}) \\
\delta &= \frac{1}{2} w^\top (u_{i_0} - v_{j_0}).
\end{aligned}$$

The second more successful approach is to add the term $(1/2)w^\top w$ to the objective function and to drop the constraint $w^\top w \leq 1$. There are several variants of this method, depending on the choice of the regularizing term involving ϵ and ξ (linear or quadratic), how the margin is dealt with (implicitly with the term 1 or explicitly with a term η), and whether the term $(1/2)b^2$ is added to the objective function or not.

These methods all share the property that if the primal problem has an optimal solution with $w \neq 0$, then the dual problem always determines w, and then under mild conditions which we call standard margin hypotheses, b and η can be determined. Then ϵ and ξ can be determined using the constraints that are active. When $(1/2)b^2$ is added to the objective function, b is determined by the equation

$$b = -(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu).$$

All these problems are convex and the constraints are qualified, so the duality gap is zero, and if the primal has an optimal solution with $w \neq 0$, then it follows that $\eta \geq 0$.

We now consider five variants in more details.

(1) **Basic soft margin SVM**: (SVM$_{s2}$).

This is the optimization problem in which the regularization term $K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$ is linear and the margin δ is given by $\delta = 1/\left\|w\right\|$:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} w^\top w + K \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad - w^\top v_j + b \geq 1 - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

This problem is the classical one discussed in all books on machine learning or pattern analysis, for instance Vapnik [182], Bishop [23], and Shawe–Taylor and Christianini [159]. It is shown in Section 54.3 that the dual program is

Dual of the Basic soft margin SVM: (SVM$_{s2}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q.
\end{aligned}$$

We can use the dual program to solve the primal. Once $\lambda \geq 0, \mu \geq 0$ have been found, w is given by

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j,$$

but b is not determined by the dual.

The complementary slackness conditions imply that if $\epsilon_i > 0$, then $\lambda_i = K$, and if $\xi_j > 0$, then $\mu_j = K$. Consequently, if $\lambda_i < K$, then $\epsilon_i = 0$ and u_i is correctly classified, and similarly if $\mu_j < K$, then $\xi_j = 0$ and v_j is correctly classified.

A priori nothing prevents the situation where $\lambda_i = K$ for all nonzero λ_i or $\mu_j = K$ for all nonzero μ_j. If this happens, we can rerun the optimization method with a larger value of K. If the following mild hypothesis holds then b can be found.

Standard Margin Hypothesis for (SVM_{s2}). There is some support vector u_{i_0} of type 1 on the blue margin, and some support vector v_{j_0} of type 1 on the red margin.

If the **Standard Margin Hypothesis** for (SVM_{s2}) holds then $\epsilon_{i_0} = 0$ and $\mu_{j_0} = 0$, and then we have the active equations

$$w^\top u_{i_0} - b = 1 \quad \text{and} \quad -w^\top v_{j_0} + b = 1,$$

and we obtain

$$b = \frac{1}{2} w^\top (u_{i_0} + v_{j_0}).$$

(2) **Basic Soft margin ν-SVM Problem** ($\mathrm{SVM}_{s2'}$).

This a generalization of Problem (SVM_{s2}) for a version of the soft margin SVM coming from Problem (SVM_{h2}), obtained by adding an extra degree of freedom, namely instead of the margin $\delta = 1/\|w\|$, we use the margin $\delta = \eta/\|w\|$ where η is some positive constant that we wish to maximize. To do so, we add a term $-K_m\eta$ to the objective function. We have the following optimization problem:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\
&\qquad \eta \geq 0,
\end{aligned}$$

where $K_m > 0$ and $K_s > 0$ are fixed constants that can be adjusted to determine the influence of η and the regularizing term.

This version of the SVM problem was first discussed in Schölkopf, Smola, Williamson, and Bartlett [147] under the name of ν-*SVC*, and also used in Schölkopf, Platt, Shawe–Taylor, and Smola [146].

In order for the problem to have a solution we must pick K_m and K_s so that

$$K_m \leq \min\{2pK_s, 2qK_s\}.$$

It is shown in Section 54.5 that the dual program is

Dual of the Basic Soft margin ν-SVM Problem (SVM$_{s2'}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m \\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

If the primal problem has an optimal solution with $w \neq 0$, then using the fact that the duality gap is zero we can show that $\eta \geq 0$. Thus constraint $\eta \geq 0$ could be omitted. As in the previous case w is given by

$$w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j,$$

but b and η are not determined by the dual.

If we drop the constraint $\eta \geq 0$, then the inequality

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m$$

is replaced by the equation

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = K_m.$$

It convenient to define $\nu > 0$ such that

$$\nu = \frac{K_m}{(p+q)K_s},$$

so that the objective function $J(w, \epsilon, \xi, b, \eta)$ is given by

$$J(w, \epsilon, \xi, b, \eta) = \frac{1}{2} w^\top w + (p+q) K_s \left(-\nu\eta + \frac{1}{p+q} \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \right).$$

Since we obtain an equivalent problem by rescaling by a common positive factor, theoretically it is convenient to normalize K_s as

$$K_s = \frac{1}{p+q},$$

in which case $K_m = \nu$. This method is called the *ν-support vector machine*.

Under the **Standard Margin Hypothesis** for (SVM$_{s2'}$), there is some support vector u_{i_0} of type 1 and some support vector v_{j_0} of type 1, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ and $\xi_{j_0} = 0$, so we have the two active constraints

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for b and η and we get

$$\begin{aligned} b &= \frac{w^\top (u_{i_0} + v_{j_0})}{2} \\ \eta &= \frac{w^\top (u_{i_0} - v_{j_0})}{2}. \end{aligned}$$

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$I_\lambda = \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\}, \qquad I_\mu = \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}.$$

Then b and η are given by the following averaging formulae:

$$\begin{aligned} b &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2 \\ \eta &= w^\top \left(\Big(\sum_{i \in I_\lambda} u_i \Big) / |I_\lambda| - \Big(\sum_{j \in I_\mu} v_j \Big) / |I_\mu| \right) /2. \end{aligned}$$

Proposition 54.1 yields bounds on ν for the method to converge, namely

$$\max \left\{ \frac{2p_f}{p+q}, \frac{2q_f}{p+q} \right\} \leq \nu \leq \min \left\{ \frac{2p_m}{p+q}, \frac{2q_m}{p+q} \right\}.$$

In Section 54.7 we investigate conditions on ν that ensure that some point u_{i_0} and some point v_{j_0} is a support vector. Theorem 54.3 shows that for every optimal solution $(w, b, \eta, \epsilon, \xi)$ of Problem (SVM$_{s2'}$) with $w \neq 0$ and $\eta > 0$, if

$$\max\{2p_f/(p+q), 2q_f/(p+q)\} < \nu < \min\{2p/(p+q), 2q/(p+q)\},$$

then some u_{i_0} and some v_{j_0} is a support vector. Under the same conditions on ν Proposition 54.4 shows that η and b can always be determined in terms of (λ, μ) using a single support vector.

(3) **Soft margin ν-SVM Problem** (SVM$_{s3}$). This is the variation of Problem (SVM$_{s2'}$) obtained by adding the term $(1/2)b^2$ to the objective function. The result is that in minimizing the Lagrangian to find the dual function G, not just w but also b is determined. We also suppress the constraint $\eta \geq 0$ which turns out to be redundant. If $\nu > (p_f+q_f)/(p+q)$, then η is also determined. The fact that b and η are determined by the dual seems to be an advantage of Problem (SVM$_{s3}$).

The optimization problem is

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + (p+q)K_s\left(-\nu\eta + \frac{1}{p+q}\begin{pmatrix}\epsilon^\top & \xi^\top\end{pmatrix}\mathbf{1}_{p+q}\right)\\
&\text{subject to}\\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p\\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q.
\end{aligned}$$

Theoretically it is convenient to assume that $K_s = 1/(p+q)$. Otherwise, ν needs to be replaced by $(p+q)K_s\nu$ in all the formulae below.

It is shown in Section 54.13 that the dual is given by

Dual of the Soft margin ν-SVM Problem (SVM$_{s3}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top\\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.
\end{aligned}$$

Once a solution for λ and μ is obtained, we have

$$\begin{aligned} w &= -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j \\ b &= -\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j. \end{aligned}$$

Note that the constraint

$$\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0$$

occurring in the dual of Program ($\mathrm{SVM}_{s2'}$) has been traded for the equation

$$b = -\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j$$

determining b.

If $\nu > (p_f + q_f)/(p + q)$, then η is determined by expressing that the duality gap is zero. We obtain

$$\begin{aligned} ((p+q)\nu - p_f - q_f)\eta &= (p_f - q_f)b + w^\top \left(\sum_{j \in K_\mu} v_j - \sum_{i \in K_\lambda} u_i \right) \\ &\quad + \frac{1}{K_s} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \end{aligned}$$

In practice another way to compute η is to assume the Standard Margin Hypothesis for (SVM_{s3}). Under the **Standard Margin Hypothesis** for (SVM_{s3}), either some u_{i_0} is a support vector of type 1 *or* some v_{j_0} is a support vector of type 1. By the complementary slackness conditions $\epsilon_{i_0} = 0$ or $\xi_{j_0} = 0$, so we have

$$w^\top u_{i_0} - b = \eta, \quad \text{or} \quad -w^\top v_{j_0} + b = \eta,$$

and we can solve for η. As in ($\mathrm{SVM}_{s2'}$) we get more numerically stable formulae by averaging over the sets I_λ and I_μ.

Proposition 54.5 gives bounds ν, namely

$$\frac{p_f + q_f}{p + q} \leq \nu \leq \frac{p_m + q_m}{p + q}.$$

In Section 54.11 we investigate conditions on ν that ensure that either there is some blue support vector u_{i_0} *or* there is some red support vector v_{j_0}. Theorem 54.7 shows that for every optimal solution $(w, b, \eta, \epsilon, \xi)$ of Problem (SVM_{s3}) with $w \neq 0$ and $\eta > 0$, if

$$(p_{sf} + q_{sf})/(p + q) < \nu < 1,$$

then some u_{i_0} or some v_{j_0} is a support vector.

(4) **Basic Quadratic Soft margin ν-SVM Problem** (SVM_{s4}). This is the version of Problem ($\mathrm{SVM}_{s2'}$) in which instead of using the linear function $K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q}$ as a regularizing function we use the quadratic function $K(\|\epsilon\|_2^2 + \|\xi\|_2^2)$. The optimization problem is

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} w^\top w + (p+q) K_s \left(-\nu\eta + \frac{1}{p+q} (\epsilon^\top \epsilon + \xi^\top \xi) \right) \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\ &\qquad - w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q \\ &\qquad \eta \geq 0, \end{aligned}$$

where ν and K_s are two given positive constants. As we saw earlier, theoretically, it is convenient to pick $K_s = 1/(p+q)$. When writing a computer program, it is preferable to assume that K_s is arbitrary. In this case ν needs to be replaced by $(p+q)K_s\nu$ in all the formulae obtained with $K_s = 1/(p+q)$.

In this method, it is no longer necessary to require $\epsilon \geq 0$ and $\xi \geq 0$, because an optimal solution satisfies these conditions.

One of the advantages of this methods is that ϵ is determined by λ, ξ is determined by μ, and η and b are determined by λ and μ. We can omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$; see Section 54.14. For K_s and ν fixed, if Program (SVM_{s4}) has an optimal solution, then it is unique; see Theorem 54.8.

A drawback of Program (SVM_{s4}) is that for fixed K_s, the quantity $\delta = \eta / \|w\|$ and the hyperplanes $H_{w,b}$, $H_{w,b+\eta}$ and $H_{w,b-\eta}$ are *independent* of ν. This is shown in Theorem 54.8. Thus this method is less flexible than ($\mathrm{SVM}_{s2'}$) and (SVM_{s3}).

It is shown in Section 54.9 that the dual is given by

Dual of the Basic Quadratic Soft margin ν-SVM Problem (SVM$_{s4}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0\\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq \nu\\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p\\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q.
\end{aligned}$$

The above program is similar to the program that was obtained for Problem (SVM$_{s2'}$) but the matrix $X^\top X$ is replaced by the matrix $X^\top X + (1/2K)I_{p+q}$, which is positive definite since $K > 0$, and also the inequalities $\lambda_i \leq K$ and $\mu_j \leq K$ no longer hold. If the constraint $\eta \geq 0$ is dropped, then the inequality

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq \nu$$

is replaced by the equation

$$\sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu.$$

We obtain w from λ and μ, and γ, as in Problem (SVM$_{s2'}$); namely,

$$w = -X\begin{pmatrix}\lambda \\ \mu\end{pmatrix} = \sum_{i=1}^p \lambda_i u_i - \sum_{j=1}^q \mu_j v_j$$

and η is given by

$$(p+q)K_s\nu\eta = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \frac{1}{2K_s}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}.$$

The constraints imply that there is some i_o such that $\lambda_{i_0} > 0$ *and* some j_0 such that $\mu_{j_0} > 0$, which means that at least two points are misclassified, so Problem (SVM$_{s4}$) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable. We can solve for b using the active constraints corresponding to any i_0 such that $\lambda_{i_0} > 0$ and any j_0 such that $\mu_{j_0} > 0$. To improve numerical stability we average over the sets of indices I_λ and I_μ.

(5) **Quadratic Soft margin ν-SVM Problem** (SVM_{s5}). This is the variant of Problem (SVM_{s4}) in which we add the term $(1/2)b^2$ to the objective function. We also drop the constraint $\eta \geq 0$ which is redundant. We have the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + (p+q)K_s\left(-\nu\eta + \frac{1}{p+q}(\epsilon^\top\epsilon + \xi^\top\xi)\right) \\ &\text{subject to} \\ &\qquad w^\top u_i - b \geq \eta - \epsilon_i, \qquad i = 1, \ldots, p \\ &\qquad -w^\top v_j + b \geq \eta - \xi_j, \qquad j = 1, \ldots, q, \end{aligned}$$

where ν and K_s are two given positive constants. As we saw earlier, it is convenient to pick $K_s = 1/(p+q)$. When writing a computer program, it is preferable to assume that K_s is arbitrary. In this case ν must be replaced by $(p+q)K_s\nu$ in all the formulae.

One of the advantages of this methods is that ϵ is determined by λ, ξ is determined by μ (as in (SVM_{s4})), and both η and b determined by λ and μ. We can omit the constraint $\eta \geq 0$, because for an optimal solution it can be shown using duality that $\eta \geq 0$. For K_s and ν fixed, if Program (SVM_{s5}) has an optimal solution, then it is unique; see Theorem 54.9.

A drawback of Program (SVM_{s5}) is that for fixed K_s, the quantity $\delta = \eta/\left\|w\right\|$ and the hyperplanes $H_{w,b}, H_{w,b+\eta}$ and $H_{w,b-\eta}$ are *independent* of ν. This is shown in Theorem 54.9. Thus this method is less flexible than ($\mathrm{SVM}_{s2'}$) and (SVM_{s3}).

It is shown in Section 54.15 that the dual of Program (SVM_{s5}) is given by

Dual of the Quadratic Soft margin ν-SVM Problem (SVM_{s5}):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(X^\top X + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{1}{2K}I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\ &\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\ &\qquad \mu_j \geq 0, \quad j = 1, \ldots, q. \end{aligned}$$

This time we obtain w, b, η, ϵ and ξ from λ and μ:

$$\begin{aligned}
w &= \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j \\
b &= -\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \\
\epsilon &= \frac{\lambda}{2K} \\
\xi &= \frac{\mu}{2K},
\end{aligned}$$

and

$$(p+q)K_s \nu \eta = \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \left(X^\top X + \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p^\top & -\mathbf{1}_p \mathbf{1}_q^\top \\ -\mathbf{1}_q \mathbf{1}_p^\top & \mathbf{1}_q \mathbf{1}_q^\top \end{pmatrix} + \frac{1}{2K_s} I_{p+q} \right) \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

The constraint

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j = \nu$$

implies that either there is some i_0 such that $\lambda_{i_0} > 0$ or there is some j_0 such that $\mu_{j_0} > 0$, we have $\epsilon_{i_0} > 0$ or $\xi_{j_0} > 0$, which means that at least one point is misclassified, so Problem (SVM$_{s5}$) should only be used when the sets $\{u_i\}$ and $\{v_j\}$ are *not* linearly separable.

These methods all have a kernelized version.

We implemented all these methods in `Matlab`, solving the dual using ADMM.

From a theoretical point of view, Problems (SVM$_{s4}$) and (SVM$_{s5}$) seem to have more advantages than the others since they determine w, b, η and b without requiring any condition about support vectors of type 1. However, from a practical point of view, Problems (SVM$_{s4}$) and (SVM$_{s5}$) are less flexible that (SVM$_{s2'}$) and (SVM$_{s3}$), and we have observed that (SVM$_{s4}$) and (SVM$_{s5}$) are unable to produce as small a margin δ as (SVM$_{s2'}$) and (SVM$_{s3}$).

54.18 Problems

Problem 54.1. Prove the following inequality

$$\max\left\{\frac{1}{2p_m}, \frac{1}{2q_m}\right\} \leq K \leq \min\left\{\frac{1}{2p_f}, \frac{1}{2q_f}\right\}$$

stated just after Definition 54.1.

Problem 54.2. Prove the averaging formulae

$$b = w^\top \left(\Big(\sum_{i \in I_\lambda} u_i\Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j\Big) / |I_\mu| \right) /2$$

$$\delta = w^\top \left(\Big(\sum_{i \in I_\lambda} u_i\Big) / |I_\lambda| - \Big(\sum_{j \in I_\mu} v_j\Big) / |I_\mu| \right) /2$$

stated at the end of Section 54.1.

Problem 54.3. Prove that the matrix

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}$$

has rank $p + q + 2$.

Problem 54.4. Prove that the dual program of the kernel version of (SVM$_{s1}$) is given by:

Dual of Soft margin kernel SVM (SVM$_{s1}$):

$$\begin{aligned} &\text{minimize} \quad \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} \mathbf{K} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{q} \mu_j = \frac{1}{2} \\ &\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \dots, p \\ &\qquad 0 \leq \mu_j \leq K, \quad j = 1, \dots, q, \end{aligned}$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given by

$$\mathbf{K}_{ij} = \begin{cases} \kappa(u_i, u_j) & 1 \leq i \leq p,\ 1 \leq j \leq q \\ -\kappa(u_i, v_{j-p}) & 1 \leq i \leq p,\ p+1 \leq j \leq p+q \\ -\kappa(v_{i-p}, u_j) & p+1 \leq i \leq p+q,\ 1 \leq j \leq p \\ \kappa(v_{i-p}, v_{j-q}) & p+1 \leq i \leq p+q,\ p+1 \leq j \leq p+q. \end{cases}$$

Problem 54.5. Prove the averaging formula

$$b = w^\top \left(\Big(\sum_{i \in I_\lambda} u_i\Big) / |I_\lambda| + \Big(\sum_{j \in I_\mu} v_j\Big) / |I_\mu| \right) /2$$

stated in Section 54.3.

Problem 54.6. Prove that the kernel version of Program (SVM$_{s2}$) is given by:

Dual of Soft margin kernel SVM (SVM$_{s2}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{K}\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad 0 \leq \lambda_i \leq K, \quad i = 1, \ldots, p \\
&\qquad 0 \leq \mu_j \leq K, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the $\ell \times \ell$ kernel symmetric matrix (with $\ell = p + q$) given at the end of Section 54.1.

Problem 54.7. Prove that the matrix

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}$$

has rank $p + q + 1$.

Problem 54.8. Prove that the matrices

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top & 0 \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top & -1 \\ I_p & 0_{p,q} & I_p & 0_{p,q} & 0_p \\ 0_{q,p} & I_q & 0_{q,p} & I_q & 0_q \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}$$

have rank $p + q + 2$.

Problem 54.9. Prove that the kernel version of Program (SVM$_{s2'}$) is given by:

Dual of the Soft margin kernel SVM ($\mathrm{SVM}_{s2'}$):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\mathbf{K}\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0\\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq K_m\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

Problem 54.10. Prove the formulae determining b in terms of η stated just before Theorem 54.8.

Problem 54.11. Prove that the matrix

$$A = \begin{pmatrix} \mathbf{1}_p^\top & \mathbf{1}_q^\top & 0_p^\top & 0_q^\top \\ I_p & 0_{p,q} & I_p & 0_{p,q} \\ 0_{q,p} & I_q & 0_{q,p} & I_q \end{pmatrix}$$

has rank $p + q + 1$.

Problem 54.12. Prove that the kernel version of Program (SVM_{s3}) is given by:

Dual of the Soft margin kernel SVM (SVM_{s3}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix}\right)\begin{pmatrix}\lambda\\ \mu\end{pmatrix}\\
&\text{subject to}\\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu\\
&\qquad 0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p\\
&\qquad 0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

Problem 54.13. Prove that the matrices

$$A = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top & 0 \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top & -1 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} \mathbf{1}_p^\top & -\mathbf{1}_q^\top \\ \mathbf{1}_p^\top & \mathbf{1}_q^\top \end{pmatrix}$$

have rank 2.

Problem 54.14. Implement Program (SVM_{s4}) in `Matlab`. You may adapt the programs given in Section B.2 and Section B.3.

Problem 54.15. Prove that the kernel version of Program (SVM_{s4}) is given by:

Dual of the Soft margin kernel SVM (SVM_{s4}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \frac{p+q}{2} I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i - \sum_{j=1}^q \mu_j = 0 \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j \geq \nu \\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

Problem 54.16. Implement Program (SVM_{s5}) in `Matlab`. You may adapt the programs given in Section B.2 and Section B.3.

Problem 54.17. Prove that the kernel version of Program (SVM_{s5}) is given by:

Dual of the Soft margin kernel SVM (SVM_{s5}):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(\mathbf{K} + \begin{pmatrix}\mathbf{1}_p\mathbf{1}_p^\top & -\mathbf{1}_p\mathbf{1}_q^\top \\ -\mathbf{1}_q\mathbf{1}_p^\top & \mathbf{1}_q\mathbf{1}_q^\top\end{pmatrix} + \frac{p+q}{2} I_{p+q}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^p \lambda_i + \sum_{j=1}^q \mu_j = \nu \\
&\qquad \lambda_i \geq 0, \quad i = 1, \ldots, p \\
&\qquad \mu_j \geq 0, \quad j = 1, \ldots, q,
\end{aligned}$$

where $\mathbf{K}$ is the kernel matrix of Section 54.1.

Chapter 55

Ridge Regression, Lasso, Elastic Net

In this chapter we discuss linear regression. This problem can be cast as a learning problem. We observe a sequence of (distinct) pairs $((x_1, y_1), \ldots, (x_m, y_m))$ called a *set of training data* (or *predictors*), where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, viewed as input-output pairs of some unknown function f that we are trying to infer. The simplest kind of function is a linear function $f(x) = x^\top w$, where $w \in \mathbb{R}^n$ is a vector of coefficients usually called a *weight vector*. Since the problem is overdetermined and since our observations may be subject to errors, we can't solve for w exactly as the solution of the system $Xw = y$, where X is the $m \times n$ matrix

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

where the row vectors $x_i^\top$ are the rows of X, and thus the $x_i \in \mathbb{R}^n$ are column vectors. So instead we solve the least-squares problem of minimizing $\|Xw - y\|_2^2$. In general there are still infinitely many solutions so we add a regularizing term. If we add the term $K \|w\|_2^2$ to the objective function $J(w) = \|Xw - y\|_2^2$, then we have *ridge regression*. This problem is discussed in Section 55.1 where we derive the dual program. The dual has a unique solution which yields a solution of the primal. However, the solution of the dual is given in terms of the matrix $XX^\top$ (whereas the solution of the primal is given in terms of $X^\top X$), and since our data points x_i are represented by the rows of the matrix X, we see that this solution only involves inner products of the x_i. This observation is the core of the idea of kernel functions, which were discussed in Chapter 53. We also explain how to solve the problem of learning an affine function $f(x) = x^\top w + b$.

In general the vectors w produced by ridge regression have few zero entries. In practice it is highly desirable to obtain sparse solutions, that is vectors w with many components equal to zero. This can be achieved by replacing the regularizing term $K \|w\|_2^2$ by the regularizing term $K \|w\|_1$; that is, to use the ℓ^1-norm instead of the ℓ^2-norm; see Section 55.4. This method has the exotic name of *lasso regression*. This time there is no closed-form solution, but this is a convex optimization problem and there are efficient iterative methods to solve it. One of the best methods relies on ADMM (see Section 52.8) and is discussed in Section 55.4. The lasso method has some limitations, in particular when the number m of data is smaller than the dimension n of the data. This happens in some applications in genetics and medicine. Fortunately there is a way to combine the best features of ridge regression and lasso, which is to use *two* regularizing terms:

1. An ℓ^2-term $(1/2)K \|w\|_2^2$ as in ridge regression (with $K > 0$).

2. An ℓ^1-term $\tau \|w\|_1$ as in lasso.

This method is known as *elastic net regression* and is discussed in Section 55.6. It retains most of the desirable features of ridge regression and lasso, and eliminates some of their weaknesses. Furthermore, it is effectively solved by ADMM.

55.1 Ridge Regression

The problem of solving an overdetermined or underdetermined linear system $Aw = y$, where A is an $m \times n$ matrix, arises as a "learning problem" in which we observe a sequence of data $((a_1, y_1), \ldots, (a_m, y_m))$, viewed as input-output pairs of some unknown function f that we are trying to infer, where the a_i are the *rows* of the matrix A and $y_i \in \mathbb{R}$. The values y_i are sometimes called *labels* or *responses*. The simplest kind of function is a linear function $f(x) = x^\top w$, where $w \in \mathbb{R}^n$ is a vector of coefficients usually called a *weight vector*, or sometimes an *estimator*. In the statistical literature w is often denoted by β. Since the problem is overdetermined and since our observations may be subject to errors, we can't solve for w exactly as the solution of the system $Aw = y$, so instead we solve the least-square problem of minimizing $\|Aw - y\|_2^2$.

In Section 23.1 we showed that this problem can be solved using the pseudo-inverse. We know that the minimizers w are solutions of the normal equations $A^\top Aw = A^\top y$, but when $A^\top A$ is not invertible, such a solution is not unique so some criterion has to be used to choose among these solutions.

One solution is to pick the unique vector w^+ of smallest Euclidean norm $\|w^+\|_2$ that minimizes $\|Aw - y\|_2^2$. The solution w^+ is given by $w^+ = A^+ y$, where A^+ is the pseudo-inverse of A. The matrix A^+ is obtained from an SVD of A, say $A = V\Sigma U^\top$. Namely, $A^+ = U\Sigma^+ V^\top$, where Σ^+ is the matrix obtained from Σ by replacing every nonzero singular value σ_i in Σ by σ_i^{-1}, leaving all zeros in place, and then transposing. The difficulty with this approach is that it requires knowing whether a singular value is zero or very small but nonzero. A very small nonzero singular value σ in Σ yields a very large value σ^{-1} in Σ^+, but $\sigma = 0$ remains 0 in Σ^+.

This discontinuity phenomenon is not desirable and another way is to control the size of w by adding a regularization term to $\|Aw - y\|^2$, and a natural candidate is $\|w\|^2$.

It is customary to rename each column vector $a_i^\top$ as x_i (where $x_i \in \mathbb{R}^n$) and to rename the input data matrix A as X, so that the row vector $x_i^\top$ are the *rows* of the $m \times n$ matrix X

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix}.$$

Our optimization problem, called *ridge regression*, is

Program (RR1):

$$\text{minimize} \quad \|y - Xw\|^2 + K\|w\|^2,$$

which by introducing the new variable $\xi = y - Xw$ can be rewritten as

Program (RR2):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad\qquad y - Xw = \xi, \end{aligned}$$

where $K > 0$ is some constant determining the influence of the regularizing term $w^\top w$, and we minimize over ξ and w.

The objective function of the first version of our minimization problem can be expressed as

$$\begin{aligned} J(w) &= \|y - Xw\|^2 + K\|w\|^2 \\ &= (y - Xw)^\top (y - Xw) + K w^\top w \\ &= y^\top y - 2w^\top X^\top y + w^\top X^\top X w + K w^\top w \\ &= w^\top (X^\top X + K I_n) w - 2w^\top X^\top y + y^\top y. \end{aligned}$$

The matrix $X^\top X$ is symmetric positive semidefinite and $K > 0$, so the matrix $X^\top X + KI_n$ is positive definite. It follows that

$$J(w) = w^\top (X^\top X + K I_n) w - 2w^\top X^\top y + y^\top y$$

is strictly convex, so by Theorem 40.13(2)-(4), it has a unique minimum iff $\nabla J_w = 0$. Since

$$\nabla J_w = 2(X^\top X + K I_n) w - 2X^\top y,$$

we deduce that

$$w = (X^\top X + K I_n)^{-1} X^\top y. \tag{$*_{wp}$}$$

There is an interesting connection between the matrix $(X^\top X + KI_n)^{-1}X^\top$ and the pseudo-inverse X^+ of X.

Proposition 55.1. *The limit of the matrix* $(X^\top X + KI_n)^{-1}X^\top$ *when* $K > 0$ *goes to zero is the pseudo-inverse* X^+ *of* X.

Proof. To show this let $X = V\Sigma U^\top$ be a SVD of X. Then

$$(X^\top X + KI_n) = U\Sigma^\top V^\top V\Sigma U^\top + KI_n = U(\Sigma^\top\Sigma + KI_n)U^\top,$$

so

$$(X^\top X + KI_n)^{-1}X^\top = U(\Sigma^\top\Sigma + KI_n)^{-1}U^\top U\Sigma^\top V^\top = U(\Sigma^\top\Sigma + KI_n)^{-1}\Sigma^\top V^\top.$$

The diagonal entries in the matrix $(\Sigma^\top\Sigma + KI_n)^{-1}\Sigma^\top$ are

$$\frac{\sigma_i}{\sigma_i^2 + K}, \quad \text{if } \sigma_i > 0,$$

and zero if $\sigma_i = 0$. All nondiagonal entries are zero. When $\sigma_i > 0$ and $K > 0$ goes to 0,

$$\lim_{K\mapsto 0} \frac{\sigma_i}{\sigma_i^2 + K} = \sigma_i^{-1},$$

so

$$\lim_{K\mapsto 0}(\Sigma^\top\Sigma + KI_n)^{-1}\Sigma^\top = \Sigma^+,$$

which implies that

$$\lim_{K\mapsto 0}(X^\top X + KI_n)^{-1}X^\top = X^+. \qquad \square$$

The dual function of the first formulation of our problem is a constant function (with value the minimum of J) so it is not useful, but the second formulation of our problem yields an interesting dual problem. The Lagrangian is

$$\begin{aligned} L(\xi, w, \lambda) &= \xi^\top\xi + Kw^\top w + (y - Xw - \xi)^\top\lambda \\ &= \xi^\top\xi + Kw^\top w - w^\top X^\top\lambda - \xi^\top\lambda + \lambda^\top y, \end{aligned}$$

with $\lambda, \xi, y \in \mathbb{R}^m$. The Lagrangian $L(\xi, w, \lambda)$, as a function of ξ and w with λ held fixed, is obviously convex, in fact strictly convex.

To derive the dual function $G(\lambda)$ we minimize $L(\xi, w, \lambda)$ with respect to ξ and w. Since $L(\xi, w, \lambda)$ is (strictly) convex as a function of ξ and w, by Theorem 40.13(4), it has a minimum iff its gradient $\nabla L_{\xi,w}$ is zero (in fact, by Theorem 40.13(2), a unique minimum since the function is strictly convex). Since

$$\nabla L_{\xi,w} = \begin{pmatrix} 2\xi - \lambda \\ 2Kw - X^\top\lambda \end{pmatrix},$$

we get

$$\begin{aligned} \lambda &= 2\xi \\ w &= \frac{1}{2K}X^\top\lambda = X^\top\frac{\xi}{K}. \end{aligned}$$

The above suggests defining the variable α so that $\xi = K\alpha$, so we have $\lambda = 2K\alpha$ and $w = X^\top \alpha$. Then we obtain the dual function as a function of α by substituting the above values of ξ, λ and w back in the Lagrangian and we get

$$\begin{aligned} G(\alpha) &= K^2\alpha^\top\alpha + K\alpha^\top XX^\top\alpha - 2K\alpha^\top XX^\top\alpha - 2K^2\alpha^\top\alpha + 2K\alpha^\top y \\ &= -K\alpha^\top(XX^\top + KI_m)\alpha + 2K\alpha^\top y. \end{aligned}$$

This is a strictly concave function so by Theorem 40.13(4), its maximum is achieved iff $\nabla G_\alpha = 0$, that is,

$$2K(XX^\top + KI_m)\alpha = 2Ky,$$

which yields

$$\alpha = (XX^\top + KI_m)^{-1}y.$$

Putting everything together we obtain

$$\begin{aligned} \alpha &= (XX^\top + KI_m)^{-1}y \\ w &= X^\top\alpha \\ \xi &= K\alpha, \end{aligned}$$

which yields

$$w = X^\top(XX^\top + KI_m)^{-1}y. \tag{$*_{wd}$}$$

Earlier in $(*_{wp})$ we found that

$$w = (X^\top X + KI_n)^{-1}X^\top y,$$

and it is easy to check that

$$(X^\top X + KI_n)^{-1}X^\top = X^\top(XX^\top + KI_m)^{-1}.$$

If $n < m$ it is cheaper to use the formula on the left-hand side, but if $m < n$ it is cheaper to use the formula on the right-hand side.

55.2 Ridge Regression; Learning an Affine Function

It is easy to adapt the above method to learn an affine function $f(x) = x^\top w + b$ instead of a linear function $f(x) = x^\top w$, where $b \in \mathbb{R}$. We have the following optimization program

Program (RR3):

$$\begin{aligned} &\text{minimize} \quad \xi^\top\xi + Kw^\top w \\ &\text{subject to} \\ &\qquad\qquad y - Xw - b\mathbf{1} = \xi, \end{aligned}$$

with $y, \xi, \mathbf{1} \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Note that in Program (**RR3**) minimization is performed over ξ, w and b, but b is not penalized in the objective function. As in Section 55.1, the objective function is strictly convex.

The Lagrangian associated with this program is

$$L(\xi, w, b, \lambda) = \xi^\top \xi + K w^\top w - w^\top X^\top \lambda - \xi^\top \lambda - b\mathbf{1}^\top \lambda + \lambda^\top y.$$

Since L is (strictly) convex as a function of ξ, b, w, by Theorem 40.13(4), it has a minimum iff $\nabla L_{\xi,b,w} = 0$. We get

$$\begin{aligned} \lambda &= 2\xi \\ \mathbf{1}^\top \lambda &= 0 \\ w &= \frac{1}{2K} X^\top \lambda = X^\top \frac{\xi}{K}. \end{aligned}$$

As before, if we set $\xi = K\alpha$ we obtain $\lambda = 2K\alpha$, $w = X^\top \alpha$, and

$$G(\alpha) = -K\alpha^\top (XX^\top + KI_m)\alpha + 2K\alpha^\top y.$$

Since $K > 0$ and $\lambda = 2K\alpha$, the dual to ridge regression is the following program

Program (DRR3):

$$\begin{aligned} &\text{minimize} \quad \alpha^\top (XX^\top + KI_m)\alpha - 2\alpha^\top y \\ &\text{subject to} \\ &\qquad\qquad \mathbf{1}^\top \alpha = 0, \end{aligned}$$

where the minimization is over α.

Observe that up to the factor 1/2, this problem satisfies the conditions of Proposition 42.3 with

$$\begin{aligned} A &= (XX^\top + KI_m)^{-1} \\ b &= y \\ B &= \mathbf{1}_m \\ f &= 0, \end{aligned}$$

and x renamed as α. Therefore, it has a unique solution (α, μ) (beware that $\lambda = 2K\alpha$ is **not** the λ used in Proposition 42.3, which we rename as μ), which is the unique solution of the KKT-equations

$$\begin{pmatrix} XX^\top + KI_m & \mathbf{1}_m \\ \mathbf{1}_m^\top & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \mu \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

Since the solution given by Proposition 42.3 is

$$\mu = (B^\top A B)^{-1}(B^\top A b - f), \quad \alpha = A(b - B\mu),$$

we get

$$\mu = (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}^\top (XX^\top + KI_m)^{-1} y, \quad \alpha = (XX^\top + KI_m)^{-1}(y - \mu\mathbf{1}).$$

Note that the matrix $B^\top AB$ is the scalar $\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1}$, which is the negative of the Schur complement of $XX^\top + KI_m$.

Interestingly $b = \mu$, which is not obvious a priori.

Proposition 55.2. *We have $b = \mu$.*

Proof. To prove this result we need to express α differently. Since μ is a scalar, $\mu\mathbf{1} = \mathbf{1}\mu$, so

$$\mu\mathbf{1} = \mathbf{1}\mu = (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\top (XX^\top + KI_m)^{-1} y,$$

and we obtain

$$\alpha = (XX^\top + KI_m)^{-1}(I_m - (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\top (XX^\top + KI_m)^{-1}) y. \tag{$*_{\alpha_3}$}$$

Since $w = X^\top \alpha$, we have

$$w = X^\top (XX^\top + KI_m)^{-1}(I_m - (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\top (XX^\top + KI_m)^{-1}) y. \tag{$*_{w_3}$}$$

From $\xi = K\alpha$, we deduce that b is given by the equation

$$b\mathbf{1} = y - Xw - K\alpha.$$

Since $w = X^\top \alpha$, using $(*_{\alpha_3})$ we obtain

$$\begin{aligned} b\mathbf{1} &= y - Xw - K\alpha \\ &= y - (XX^\top + KI_m)\alpha \\ &= y - (I_m - (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\top (XX^\top + KI_m)^{-1}) y \\ &= (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^\top (XX^\top + KI_m)^{-1}) y \\ &= \mu\mathbf{1}, \end{aligned}$$

and thus

$$b = \mu = (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}^\top (XX^\top + KI_m)^{-1} y, \tag{$*_{b_3}$}$$

as claimed. ☐

In summary the KKT-equations determine both α and μ, and so $w = X^\top \alpha$ and b as well.

There is also a useful expression of b as an average.

Since $\mathbf{1}^\top \mathbf{1} = m$ and $\mathbf{1}^\top \alpha = 0$, we get

$$b = \frac{1}{m}\mathbf{1}^\top y - \frac{1}{m}\mathbf{1}^\top Xw - \frac{1}{m}K\mathbf{1}^\top \alpha = \overline{y} - \sum_{j=1}^{n} \overline{X^j} w_j,$$

where $\overline{y}$ is the mean of y and $\overline{X^j}$ is the mean of the jth column of X. Therefore,

$$b = \overline{y} - \sum_{j=1}^{n} \overline{X^j} w_j = \overline{y} - (\overline{X^1} \cdots \overline{X^n}) w,$$

where $(\overline{X^1} \cdots \overline{X^n})$ is the $1 \times n$ row vector whose jth entry is $\overline{X^j}$.

We will now show that solving the dual (**DRR3**) for α and obtaining $w = X^\top \alpha$ is equivalent to solving our previous ridge regression Problem (**RR2**) applied to the centered data $\widehat{y} = y - \overline{y}\mathbf{1}_m$ and $\widehat{X} = X - \overline{X}$, where $\overline{X}$ is the $m \times n$ matrix whose jth column is $\overline{X^j}\mathbf{1}_m$, the vector whose coordinates are all equal to the mean $\overline{X^j}$ of the jth column X^j of X.

The expression

$$b = \overline{y} - (\overline{X^1} \cdots \overline{X^n}) w$$

suggests looking for an intercept term b (also called bias) of the above form, namely

Program (RR4):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad y - Xw - b\mathbf{1} = \xi \\ &\qquad b = \widehat{b} + \overline{y} - (\overline{X^1} \cdots \overline{X^n}) w, \end{aligned}$$

with $\widehat{b} \in \mathbb{R}$. Again, in Program (**RR4**), minimization is performed over ξ, w, b and $\widehat{b}$, but b and $\widehat{b}$ are not penalized.

Since

$$b\mathbf{1} = \widehat{b}\mathbf{1} + \overline{y}\mathbf{1} - (\overline{X^1}\mathbf{1} \cdots \overline{X^n}\mathbf{1}) w,$$

if $\overline{X} = (\overline{X^1}\mathbf{1} \cdots \overline{X^n}\mathbf{1})$ is the $m \times n$ matrix whose jth column is the vector $\overline{X^j}\mathbf{1}$, then the above program is equivalent to the program

Program (RR5):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad y - Xw - \overline{y}\mathbf{1} + \overline{X} w - \widehat{b}\mathbf{1} = \xi, \end{aligned}$$

where minimization is performed over ξ, w and $\widehat{b}$. If we write $\widehat{y} = y - \overline{y}\mathbf{1}$ and $\widehat{X} = X - \overline{X}$, then the above program becomes

Program (RR6):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad \widehat{y} - \widehat{X} w - \widehat{b}\mathbf{1} = \xi, \end{aligned}$$

minimizing over ξ, w and $\widehat{b}$. If the solution to this program is $\widehat{w}$, then $\widehat{b}$ is given by

$$\widehat{b} = \overline{\widehat{y}} - (\overline{\widehat{X}^1} \cdots \overline{\widehat{X}^n})\widehat{w} = 0,$$

since the data $\widehat{y}$ and $\widehat{X}$ are centered. Therefore (**RR6**) *is equivalent to ridge regression without an intercept term applied to the centered data* $\widehat{y} = y - \overline{y}\mathbf{1}$ *and* $\widehat{X} = X - \overline{X}$,

Program (RR6′):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad\qquad \widehat{y} - \widehat{X} w = \xi, \end{aligned}$$

minimizing over ξ and w.

If $\widehat{w}$ is the optimal solution of this program given by

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + K I_m)^{-1} \widehat{y}, \tag{$*_{w6}$}$$

then b is given by

$$b = \overline{y} - (\overline{X^1} \cdots \overline{X^n})\widehat{w}.$$

Remark: Although this is not obvious a priori, the optimal solution w^* of the Program (**RR3**) given by $(*_{w_3})$ is equal to the optimal solution $\widehat{w}$ of Program (**RR6′**) given by $(*_{w_6})$. We believe that it should be possible to prove the equivalence of these formulae but a proof eludes us at this time. We leave this as an open problem. In practice the Program (**RR6′**) involving the centered data appears to be the preferred one.

Example 55.1. Consider the data set (X, y_1) with

$$X = \begin{pmatrix} -10 & 11 \\ -6 & 5 \\ -2 & 4 \\ 0 & 0 \\ 1 & 2 \\ 2 & -5 \\ 6 & -4 \\ 10 & -6 \end{pmatrix}, \quad y_1 = \begin{pmatrix} 0 \\ -2.5 \\ 0.5 \\ -2 \\ 2.5 \\ -4.2 \\ 1 \\ 4 \end{pmatrix}$$

as illustrated in Figure 55.1. We find that $\overline{y} = -0.0875$ and $(\overline{X^1}, \overline{X^2}) = (0.125, 0.875)$. For the value $K = 5$, we obtain

$$w = \begin{pmatrix} 0.9207 \\ 0.8677 \end{pmatrix}, \quad b = -0.9618,$$

for $K = 0.1$, we obtain

$$w = \begin{pmatrix} 1.1651 \\ 1.1341 \end{pmatrix}, \quad b = -1.2255,$$

and for $K = 0.01$,

$$w = \begin{pmatrix} 1.1709 \\ 1.1405 \end{pmatrix}, \quad b = -1.2318.$$

See Figure 55.2.

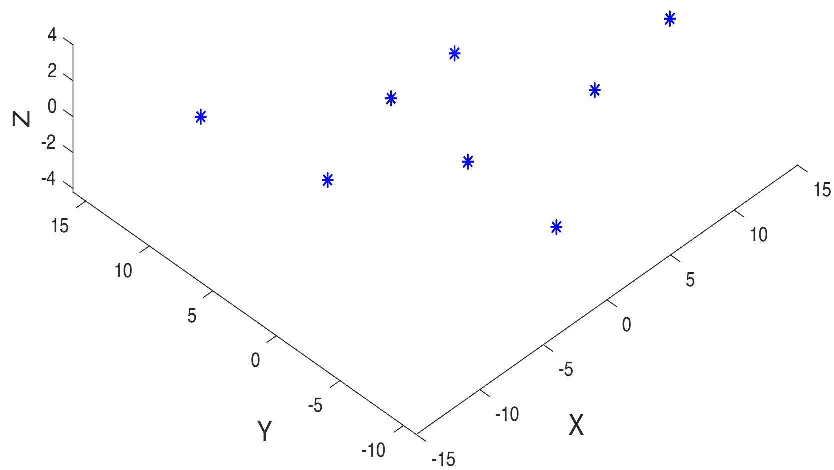


Figure 55.1: The data set (X, y_1) of Example 55.1.

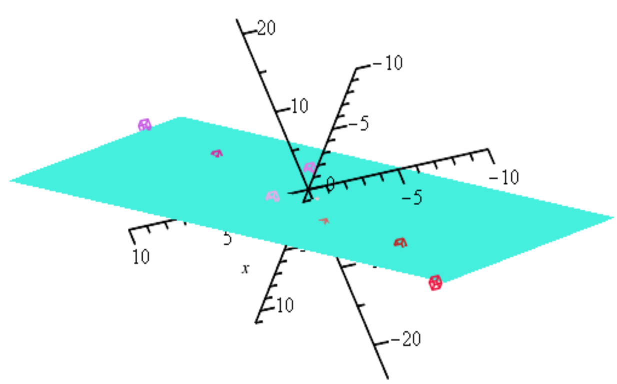


Figure 55.2: The graph of the plane $f(x, y) = 1.1709x + 1.1405y - 1.2318$ as an approximate fit to the data (X, y_1) of Example 55.1.

We conclude that the points (X_i, y_i) (where X_i is the ith row of X) almost lie on the plane of equation

$$x + y - z - 1 = 0,$$

and that f is almost the function given by $f(x, y) = 1.1x + 1.1y - 1.2$. See Figures 55.3 and 55.4.

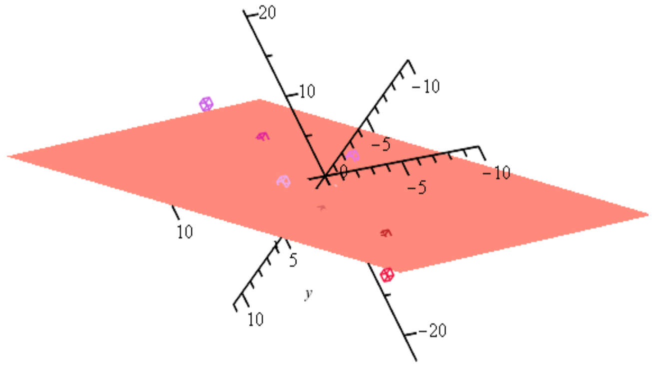


Figure 55.3: The graph of the plane $f(x, y) = 1.1x + 1.1y - 1.2$ as an approximate fit to the data (X, y_1) of Example 55.1.

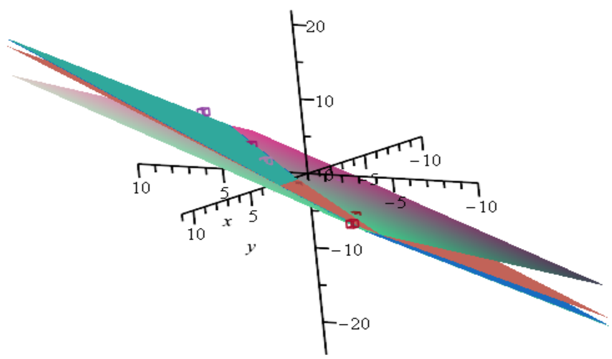


Figure 55.4: A comparison of how the graphs of the planes corresponding to $K = 1, 0.1, 0.01$ and the salmon plane of equation $f(x, y) = 1.1x + 1.1y - 1.2$ approximate the data (X, y_1) of Example 55.1.

If we change y_1 to

$$y_2 = \begin{pmatrix} 0 & -2 & 1 & -1 & 2 & -4 & 1 & 3 \end{pmatrix}^\top,$$

as evidenced by Figure 55.5, the exact solution is

$$w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = -1,$$

and for $K = 0.01$, we find that

$$w = \begin{pmatrix} 0.9999 \\ 0.9999 \end{pmatrix}, \quad b = -0.9999.$$

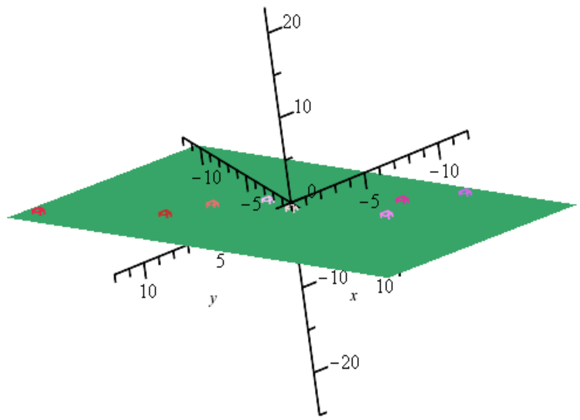


Figure 55.5: The data (X, y_2) of Example 55.1 is contained within the graph of the plane $f(x, y) = x + y - 1$.

We can see how the choice of K affects the quality of the solution (w, b) by computing the norm $\|\xi\|_2$ of the error vector $\xi = \widehat{y} - \widehat{X}w$. We notice that the smaller K is, the smaller is this norm.

It is natural to wonder what happens *if we also penalize* b in program (**RR3**). Let us add the term Kb^2 to the objective function. Then we obtain the program

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w + K b^2 \\ &\text{subject to} \\ &\qquad\qquad y - Xw - b\mathbf{1} = \xi, \end{aligned}$$

minimizing over ξ, w and b.

This suggests treating b an an extra component of the weight vector w and by forming the $m \times (n + 1)$ matrix $[X\ \mathbf{1}]$ obtained by adding a column of 1's (of dimension m) to the matrix X, we obtain

Program (RR3b):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w + K b^2 \\ &\text{subject to} \\ &\qquad y - [X\ \mathbf{1}] \begin{pmatrix} w \\ b \end{pmatrix} = \xi, \end{aligned}$$

minimizing over ξ, w and b.

This program is solved just as Program (**RR2**). In terms of the dual variable α, we get

$$\begin{aligned} \alpha &= ([X\ \mathbf{1}][X\ \mathbf{1}]^\top + K I_m)^{-1} y \\ \begin{pmatrix} w \\ b \end{pmatrix} &= [X\ \mathbf{1}]^\top \alpha \\ \xi &= K\alpha. \end{aligned}$$

Thus $b = \mathbf{1}^\top \alpha$. Observe that $[X\ \mathbf{1}][X\ \mathbf{1}]^\top = XX^\top + \mathbf{1}\mathbf{1}^\top$.

If $n < m$, it is preferable to use the formula

$$\begin{pmatrix} w \\ b \end{pmatrix} = ([X\ \mathbf{1}]^\top [X\ \mathbf{1}] + K I_{n+1})^{-1} [X\ \mathbf{1}]^\top y.$$

Since we also have the equation

$$y - Xw - b\mathbf{1} = \xi,$$

we obtain

$$\frac{1}{m}\mathbf{1}^\top y - \frac{1}{m}\mathbf{1}^\top X w - \frac{1}{m} b \mathbf{1}^\top \mathbf{1} = \frac{1}{m}\mathbf{1}^\top K\alpha,$$

so

$$\overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n}) w - b = \frac{1}{m} K b,$$

which yields

$$b = \frac{m}{m+K}(\overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n}) w).$$

Remark: As a least squares problem, the solution is given in terms of the pseudo-inverse $[X\ \mathbf{1}]^+$ of $[X\ \mathbf{1}]$ by

$$\begin{pmatrix} w \\ b \end{pmatrix} = [X\ \mathbf{1}]^+ y.$$

Example 55.2. Applying Program (**RR3b**) to the data set of Example 55.1 with $K = 0.01$ yields

$$w = \begin{pmatrix} 1.1706 \\ 1.1401 \end{pmatrix}, \quad b = -1.2298.$$

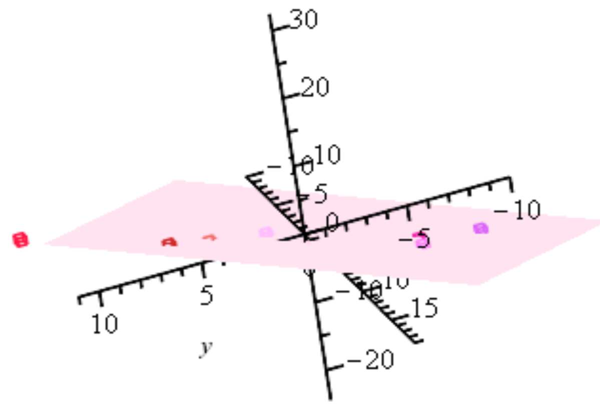


Figure 55.6: The graph of the plane $f(x, y) = 1.1706x + 1.1401y - 1.2298$ as an approximate fit to the data (X, y_1) of Example 55.1.

See Figure 55.6. We can see how the choice of K affects the quality of the solution (w, b) by computing the norm $\|\xi\|_2$ of the error vector $\xi = y - Xw - b\mathbf{1}_m$. As in Example 55.1 we notice that the smaller K is, the smaller is this norm. We also observe that for a given value of K, Program (**RR6′**) gives a slightly smaller value of $\|\xi\|_2$ than (**RR3**b) does.

As pointed out by Hastie, Tibshirani, and Friedman [88] (Section 3.4), a defect of the approach where b is also penalized is that the solution for b is not invariant under adding a constant c to each value y_i. This is not the case for the approach using Program (**RR6′**).

55.3 Kernel Ridge Regression

One interesting aspect of the dual (of either (**RR2**) or (**RR3**)) is that it shows that the solution w being of the form $X^\top \alpha$, is a linear combination

$$w = \sum_{i=1}^m \alpha_i x_i$$

of the data points x_i, with the coefficients α_i corresponding to the dual variable $\lambda = 2K\alpha$ of the dual function, and with

$$\alpha = (XX^\top + KI_m)^{-1} y.$$

If m is smaller than n, then it is more advantageous to solve for α. But what really makes the dual interesting is that with our definition of X as

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

the matrix $XX^\top$ consists of the inner products $x_i^\top x_j$, and similarly the function learned $f(x) = x^\top w$ can be expressed as

$$f(x) = \sum_{i=1}^{m} \alpha_i x_i^\top x,$$

namely that both w and $f(x)$ are given *in terms of the inner products* $x_i^\top x_j$ and $x_i^\top x$.

This fact is the key to a generalization to ridge regression in which the input space $\mathbb{R}^n$ is embedded in a larger (possibly infinite dimensional) Euclidean space F (with an inner product $\langle -, - \rangle$) usually called a *feature space*, using a function

$$\varphi \colon \mathbb{R}^n \to F.$$

The problem becomes (*kernel ridge regression*)

Program (KRR2):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K\langle w, w\rangle \\ &\text{subject to} \\ &\qquad\qquad y_i - \langle w, \varphi(x_i)\rangle = \xi_i, \quad i = 1, \ldots, m, \end{aligned}$$

minimizing over ξ and w. Note that $w \in F$. This problem is discussed in Shawe–Taylor and Christianini [159] (Section 7.3).

We will show below that the solution is exactly the same:

$$\begin{aligned} \alpha &= (\mathbf{G} + KI_m)^{-1} y \\ w &= \sum_{i=1}^{m} \alpha_i \varphi(x_i) \\ \xi &= K\alpha, \end{aligned}$$

where $\mathbf{G}$ is the Gram matrix given by $\mathbf{G}_{ij} = \langle \varphi(x_i), \varphi(x_j)\rangle$. This matrix is also called the *kernel matrix* and is often denoted by $\mathbf{K}$ instead of $\mathbf{G}$.

In this framework we have to be a little careful in using gradients since the inner product $\langle -, - \rangle$ on F is involved and F could be infinite dimensional, but this causes no problem because we can use derivatives, and by Proposition 39.5 we have

$$d\langle -, - \rangle_{(u,v)}(x, y) = \langle x, v\rangle + \langle u, y\rangle.$$

This implies that the derivative of the map $u \mapsto \langle u, u\rangle$ is

$$d\langle -, - \rangle_u(x) = 2\langle x, u\rangle. \qquad (d_1)$$

Since the map $u \mapsto \langle u, v\rangle$ (with v fixed) is linear, its derivative is

$$d\langle -, v\rangle_u(x) = \langle x, v\rangle. \tag{d_2}$$

The derivative of the Lagrangian

$$L(\xi, w, \lambda) = \xi^\top \xi + K\langle w, w\rangle - \sum_{i=1}^m \lambda_i \langle \varphi(x_i), w\rangle - \xi^\top \lambda + \lambda^\top y$$

with respect to ξ and w is

$$dL_{\xi,w}\big(\widetilde{\xi}, \widetilde{w}\big) = 2(\widetilde{\xi})^\top \xi - (\widetilde{\xi})^\top \lambda + \left\langle 2Kw - \sum_{i=1}^m \lambda_i \varphi(x_i), \widetilde{w} \right\rangle,$$

where we used (d_1) to calculate the derivative of $\xi^\top \xi + K\langle w, w\rangle$ and (d_2) to calculate the derivative of $-\sum_{i=1}^m \lambda_i \langle \varphi(x_i), w\rangle - \xi^\top \lambda$. We have $dL_{\xi,w}\big(\widetilde{\xi}, \widetilde{w}\big) = 0$ for all $\widetilde{\xi}$ and $\widetilde{w}$ iff

$$\begin{aligned} 2Kw &= \sum_{i=1}^m \lambda_i \varphi(x_i) \\ \lambda &= 2\xi. \end{aligned}$$

Again we define $\xi = K\alpha$, so we have $\lambda = 2K\alpha$, and

$$w = \sum_{i=1}^m \alpha_i \varphi(x_i).$$

Plugging back into the Lagrangian we get

$$\begin{aligned} G(\alpha) &= K^2 \alpha^\top \alpha + K \sum_{i,j=1}^m \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j)\rangle - 2K \sum_{i,j=1}^m \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j)\rangle \\ &\quad - 2K^2 \alpha^\top \alpha + 2K\alpha^\top y \\ &= -K^2 \alpha^\top \alpha - K \sum_{i,j=1}^m \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j)\rangle + 2K\alpha^\top y. \end{aligned}$$

If $\mathbf{G}$ is the matrix given by $\mathbf{G}_{ij} = \langle \varphi(x_i), \varphi(x_j)\rangle$, then we have

$$G(\alpha) = -K\alpha^\top (\mathbf{G} + KI_m)\alpha + 2K\alpha^\top y.$$

The function G is strictly concave, so by Theorem 40.13(4) it has a maximum for

$$\alpha = (\mathbf{G} + KI_m)^{-1} y,$$

as claimed earlier.

As in the standard case of ridge regression, if $F = \mathbb{R}^n$ (but the inner product $\langle -, - \rangle$ is arbitrary), we can adapt the above method to learn an affine function $f(w) = x^\top w + b$ instead of a linear function $f(w) = x^\top w$, where $b \in \mathbb{R}$. This time we assume that b is of the form

$$b = \overline{y} - \langle w, (\overline{X^1} \cdots \overline{X^n}) \rangle,$$

where X^j is the j column of the $m \times n$ matrix X whose ith row is the transpose of the column vector $\varphi(x_i)$, and where $(\overline{X^1} \cdots \overline{X^n})$ is viewed as a column vector. We have the minimization problem

Program (KRR6′):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K\langle w, w \rangle \\ &\text{subject to} \\ &\qquad\qquad \widehat{y_i} - \langle w, \widehat{\varphi(x_i)} \rangle = \xi_i, \quad i = 1, \ldots, m, \end{aligned}$$

minimizing over ξ and w, where $\widehat{\varphi(x_i)}$ is the n-dimensional vector $\varphi(x_i) - (\overline{X^1} \cdots \overline{X^n})$.

The solution is given in terms of the matrix $\widehat{\mathbf{G}}$ defined by

$$\widehat{\mathbf{G}}_{ij} = \langle \widehat{\varphi(x_i)}, \widehat{\varphi(x_j)} \rangle,$$

as before. We get

$$\alpha = (\widehat{\mathbf{G}} + KI_m)^{-1} \widehat{y},$$

and according to a previous computation, b is given by

$$b = \overline{y} - \frac{1}{m} \mathbf{1} \widehat{\mathbf{G}} \alpha.$$

We explained in Section 53.4 how to compute the matrix $\widehat{\mathbf{G}}$ from the matrix $\mathbf{G}$.

Since the dimension of the feature space F may be very large, one might worry that computing the inner products $\langle \varphi(x_i), \varphi(x_j) \rangle$ might be very expensive. This is where kernel functions come to the rescue. A *kernel function* κ for an embedding $\varphi \colon \mathbb{R}^n \to F$ is a map $\kappa \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the property that

$$\kappa(u, v) = \langle \varphi(u), \varphi(v) \rangle \quad \text{for all } u, v \in \mathbb{R}^n.$$

If $\kappa(u, v)$ can be computed in a reasonably cheap way, and if $\varphi(u)$ can be computed cheaply, then the inner products $\langle \varphi(x_i), \varphi(x_j) \rangle$ (and $\langle \varphi(x_i), \varphi(x) \rangle$) can be computed cheaply; see Chapter 53. Fortunately there are good kernel functions. Two very good sources on kernel methods are Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159].

55.4 Lasso Regression (ℓ^1-Regularized Regression)

The main weakness of ridge regression is that the estimated weight vector w usually has many nonzero coefficients. As a consequence, ridge regression does not scale up well. In practice we need methods capable of handling millions of parameters, or more. A way to encourage sparsity of the vector w, which means that many coordinates of w are zero, is to replace the quadratic penalty function $\tau w^\top w = \tau \|w\|_2^2$ by the penalty function $\tau \|w\|_1$, with the ℓ^2-norm replaced by the ℓ^1-norm.

This method was first proposed by Tibshirani around 1996, under the name *lasso*, which stands for “least absolute selection and shrinkage operator.” This method is also known as *ℓ^1-regularized regression*, but this is not as cute as “lasso,” which is used predominantly.

Given a set of training data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, if X is the $m \times n$ matrix

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

in which the row vectors $x_i^\top$ are the rows of X, then *lasso regression* is the following optimization problem

Program (lasso1):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \|w\|_1 \\ &\text{subject to} \\ &\qquad\qquad y - Xw = \xi, \end{aligned}$$

minimizing over ξ and w, where $\tau > 0$ is some constant determining the influence of the regularizing term $\|w\|_1$.

The difficulty with the regularizing term $\|w\|_1 = |w_1| + \cdots + |w_n|$ is that the map $w \mapsto \|w\|_1$ is not differentiable for all w. This difficulty can be overcome by using subgradients, but the dual of the above program can also be obtained in an elementary fashion by using a trick that we already used, which is that if $x \in \mathbb{R}$, then

$$|x| = \max\{x, -x\}.$$

Using this trick, by introducing a vector $\epsilon \in \mathbb{R}^n$ of nonnegative variables, we can rewrite lasso minimization as follows:

Program lasso regularization (lasso2):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\
&\text{subject to} \\
&\qquad\qquad y - Xw = \xi \\
&\qquad\qquad\quad w \leq \epsilon \\
&\qquad\qquad - w \leq \epsilon.
\end{aligned}
$$

minimizing over ξ, w and ϵ, with $y, \xi \in \mathbb{R}^m$, and $w, \epsilon, \mathbf{1}_n \in \mathbb{R}^n$.

The constraints $w \leq \epsilon$ and $-w \leq \epsilon$ are equivalent to $|w_i| \leq \epsilon_i$ for $i = 1, \ldots, n$, so for an optimal solution we must have $\epsilon \geq 0$ and $|w_i| = \epsilon_i$, that is, $\|w\|_1 = \epsilon_1 + \cdots + \epsilon_n$.

The Lagrangian $L(\xi, w, \epsilon, \lambda, \alpha_+, \alpha_-)$ is given by

$$
\begin{aligned}
L(\xi, w, \epsilon, \lambda, \alpha_+, \alpha_-) &= \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon + \lambda^\top (y - Xw - \xi) \\
&\quad + \alpha_+^\top (w - \epsilon) + \alpha_-^\top (-w - \epsilon) \\
&= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y \\
&\quad + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top (\alpha_+ - \alpha_- - X^\top \lambda),
\end{aligned}
$$

with $\lambda \in \mathbb{R}^m$ and $\alpha_+, \alpha_- \in \mathbb{R}^n_+$. Since the objective function is convex and the constraints are affine (and thus qualified), the Lagrangian L has a minimum with respect to the primal variables, ξ, w, ϵ iff $\nabla L_{\xi, w, \epsilon} = 0$. Since the gradient $\nabla L_{\xi, w, \epsilon}$ is given by

$$
\nabla L_{\xi, w, \epsilon} = \begin{pmatrix} \xi - \lambda \\ \alpha_+ - \alpha_- - X^\top \lambda \\ \tau \mathbf{1}_n - \alpha_+ - \alpha_- \end{pmatrix},
$$

we obtain the equations

$$
\begin{aligned}
\xi &= \lambda \\
\alpha_+ - \alpha_- &= X^\top \lambda \\
\alpha_+ + \alpha_- &= \tau \mathbf{1}_n.
\end{aligned}
$$

Using these equations, the dual function $G(\lambda, \alpha_+, \alpha_-) = \min_{\xi, w, \epsilon} L$ is given by

$$
\begin{aligned}
G(\lambda, \alpha_+, \alpha_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y = \frac{1}{2}\lambda^\top \lambda - \lambda^\top \lambda + \lambda^\top y \\
&= -\frac{1}{2}\lambda^\top \lambda + \lambda^\top y = -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2\right),
\end{aligned}
$$

so

$$
G(\lambda, \alpha_+, \alpha_-) = -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2\right).
$$

Since $\alpha_+, \alpha_- \geq 0$, for any $i \in \{1, \ldots, n\}$ the minimum of $(\alpha_+)_i - (\alpha_-)_i$ is $-\tau$, and the maximum is τ. If we recall that for any $z \in \mathbb{R}^n$,

$$\|z\|_\infty = \max_{1 \leq i \leq n} |z_i|,$$

it follows that the constraints

$$\begin{aligned} \alpha_+ + \alpha_- &= \tau \mathbf{1}_n \\ X^\top \lambda &= \alpha_+ - \alpha_- \end{aligned}$$

are equivalent to

$$\left\|X^\top \lambda\right\|_\infty \leq \tau.$$

The above is equivalent to the $2n$ constraints

$$-\tau \leq (X^\top \lambda)_i \leq \tau, \quad 1 \leq i \leq n.$$

Therefore, the dual lasso program is given by

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2\right) \\ &\text{subject to} \\ &\qquad\qquad \left\|X^\top \lambda\right\|_\infty \leq \tau, \end{aligned}$$

which (since $\|y\|_2^2$ is a constant term) is equivalent to

Program (Dlasso2):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|y - \lambda\|_2^2 \\ &\text{subject to} \\ &\qquad\qquad \left\|X^\top \lambda\right\|_\infty \leq \tau, \end{aligned}$$

minimizing over $\lambda \in \mathbb{R}^m$.

One way to solve lasso regression is to use the dual program to find $\lambda = \xi$, and then to use linear programming to find w by solving the linear program arising from the lasso primal by holding ξ constant. The best way is to use ADMM as explained in Section 52.8(4). There are also a number of variations of gradient descent; see Hastie, Tibshirani, and Wainwright [89].

In theory, if we know the support of w and the signs of its components, then w is determined as we now explain.

In view of the constraint $y - Xw = \xi$ and the fact that for an optimal solution we must have $\xi = \lambda$, the following condition must hold:

$$\left\|X^\top (Xw - y)\right\|_\infty \leq \tau. \qquad (*)$$

Also observe that for an optimal solution, we have

$$\begin{aligned}\frac{1}{2}\left\|y - Xw\right\|_2^2 + w^\top X^\top (y - Xw) &= \frac{1}{2}\left\|y\right\|^2 - w^\top X^\top y + \frac{1}{2} w^\top X^\top X w + w^\top X^\top y - w^\top X^\top X w \\ &= \frac{1}{2}\left(\left\|y\right\|_2^2 - \left\|Xw\right\|_2^2\right) \\ &= \frac{1}{2}\left(\left\|y\right\|_2^2 - \left\|y - \lambda\right\|_2^2\right) = G(\lambda).\end{aligned}$$

Since the objective function is convex and the constaints are qualified, by Theorem 50.19(2) the duality gap is zero, so for optimal solutions of the primal and the dual, $G(\lambda) = L(\xi, w, \epsilon)$, that is

$$\frac{1}{2}\left\|y - Xw\right\|_2^2 + w^\top X^\top (y - Xw) = \frac{1}{2}\left\|\xi\right\|_2^2 + \tau\left\|w\right\|_1 = \frac{1}{2}\left\|y - Xw\right\|_2^2 + \tau\left\|w\right\|_1,$$

which yields the equation

$$w^\top X^\top (y - Xw) = \tau\left\|w\right\|_1. \tag{$**_1$}$$

The above is the inner product of w and $X^\top(y - Xw)$, so whenever $w_i \neq 0$, since $\left\|w\right\|_1 = |w_1| + \cdots + |w_n|$, in view of $(*)$, we must have $(X^\top(y - Xw))_i = \tau \mathrm{sgn}(w_i)$. If

$$S = \{i \in \{1, \ldots, n\} \mid w_i \neq 0\}, \tag{$\dagger$}$$

if X_S denotes the matrix consisting of the columns of X indexed by S, and if w_S denotes the vector consisting of the nonzero components of w, then we have

$$X_S^\top (y - X_S w_S) = \tau \mathrm{sgn}(w_S). \tag{$**_2$}$$

We also have

$$\left\|X_{\overline{S}}^\top (y - X_S w_S)\right\|_\infty \leq \tau, \tag{$**_3$}$$

where $\overline{S}$ is the complement of S.

Equation $(**_2)$ yields

$$X_S^\top X_S w_S = X_S^\top y - \tau \mathrm{sgn}(w_S),$$

so if $X_S^\top X_S$ is invertible (which will be the case if the columns of X are linearly independent), we get

$$w_S = (X_S^\top X_S)^{-1}(X_S^\top y - \tau \mathrm{sgn}(w_S)). \tag{$**_4$}$$

In theory, if we know the support of w and the signs of its components, then w_S is determined, but in practice this is useless since the problem is to find the support and the sign of the solution.

55.5 Lasso Regression; Learning an Affine Function

In the preceding section we made the simplifying assumption that we were trying to learn a linear function $f(x) = x^\top w$. To learn an affine function $f(x) = x^\top w + b$, we solve the following optimization problem

Program (lasso3):

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\ &\text{subject to} && \\ &&& y - Xw - b\mathbf{1}_m = \xi \\ &&& w \leq \epsilon \\ &&& -w \leq \epsilon. \end{aligned}$$

Observe that as in the case of ridge regression, minimization is performed over ξ, w, ϵ and b, but b is not penalized in the objective function.

The Lagrangian associated with this optimization problem is

$$\begin{aligned} L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) = {} & \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda \\ & + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top (\alpha_+ - \alpha_- - X^\top \lambda), \end{aligned}$$

so by setting the gradient $\nabla L_{\xi, w, \epsilon, b}$ to zero we obtain the equations

$$\begin{aligned} \xi &= \lambda \\ \alpha_+ - \alpha_- &= X^\top \lambda \\ \alpha_+ + \alpha_- &= \tau \mathbf{1}_n \\ \mathbf{1}_m^\top \lambda &= 0. \end{aligned}$$

Using these equations, we find that the dual function is also given by

$$G(\lambda, \alpha_+, \alpha_-) = -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2 \right),$$

and the dual lasso program is given by

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2 \right) \\ &\text{subject to} && \\ &&& \left\| X^\top \lambda \right\|_\infty \leq \tau \\ &&& \mathbf{1}_m^\top \lambda = 0, \end{aligned}$$

which is equivalent to

Program (Dlasso3):

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|y-\lambda\|_2^2 \\ &\text{subject to} && \\ & && \left\|X^\top\lambda\right\|_\infty \le \tau \\ & && \mathbf{1}_m^\top \lambda = 0, \end{aligned}$$

minimizing over $\lambda \in \mathbb{R}^m$.

Once $\lambda = \xi$ and w are determined, we obtain b using the equation

$$b\mathbf{1}_m = y - Xw - \xi,$$

and since $\mathbf{1}_m^\top \mathbf{1}_m = m$ and $\mathbf{1}_m^\top \xi = \mathbf{1}_m^\top \lambda = 0$, the above yields

$$b = \frac{1}{m}\mathbf{1}_m^\top y - \frac{1}{m}\mathbf{1}_m^\top Xw - \frac{1}{m}\mathbf{1}_m^\top \xi = \overline{y} - \sum_{j=1}^n \overline{X^j} w_j,$$

where $\overline{y}$ is the mean of y and $\overline{X^j}$ is the mean of the jth column of X.

The equation

$$b = \widehat{b} + \overline{y} - \sum_{j=1}^n \overline{X^j} w_j = \widehat{b} + \overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n})w,$$

can be used as in ridge regression, (see Section 55.2), to show that the Program (**lasso3**) is equivalent to applying lasso regression (**lasso2**) without an intercept term to the centered data, by replacing y by $\widehat{y} = y - \overline{y}\mathbf{1}$ and X by $\widehat{X} = X - \overline{X}$. Then b is given by

$$b = \overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n})\widehat{w},$$

where $\widehat{w}$ is the solution given by (**lasso2**). This is the method described by Hastie, Tibshirani, and Wainwright [89] (Section 2.2).

Example 55.3. We can create a data set (X, y) where X a 100×5 matrix and y is a 100×1 vector using the following `Matlab` program in which the command `randn` creates an array of normally distributed numbers.

```
X = randn(100,5);
ww = [0; 2; 0; -3; 0];
y = X*ww + randn(100,1)*0.1;
```

The purpose of the third line is to add some small noise to the "output" $X * ww$. The first five rows of X are

$$\begin{pmatrix} -1.1658 & -0.0679 & -1.6118 & 0.3199 & 0.4400 \\ -1.1480 & -0.1952 & -0.0245 & -0.5583 & -0.6169 \\ 0.1049 & -0.2176 & -1.9488 & -0.3114 & 0.2748 \\ 0.7223 & -0.3031 & 1.0205 & -0.5700 & 0.6011 \\ 2.5855 & 0.0230 & 0.8617 & -1.0257 & 0.0923 \end{pmatrix},$$

and the first five rows of y are

$$y = \begin{pmatrix} -1.0965 \\ 1.2155 \\ 0.4324 \\ 1.1902 \\ 3.1346 \end{pmatrix}.$$

We ran the program for lasso using ADMM (see Problem 52.7) with various values of ρ and τ, including $\rho = 1$ and $\rho = 10$. We observed that the program converges a lot faster for $\rho = 10$ than for $\rho = 1$. We plotted the values of the five components of $w(\tau)$ for values of τ from $\tau = 0$ to $\tau = 0.5$ by increment of 0.02, and observed that the first, third, and fifth coordinate drop basically linearly to zero (a value less that 10^{-4}) around $\tau = 0.2$. See Figures 55.7, 55.8, and 55.9. This behavior is also observed in Hastie, Tibshirani, and Wainwright [89] (Section 2.2).

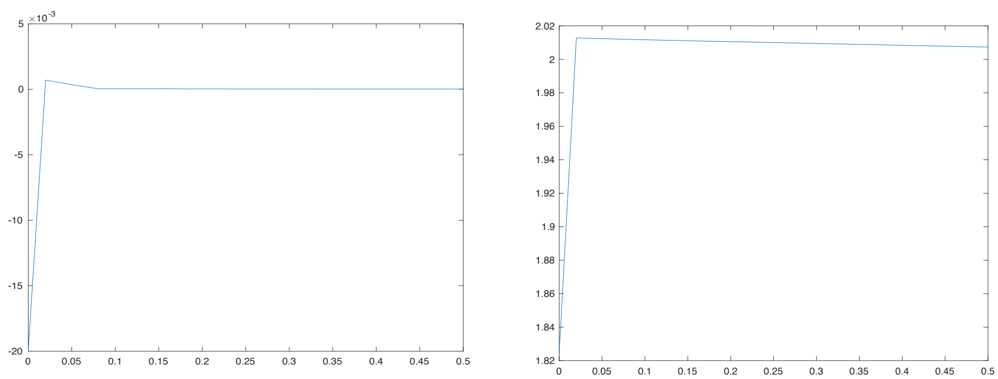


Figure 55.7: First and second component of w.

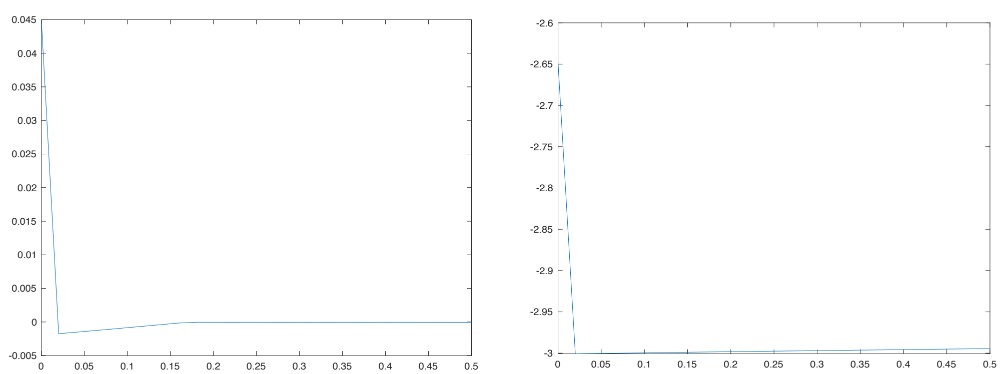


Figure 55.8: Third and fourth component of w.

For $\tau = 0.02$, we have

$$w = \begin{pmatrix} 0.00003 \\ 2.01056 \\ -0.00004 \\ -2.99821 \\ 0.00000 \end{pmatrix}, \quad b = 0.00135.$$

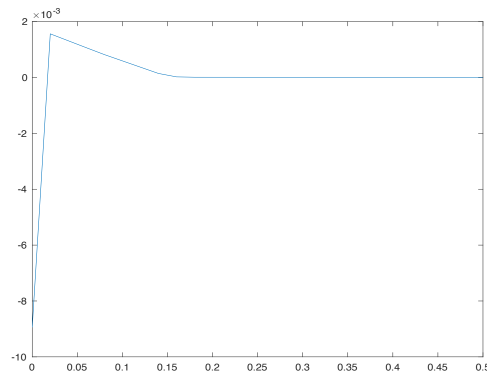


Figure 55.9: Fifth component of w.

This weight vector w is very close to the original vector $ww = [0; 2; 0; -3; 0]$ that we used to create y. For large values of τ, the weight vector is essentially the zero vector. This happens for $\tau = 235$, where every component of w is less than 10^{-5}.

Another way to find b is to add the term $(C/2)b^2$ to the objective function, for some positive constant C, obtaining the program

Program(lasso4):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon + \frac{1}{2} C b^2 \\ &\text{subject to} \\ &\qquad y - Xw - b\mathbf{1}_m = \xi \\ &\qquad w \leq \epsilon \\ &\qquad -w \leq \epsilon, \end{aligned}$$

minimizing over ξ, w, ϵ and b.

This time the Lagrangian is

$$\begin{aligned} L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y + \frac{C}{2} b^2 - b\mathbf{1}_m^\top \lambda \\ &\quad + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top (\alpha_+ - \alpha_- - X^\top \lambda), \end{aligned}$$

so by setting the gradient $\nabla L_{\xi,w,\epsilon,b}$ to zero we obtain the equations

$$\begin{aligned}
\xi &= \lambda \\
\alpha_+ - \alpha_- &= X^\top \lambda \\
\alpha_+ + \alpha_- &= \tau \mathbf{1}_n \\
Cb &= \mathbf{1}_m^\top \lambda.
\end{aligned}$$

Thus b is also determined, and the dual lasso program is identical to the first lasso dual (**Dlasso2**), namely

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\|y-\lambda\|_2^2 \\
&\text{subject to} \\
&\qquad\qquad \left\|X^\top \lambda\right\|_\infty \leq \tau,
\end{aligned}$$

minimizing over λ.

Since the equations $\xi = \lambda$ and

$$y - Xw - b\mathbf{1}_m = \xi$$

hold, from $Cb = \mathbf{1}_m^\top \lambda$ we get

$$\frac{1}{m}\mathbf{1}_m^\top y - \frac{1}{m}\mathbf{1}_m^\top X w - b\frac{1}{m}\mathbf{1}_m^\top \mathbf{1} = \frac{1}{m}\mathbf{1}_m^\top \lambda,$$

that is

$$\overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n})w - b = \frac{C}{m}b,$$

which yields

$$b = \frac{m}{m+C}(\overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n})w).$$

As in the case of ridge regression, a defect of the approach where b is also penalized is that the solution for b is not invariant under adding a constant c to each value y_i

It is interesting to compare the behavior of the methods:

1. Ridge regression (**RR6′**) (which is equivalent to (**RR3**)).
2. Ridge regression (**RR3**b), with b penalized (by adding the term Kb^2 to the objective function).
3. Least squares applied to $[X \ \ \mathbf{1}]$.
4. (**lasso3**).

When $n \leq 2$ and K and τ are small and of the same order of magnitude, say 0.1 or 0.01, there is no noticeable difference. We ran out programs on the data set of 200 points generated by the following `Matlab` program:

```
X14 = 15*randn(200,1);
ww14 = 1;
y14 = X14*ww14 + 10*randn(200,1) + 20;
```

The result is shown in Figure 55.10, with the following colors: Method (1) in magenta, Method (2) in red, Method (3) in blue, and Method (4) in cyan. All four lines are identical.

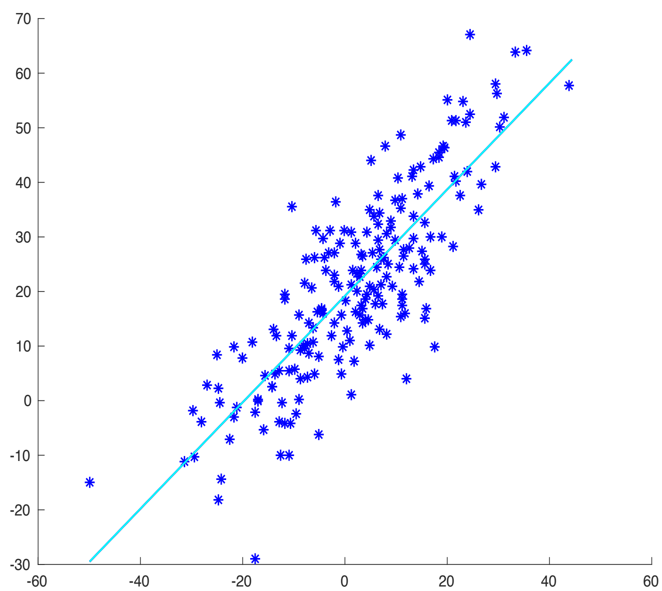


Figure 55.10: Comparison of the four methods with $K = \tau = 0.1$.

In order to see a difference we also ran our programs with $K = 1000$ and $\tau = 10000$; see Figure 55.11.

As expected, due to the penalization of b, Method (3) yields a significantly lower line (in red), and due to the large value of τ, the line corresponding to lasso (in cyan) has a smaller slope. Method (1) (in magenta) also has a smaller slope but still does not deviate that much from least squares (in blue). It is also interesting to experiment on data sets where n is larger (say $20, 50$).

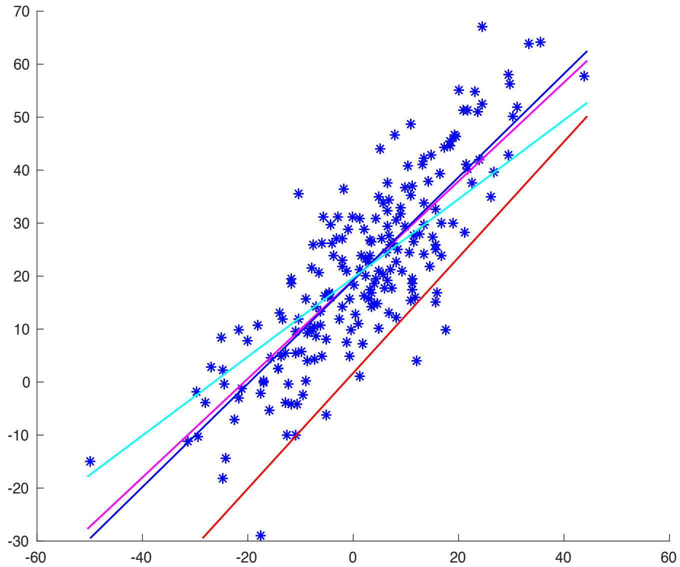


Figure 55.11: Comparison of the four methods with $K = 1000, \tau = 10000$.

55.6 Elastic Net Regression

The lasso method is unsatisfactory when n (the dimension of the data) is much larger than the number m of data, because it only selects m coordinates and sets the others to values close to zero. It also has problems with groups of highly correlated variables. A way to overcome this problem is to add a "ridge-like" term $(1/2)Kw^\top w$ to the objective function. This way we obtain a hybrid of lasso and ridge regression called the *elastic net method* and defined as follows:

Program (elastic net):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \frac{1}{2}K w^\top w + \tau \mathbf{1}_n^\top \epsilon \\
&\text{subject to} \\
&\qquad y - Xw - b\mathbf{1}_m = \xi \\
&\qquad w \leq \epsilon \\
&\qquad -w \leq \epsilon,
\end{aligned}$$

where $K > 0$ and $\tau \geq 0$ are two constants controlling the influence of the ℓ^2-regularization and the ℓ^1-regularization.[1] Observe that as in the case of ridge regression, minimization is performed over ξ, w, ϵ and b, but b is not penalized in the objective function. The objective

[1]Some of the literature denotes K by λ_2 and τ by λ_1, but we prefer not to adopt this notation since we use λ, μ *etc.* to denote Lagrange multipliers.

function is strictly convex so if an optimal solution exists, then it is unique; the proof is left as an exercise.

The Lagrangian associated with this optimization problem is

$$\begin{aligned} L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda \\ &\quad + \epsilon^\top(\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2} K w^\top w, \end{aligned}$$

so by setting the gradient $\nabla L_{\xi,w,\epsilon,b}$ to zero we obtain the equations

$$\begin{aligned} \xi &= \lambda \\ Kw &= -(\alpha_+ - \alpha_- - X^\top \lambda) \qquad (*_w) \\ \alpha_+ + \alpha_- - \tau \mathbf{1}_n &= 0 \\ \mathbf{1}_m^\top \lambda &= 0. \end{aligned}$$

We find that $(*_w)$ determines w. Using these equations, we can find the dual function but in order to solve the dual using ADMM, since $\lambda \in \mathbb{R}^m$, it is more convenient to write $\lambda = \lambda_+ - \lambda_-$, with $\lambda_+, \lambda_- \in \mathbb{R}^m_+$ (recall that $\alpha_+, \alpha_- \in \mathbb{R}^n_+$). As in the derivation of the dual of ridge regression, we rescale our variables by defining $\beta_+, \beta_-, \mu_+, \mu_-$ such that

$$\alpha_+ = K\beta_+, \;\; \alpha_- = K\beta_-, \;\; \lambda_+ = K\mu_+, \;\; \lambda_- = K\mu_-.$$

We also let $\mu = \mu_+ - \mu_-$ so that $\lambda = K\mu$. Then $\mathbf{1}_m^\top \lambda = 0$ is equivalent to

$$\mathbf{1}_m^\top \mu_+ - \mathbf{1}_m^\top \mu_- = 0,$$

and since $\xi = \lambda = K\mu$, we have

$$\begin{aligned} \xi &= K(\mu_+ - \mu_-) \\ \beta_+ + \beta_- &= \frac{\tau}{K} \mathbf{1}_n. \end{aligned}$$

Using $(*_w)$ we can write

$$\begin{aligned} w &= -(\beta_+ - \beta_- - X^\top \mu) = -\beta_+ + \beta_- + X^\top \mu_+ - X^\top \mu_- \\ &= \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}. \end{aligned}$$

Then we have

$$\begin{pmatrix}-I_n & I_n & X^\top & -X^\top\end{pmatrix}^\top \begin{pmatrix}-I_n & I_n & X^\top & -X^\top\end{pmatrix} = \begin{pmatrix}-I_n \\ I_n \\ X \\ -X\end{pmatrix} \begin{pmatrix}-I_n & I_n & X^\top & -X^\top\end{pmatrix}$$
$$= \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top & -XX^\top \\ X & -X & -XX^\top & XX^\top \end{pmatrix}.$$

If we define the symmetric positive semidefinite $2(n+m) \times 2(n+m)$ matrix Q as

$$Q = \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top & -XX^\top \\ X & -X & -XX^\top & XX^\top \end{pmatrix},$$

then

$$\frac{1}{2} w^\top w = \frac{1}{2} \begin{pmatrix}\beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top\end{pmatrix} Q \begin{pmatrix}\beta_+ \\ \beta_- \\ \mu_+ \\ \mu_-\end{pmatrix}.$$

As a consequence, using $(*_w)$ and the fact that $\xi = K\mu$, we find that the dual function is given by

$$\begin{aligned} G(\mu, \beta_+, \beta_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2} K w^\top w \\ &= \frac{1}{2}\xi^\top \xi - K\xi^\top \mu + K\mu^\top y + K w^\top(\beta_+ - \beta_- - X^\top \mu) + \frac{1}{2} K w^\top w \\ &= \frac{1}{2} K^2 \mu^\top \mu - K^2 \mu^\top \mu + K y^\top \mu - K w^\top w + \frac{1}{2} K w^\top w \\ &= -\frac{1}{2} K^2 \mu^\top \mu - \frac{1}{2} K w^\top w + K y^\top \mu. \end{aligned}$$

But

$$\mu = \begin{pmatrix} I_m & -I_m \end{pmatrix} \begin{pmatrix}\mu_+ \\ \mu_-\end{pmatrix},$$

so

$$\frac{1}{2}\mu^\top \mu = \frac{1}{2} \begin{pmatrix}\mu_+^\top & \mu_-^\top\end{pmatrix} \begin{pmatrix} I_m & -I_m \\ -I_m & I_m \end{pmatrix} \begin{pmatrix}\mu_+ \\ \mu_-\end{pmatrix},$$

so we get

$$G(\beta_+, \beta_-, \mu_+, \mu_-) = -\frac{1}{2}K \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} P \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix} - Kq^\top \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}$$

with

$$\begin{aligned} P &= Q + K \begin{pmatrix} 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{m,n} & 0_{m,n} & I_m & -I_m \\ 0_{m,n} & 0_{m,n} & -I_m & I_m \end{pmatrix} \\ &= \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top + KI_m & -XX^\top - KI_m \\ X & -X & -XX^\top - KI_m & XX^\top + KI_m \end{pmatrix}, \end{aligned}$$

and

$$q = \begin{pmatrix} 0_n \\ 0_n \\ -y \\ y \end{pmatrix}.$$

The constraints are the equations

$$\begin{aligned} \beta_+ + \beta_- &= \frac{\tau}{K}\mathbf{1}_n \\ \mathbf{1}_m^\top \mu_+ - \mathbf{1}_m^\top \mu_- &= 0, \end{aligned}$$

which correspond to the $(n+1) \times 2(n+m)$ matrix

$$A = \begin{pmatrix} I_n & I_n & 0_{n,m} & 0_{n,m} \\ 0_n^\top & 0_n^\top & \mathbf{1}_m^\top & -\mathbf{1}_m^\top \end{pmatrix}$$

and the right-hand side

$$c = \begin{pmatrix} \frac{\tau}{K}\mathbf{1}_n \\ 0 \end{pmatrix}.$$

Since $K > 0$, the dual of elastic net is equivalent to

Program (Dual Elastic Net):

$$\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top\end{pmatrix} P \begin{pmatrix}\beta_+ \\ \beta_- \\ \mu_+ \\ \mu_-\end{pmatrix} + q^\top \begin{pmatrix}\beta_+ \\ \beta_- \\ \mu_+ \\ \mu_-\end{pmatrix}$$

subject to

$$A \begin{pmatrix}\beta_+ \\ \beta_- \\ \mu_+ \\ \mu_-\end{pmatrix} = c,$$

$\beta_+, \beta_- \in \mathbb{R}^n_+$, $\mu_+, \mu_- \in \mathbb{R}^m_+$.

Once $\xi = K\mu = K(\mu_+ - \mu_-)$ and w are determined by $(*_w)$, we obtain b using the equation

$$b\mathbf{1}_m = y - Xw - \xi,$$

which as in Section 55.5 yields

$$b = \overline{y} - \sum_{j=1}^{n} \overline{X^j} w_j,$$

where $\overline{y}$ is the mean of y and $\overline{X^j}$ is the mean of the jth column of X.

We leave it as an easy exercise to show that A has rank $n+1$. Then we can solve the above program using ADMM, and we have done so. This very similar to what is done in Section 56.3, and hence the details are left as an exercise.

Observe that when $\tau = 0$, the elastic net method reduces to ridge regression. As K tends to 0 the elastic net method tends to lasso, but $K = 0$ is not an allowable value since $\tau/0$ is undefined. Anyway, if we get rid of the constraint

$$\beta_+ + \beta_- = \frac{\tau}{K}\mathbf{1}_n$$

the corresponding optimization program not does determine w. Experimenting with our program we found that convergence becomes very slow for $K < 10^{-3}$. What we can do if K is small, say $K < 10^{-3}$, is to run lasso. A nice way to "blend" ridge regression and lasso is to call the elastic net method with $K = C(1-\theta)$ and $\tau = C\theta$, where $0 \leq \theta < 1$ and $C > 0$.

Running the elastic net method on the data set $(X14, y14)$ of Section 55.5 with $K = \tau = 0.5$ shows absolutely no difference, but the reader should conduct more experiments to see how elastic net behaves as K and τ are varied (the best way to do this is to use θ as explained above). Here is an example involving a data set $(X20, y20)$ where $X20$ is a 200×30 matrix generated as follows:

```
X20 = randn(50,30);
ww20 = [0; 2; 0; -3; 0; -4; 1; 0; 2; 0; 2; 3; 0; -5; 6; 0; 1; 2; 0; 10;
0; 0; 3; 4; 5; 0; 0; -6; -8; 0];
y20 = X20*ww20 + randn(50,1)*0.1 + 5;
```

Running our program with $K = 0.01$ and $K = 0.99$, and then with $K = 0.99$ and $K = 0.01$, we get the following weight vectors (in the left column is the weight vector corresponding to $K = 0.01$ and $K = 0.99$):

```
0.0254          0.2007
1.9193          2.0055
0.0766          0.0262
-3.0014         -2.8008
0.0512          0.0089
-3.8815         -3.7670
0.9591          0.8552
-0.0086         -0.3243
1.9576          1.9080
-0.0077         -0.1041
1.9881          2.0566
2.9223          2.8346
-0.0046         -0.0832
-4.9989         -4.8332
5.8640          5.4598
-0.0207         -0.2141
0.8285          0.8585
1.9310          1.8559
0.0046          0.0413
9.9232          9.4836
-0.0216         0.0303
0.0453          -0.0193
2.9384          3.0004
4.0525          3.9753
4.8723          4.6530
0.0767          0.1192
0.0132          -0.0203
-5.9750         -5.7537
-7.9764         -7.7594
-0.0054         0.0528
```

Generally, the numbers in the left column, which are more "lasso-like," have clearer zeros and nonzero values closer to those of the weight vector $ww20$ that was used to create the data set. The value of b corresponding to the first call is $b = 5.1372$, and the value of b corresponding to the second call is $b = 5.208$.

We have observed that lasso seems to converge much faster than elastic net when $K < 10^{-3}$. For example, running the above data set with $K = 10^{-3}$ and $\tau = 0.999$ requires 140520 steps to achieve primal and dual residuals less than 10^{-7}, but lasso only takes 86 steps to achieve the same degree of convergence. We observed that the larger K is the faster is the convergence. This could be attributed to the fact that the matrix P becomes more "positive definite." Another factor is that ADMM for lasso solves an $n \times n$ linear system, but ADMM for elastic net solves a $2(n + m) \times 2(n + m)$ linear system. So even though elastic net does not suffer from some of the undesirable properties of ridge regression and lasso, it appears to have a slower convergence rate, in fact much slower when K is small (say $K < 10^{-3}$). It also appears that elastic net may be too expensive a choice if m is much larger than n. Further investigations are required to gain a better understanding of the convergence issue.

55.7 Summary

The main concepts and results of this chapter are listed below:

- Ridge regression.
- Kernel ridge regression.
- Kernel functions.
- Lasso regression.
- Elastic net regression.

55.8 Problems

Problem 55.1. Check the formula

$$(X^\top X + KI_n)^{-1}X^\top = X^\top(XX^\top + KI_m)^{-1},$$

stated in Section 55.1.

Problem 55.2. Implement the ridge regression method described in Section 55.1 in `Matlab`. Also implement ridge regression with intercept and compare solving Program (**DRR3**) and Program (**RR6′**) using centered data.

Problem 55.3. Implement the ridge regression with intercept method (**RR3b**) in `Matlab` and compare it with solving (**RR6′**) using centered data.

Problem 55.4. Verify that (**lasso3**) is equivalent to (**lasso2**) applied to the centered data $\widehat{y} = y - \overline{y}\mathbf{1}$ and $\widehat{X} = X - \overline{X}$.

Problem 55.5. Verify the fomulae obtained for the kernel ridge regression program (**KRR6′**).

Problem 55.6. Implement in `Matlab` and test (**lasso3**) for various values of ρ and τ. Write a program to plot the coordinates of w as a function of τ. Compare the behavior of lasso with ridge regression (**RR6**$'$), (**RR3**b) (b penalized), and with least squares.

Problem 55.7. Check the details of the derivation of the dual of elastic net.

Problem 55.8. Write a `Matlab` program, solving the dual of elastic net; use inspiration from Section 56.3. Run tests to compare the behavior of ridge regression, lasso, and elastic net.

Problem 55.9. Prove that if an optimal solution exists for the elastic net method, then it is unique.

Problem 55.10. Prove that the matrix

$$P = \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top + KI_m & -XX^\top - KI_m \\ X & -X & -XX^\top - KI_m & XX^\top + KI_m \end{pmatrix}$$

is almost positive definite, in the sense that

$$\begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} P \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix} = 0$$

if and only if $\beta_+ = \beta_-$ and $\mu_+ = \mu_-$, that is, $\beta = 0$ and $\mu = 0$.

Chapter 56

ν-SV Regression

56.1 ν-SV Regression; Derivation of the Dual

Let $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a set of observed data usually called a set of *training data*, with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. As in Chapter 55, we form the $m \times n$ matrix X where the row vectors $x_i^\top$ are the *rows* of X. Our goal is to learn an affine function f of the form $f(x) = x^\top w + b$ that fits the set of training data, but does not penalize errors below some given $\epsilon \geq 0$. Geometrically, we view the pairs (x_i, y_i) are points in $\mathbb{R}^{n+1}$, and we try to fit a hyperplane $H_{w,b}$ of equation

$$(w^\top \ \ -1) \begin{pmatrix} x \\ z \end{pmatrix} + b = w^\top x - z + b = 0$$

that best fits the set of points (x_i, y_i) (where $(x, z) \in \mathbb{R}^{n+1}$). We seek an $\epsilon > 0$ such that most points (x_i, y_i) are inside the slab (or tube) of width 2ϵ bounded by the hyperplane $H_{w,b-\epsilon}$ of equation

$$(w^\top \ \ -1) \begin{pmatrix} x \\ z \end{pmatrix} + b - \epsilon = w^\top x - z + b - \epsilon = 0$$

and the hyperplane $H_{w,b+\epsilon}$ of equation

$$(w^\top \ \ -1) \begin{pmatrix} x \\ z \end{pmatrix} + b + \epsilon = w^\top x - z + b + \epsilon = 0.$$

Observe that the hyperplanes $H_{w,b-\epsilon}, H_{w,b}$ and $H_{w,b+\epsilon}$ intersect the z-axis when $x = 0$ for the values $(b - \epsilon, b, b + \epsilon)$. Since $\epsilon \geq 0$, the hyperplane $H_{w,b-\epsilon}$ is below the hyperplane $H_{w,b}$ which is below the hyperplane $H_{w,b+\epsilon}$. We refer to the lower hyperplane $H_{w,b-\epsilon}$ as the *blue margin*, to the upper hyperplane $H_{w,b+\epsilon}$ as the *red margin*, and to the hyperplane $H_{w,b}$ as the *best fit hyperplane*. Also note that since the term $-z$ appears in the equations of these hyperplanes, points for which $w^\top x - z + b \leq 0$ are *above* the hyperplane $H_{w,b}$, and points for which $w^\top x - z + b \geq 0$ are *below* the hyperplane $H_{w,b}$ (and similarly for $H_{w,b-\epsilon}$ and

$H_{b+\epsilon}$). The region bounded by the hyperplanes $H_{w,b-\epsilon}$ and $H_{b+\epsilon}$ (which contains the best fit hyperplane $H_{w,b}$) is called the *ϵ-slab*.

We also allow *errors* by allowing the point (x_i, y_i) to be outside of the ϵ-slab but in the slab between the hyperplane $H_{w,b-\epsilon-\xi_i}$ of equation

$$(w^\top \quad -1)\begin{pmatrix} x \\ z \end{pmatrix} + b - \epsilon - \xi_i = w^\top x - z + b - \epsilon - \xi_i = 0$$

for some $\xi_i > 0$ (which is below the blue margin hyperplane $H_{w,b-\epsilon}$) and the hyperplane $H_{w,b+\epsilon+\xi'_i}$ of equation

$$(w^\top \quad -1)\begin{pmatrix} x \\ z \end{pmatrix} + b + \epsilon + \xi'_i = w^\top x_i - z + b + \epsilon + \xi'_i = 0$$

for some $\xi'_i > 0$ (which is above the red margin hyperplane $H_{w,b+\epsilon}$),
so that $w^\top x_i - y_i + b - \epsilon - \xi_i \leq 0$ and $w^\top x_i - y_i + b + \epsilon + \xi'_i \geq 0$, that is,

$$\begin{aligned} f(x) - y_i = w^\top x_i + b - y_i &\leq \epsilon + \xi_i, \\ -(f(x) - y_i) = -w^\top x_i - b + y_i &\leq \epsilon + \xi'_i. \end{aligned}$$

Our goal is to minimize ϵ and the errors ξ_i and ξ'_i. See Figure 56.1. The trade off between

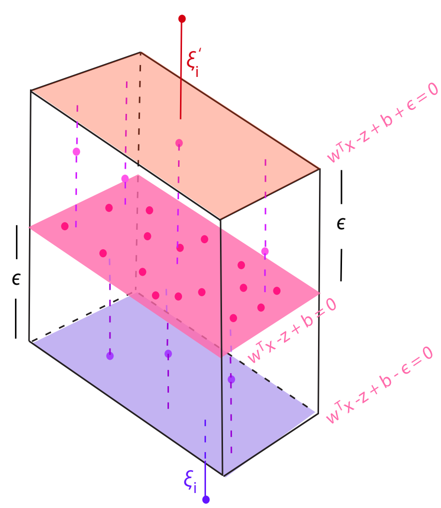


Figure 56.1: The ϵ-slab around the graph of the best fit affine function $f(x) = x^\top w + b$.

the size of ϵ and the size of the slack variables ξ_i and ξ'_i is achieved by using two constants $\nu \geq 0$ and $C > 0$. The method of *ν-support vector regression*, for short *ν-SV regression*, is specified by the following minimization problem:

Program ν-SV Regression:

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i')\right)\\
&\text{subject to}\\
&\qquad w^\top x_i + b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m\\
&\quad - w^\top x_i - b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m\\
&\quad \epsilon \geq 0,
\end{aligned}
$$

minimizing over the variables w, b, ϵ, ξ, and ξ'. The constraints are affine. The problem is to minimize ϵ and the errors ξ_i, ξ_i' so that the ℓ^1-error is "squeezed down" to zero as much as possible, in the sense that the right-hand side of the inequality

$$
\sum_{i=1}^{m} |y_i - x_i^\top w - b| \leq m\epsilon + \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \xi_i'
$$

is as small as possible. As shown by Figure 56.2, the region associated with the constraint $w^\top x_i - z + b \leq \epsilon$ contains the ϵ-slab. Similarly, as illustrated by Figure 56.3, the region associated with the constraint $w^\top x_i - z + b \geq -\epsilon$, equivalently $-w^\top x_i + z - b \leq \epsilon$, also contains the ϵ-slab.

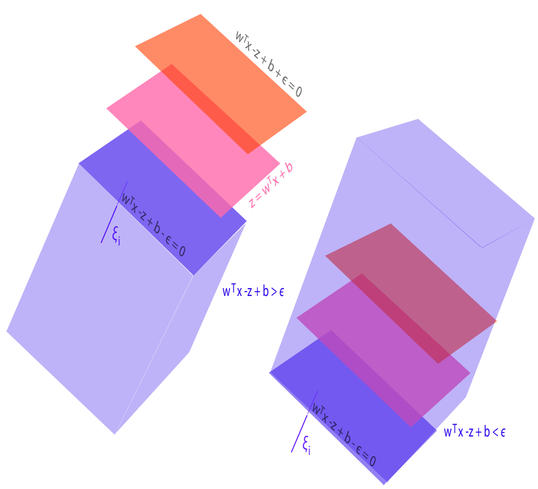


Figure 56.2: The two blue half spaces associated with the hyperplane $w^\top x_i - z + b = \epsilon$.

Observe that if we require $\epsilon = 0$, then the problem is equivalent to minimizing

$$
\|y - Xw - b\mathbf{1}\|_1 + \frac{1}{2}w^\top w.
$$

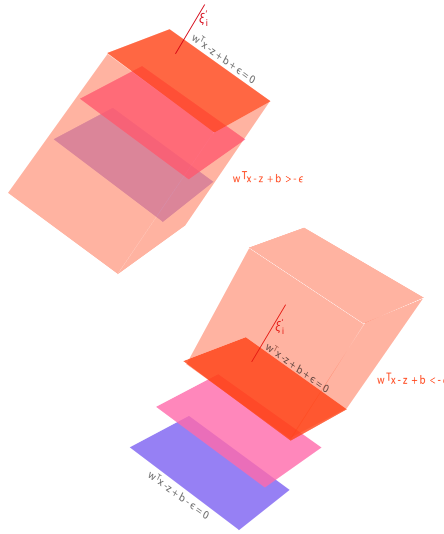


Figure 56.3: The two red half spaces associated with the hyperplane $w^\top x_i - z + b = -\epsilon$.

Thus it appears that the above problem is the version of Program (**RR3**) (see Section 55.2) in which the ℓ^2-norm of $y - Xw - b\mathbf{1}$ is replaced by its ℓ^1-norm. This a sort of "dual" of lasso (see Section 55.5) where $(1/2)w^\top w = (1/2)\left\|w\right\|_2^2$ is replaced by $\tau\left\|w\right\|_1$, and $\left\|y - Xw - b\mathbf{1}\right\|_1$ is replaced by $\left\|y - Xw - b\mathbf{1}\right\|_2^2$.

Proposition 56.1. *For any optimal solution, the equations*

$$\xi_i \xi_i' = 0, \quad i = 1, \ldots, m \tag{$\xi\xi'$}$$

hold. If $\epsilon > 0$, then the equations

$$\begin{aligned} w^\top x_i + b - y_i &= \epsilon + \xi_i \\ -w^\top x_i - b + y_i &= \epsilon + \xi_i' \end{aligned}$$

cannot hold simultaneously.

Proof. For an optimal solution we have

$$-\epsilon - \xi_i' \leq w^\top x_i + b - y_i \leq \epsilon + \xi_i.$$

If $w^\top x_i + b - y_i \geq 0$, then $\xi_i' = 0$ since the inequality

$$-\epsilon - \xi_i' \leq w^\top x_i + b - y_i$$

is trivially satisfied (because $\epsilon, \xi_i' \geq 0$), and if $w^\top x_i + b - y_i \leq 0$, then similarly $\xi_i = 0$. See Figure 56.4.

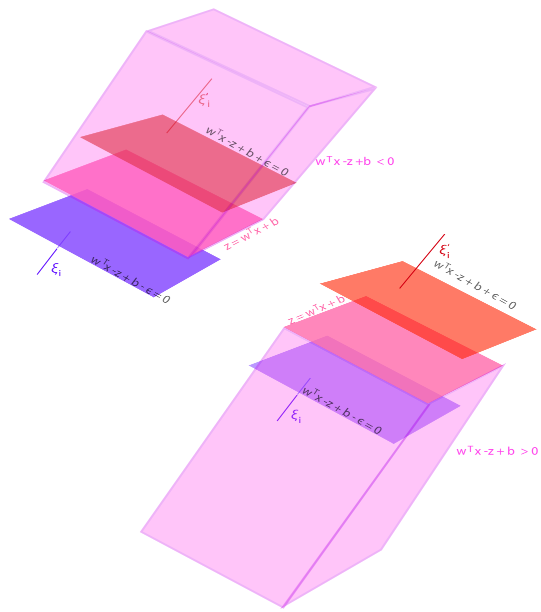


Figure 56.4: The two pink open half spaces associated with the hyperplane $w^\top x_i - z + b = 0$. Note, $\xi_i > 0$ is outside of the half space $w^\top x_i - z + b - \epsilon < 0$, and $\xi_i' > 0$ is outside of the half space $w^\top x_i - z + b + \epsilon > 0$.

Observe that the equations

$$\begin{aligned} w^\top x_i + b - y_i &= \epsilon + \xi_i \\ -w^\top x_i - b + y_i &= \epsilon + \xi_i' \end{aligned}$$

can only hold simultaneously if

$$\epsilon + \xi_i = -\epsilon - \xi',$$

that is,

$$2\epsilon + \xi_i + \xi_i' = 0,$$

and since $\epsilon, \xi_i, \xi_i' \geq 0$, this can happen only if $\epsilon = \xi_i = \xi_i' = 0$, and then

$$w^\top x_i + b = y_i.$$

In particular, if $\epsilon > 0$, then the equations

$$\begin{aligned} w^\top x_i + b - y_i &= \epsilon + \xi_i \\ -w^\top x_i - b + y_i &= \epsilon + \xi_i' \end{aligned}$$

cannot hold simultaneously. $\square$

Observe that if $\nu > 1$, then an optimal solution of the above program must yield $\epsilon = 0$. Indeed, if $\epsilon > 0$, we can reduce it by a small amount $\delta > 0$ and increase $\xi_i + \xi_i'$ by δ to still satisfy the constraints, but the objective function changes by the amount $-\nu\delta + \delta$, which is negative since $\nu > 1$, so $\epsilon > 0$ is not optimal.

Driving ϵ to zero is not the intended goal, because typically the data is not noise free so very few pairs (x_i, y_i) will satisfy the equation $w^\top x_i + b = y_i$, and then many pairs (x_i, y_i) will correspond to an error ($\xi_i > 0$ or $\xi_i' > 0$). Thus, *typically we assume that* $0 < \nu \leq 1$.

To construct the Lagrangian, we assign Lagrange multipliers $\lambda_i \geq 0$ to the constraints $w^\top x_i + b - y_i \leq \epsilon + \xi_i$, Lagrange multipliers $\mu_i \geq 0$ to the constraints $-w^\top x_i - b + y_i \leq \epsilon + \xi_i'$, Lagrange multipliers $\alpha_i \geq 0$ to the constraints $\xi_i \geq 0$, Lagrange multipliers $\beta_i \geq 0$ to the constraints $\xi_i' \geq 0$, and the Lagrange multiplier $\gamma \geq 0$ to the constraint $\epsilon \geq 0$. The Lagrangian is

$$
\begin{aligned}
L(w, b, \lambda, \mu, \gamma, \xi, \xi', \epsilon, \alpha, \beta) &= \frac{1}{2} w^\top w + C\bigg(\nu\epsilon + \frac{1}{m}\sum_{i=1}^m (\xi_i + \xi_i')\bigg) - \gamma\epsilon - \sum_{i=1}^m (\alpha_i \xi_i + \beta_i \xi_i') \\
&\quad + \sum_{i=1}^m \lambda_i (w^\top x_i + b - y_i - \epsilon - \xi_i) + \sum_{i=1}^m \mu_i (-w^\top x_i - b + y_i - \epsilon - \xi_i').
\end{aligned}
$$

The Lagrangian can also be written as

$$
\begin{aligned}
L(w, b, \lambda, \mu, \gamma, \xi, \xi', \epsilon, \alpha, \beta) &= \frac{1}{2} w^\top w + w^\top \left(\sum_{i=1}^m (\lambda_i - \mu_i) x_i\right) + \epsilon \left(C\nu - \gamma - \sum_{i=1}^m (\lambda_i + \mu_i)\right) \\
&+ \sum_{i=1}^m \xi_i \left(\frac{C}{m} - \lambda_i - \alpha_i\right) + \sum_{i=1}^m \xi_i' \left(\frac{C}{m} - \mu_i - \beta_i\right) + b\left(\sum_{i=1}^m (\lambda_i - \mu_i)\right) - \sum_{i=1}^m (\lambda_i - \mu_i) y_i.
\end{aligned}
$$

To find the dual function $G(\lambda, \mu, \gamma, \alpha, \beta)$, we minimize $L(w, b, \lambda, \mu, \gamma, \xi, \xi', \epsilon, \alpha, \beta)$ with respect to the primal variables w, ϵ, b, ξ and ξ'. Observe that the Lagrangian is convex, and since $(w, \epsilon, \xi, \xi', b) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$, a convex open set, by Theorem 40.13, the Lagrangian has a minimum iff $\nabla L_{w,\epsilon,b,\xi,\xi'} = 0$, so we compute the gradient $\nabla L_{w,\epsilon,b,\xi,\xi'}$. We obtain

$$
\nabla L_{w,\epsilon,b,\xi,\xi'} = \begin{pmatrix} w + \sum_{i=1}^m (\lambda_i - \mu_i) x_i \\ C\nu - \gamma - \sum_{i=1}^m (\lambda_i + \mu_i) \\ \sum_{i=1}^m (\lambda_i - \mu_i) \\ \frac{C}{m} - \lambda - \alpha \\ \frac{C}{m} - \mu - \beta \end{pmatrix},
$$

where

$$
\left(\frac{C}{m} - \lambda - \alpha\right)_i = \frac{C}{m} - \lambda_i - \alpha_i, \quad \text{and} \quad \left(\frac{C}{m} - \mu - \beta\right)_i = \frac{C}{m} - \mu_i - \beta_i.
$$

Consequently, if we set $\nabla L_{w,\epsilon,b,\xi,\xi'} = 0$, we obtain the equations

$$\begin{aligned}
w &= \sum_{i=1}^m (\mu_i - \lambda_i) x_i = X^\top (\mu - \lambda), \qquad (*_w)\\
C\nu - \gamma - \sum_{i=1}^m (\lambda_i + \mu_i) &= 0\\
\sum_{i=1}^m (\lambda_i - \mu_i) &= 0\\
\frac{C}{m} - \lambda - \alpha = 0, &\quad \frac{C}{m} - \mu - \beta = 0.
\end{aligned}$$

Substituting the above equations in the second expression for the Lagrangian, we find that the dual function G is independent of the variables γ, α, β and is given by

$$G(\lambda, \mu) = -\frac{1}{2} \sum_{i,j=1}^m (\lambda_i - \mu_i)(\lambda_j - \mu_j) x_i^\top x_j - \sum_{i=1}^m (\lambda_i - \mu_i) y_i$$

if

$$\begin{aligned}
\sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i &= 0\\
\sum_{i=1}^m \lambda_i + \sum_{i=1}^m \mu_i + \gamma &= C\nu\\
\lambda + \alpha = \frac{C}{m}, &\quad \mu + \beta = \frac{C}{m},
\end{aligned}$$

and $-\infty$ otherwise.

The dual program is obtained by maximizing $G(\alpha, \mu)$ or equivalently by minimizing $-G(\alpha, \mu)$, over $\alpha, \mu \in \mathbb{R}_+^m$. Taking into account the fact that $\alpha, \beta \geq 0$ and $\gamma \geq 0$, we obtain the following dual program:

Dual Program for ν-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2} \sum_{i,j=1}^m (\lambda_i - \mu_i)(\lambda_j - \mu_j) x_i^\top x_j + \sum_{i=1}^m (\lambda_i - \mu_i) y_i\\
&\text{subject to}\\
&\qquad \sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i = 0\\
&\qquad \sum_{i=1}^m \lambda_i + \sum_{i=1}^m \mu_i \leq C\nu\\
&\qquad 0 \leq \lambda_i \leq \frac{C}{m}, \quad 0 \leq \mu_i \leq \frac{C}{m}, \quad i = 1, \ldots, m,
\end{aligned}$$

minimizing over α and μ.

Solving the dual program (for example, using ADMM, see Section 56.3) does not determine b, and for this we use the KKT conditions. The KKT conditions (for the primal program) are

$$\begin{aligned}
\lambda_i(w^\top x_i + b - y_i - \epsilon - \xi_i) &= 0, \quad i = 1, \ldots, m \\
\mu_i(-w^\top x_i - b + y_i - \epsilon - \xi_i') &= 0, \quad i = 1, \ldots, m \\
\gamma\epsilon &= 0 \\
\alpha_i \xi_i &= 0, \quad i = 1, \ldots, m \\
\beta_i \xi_i' &= 0, \quad i = 1, \ldots, m.
\end{aligned}$$

If $\epsilon > 0$, since the equations

$$\begin{aligned}
w^\top x_i + b - y_i &= \epsilon + \xi_i \\
-w^\top x_i - b + y_i &= \epsilon + \xi_i'
\end{aligned}$$

cannot hold simultaneously, we must have

$$\lambda_i \mu_i = 0, \quad i = 1, \ldots, m. \tag{$\lambda\mu$}$$

From the equations

$$\lambda_i + \alpha_i = \frac{C}{m}, \quad \mu_i + \beta_i = \frac{C}{m}, \quad \alpha_i \xi_i = 0, \quad \beta_i \xi_i' = 0,$$

we get the equations

$$\left(\frac{C}{m} - \lambda_i\right)\xi_i = 0, \quad \left(\frac{C}{m} - \mu_i\right)\xi_i' = 0, \quad i = 1, \ldots, m. \tag{$*$}$$

Suppose we have optimal solution with $\epsilon > 0$. Using the above equations and the fact that $\lambda_i \mu_i = 0$ we obtain the following classification of the points x_i in terms of λ and μ.

(1) $0 < \lambda_i < C/m$. By $(*)$, $\xi_i = 0$, so the equation $w^\top x_i + b - y_i = \epsilon$ holds and x_i is on the blue margin hyperplane $H_{w,b-\epsilon}$. See Figure 56.5.

(2) $0 < \mu_i < C/m$. By $(*)$, $\xi_i' = 0$, so the equation $-w^\top x_i - b + y_i = \epsilon$ holds and x_i is on the red margin hyperplane $H_{w,b+\epsilon}$. See Figure 56.5.

(3) $\lambda_i = C/m$. By $(\lambda\mu)$, $\mu_i = 0$, and by $(*)$, $\xi_i' = 0$. Thus we have

$$\begin{aligned}
w^\top x_i + b - y_i &= \epsilon + \xi_i \\
-w^\top x_i - b + y_i &\leq \epsilon.
\end{aligned}$$

The second inequality is equivalent to $-\epsilon \leq w^\top x_i + b - y_i$, and since $\epsilon > 0$ and $\xi_i \geq 0$ it is trivially satisfied. If $\xi_i = 0$, then x_i is on the blue margin $H_{w,b-\epsilon}$, else x_i is an error and it lies in the open half-space bounded by the blue margin $H_{w,b-\epsilon}$ and not containing the best fit hyperplane $H_{w,b}$ (it is outside of the ϵ-slab). See Figure 56.5.

(4) $\mu_i = C/m$. By $(\lambda\mu)$, $\lambda_i = 0$, and by $(*)$, $\xi_i = 0$. Thus we have

$$\begin{aligned} w^\top x_i + b - y_i &\leq \epsilon \\ -w^\top x_i - b + y_i &= \epsilon + \xi'_i. \end{aligned}$$

The second equation is equivalent to $w^\top x_i + b - y_i = -\epsilon - \xi'_i$, and since $\epsilon > 0$ and $\xi'_i \geq 0$, the first inequality it is trivially satisfied. If $\xi'_i = 0$, then x_i is on the red margin $H_{w,b+\epsilon}$, else x_i is an error and it lies in the open half-space bounded by the red margin $H_{w,b-\epsilon}$ and not containing the best fit hyperplane $H_{w,b}$ (it is outside of the ϵ-slab). See Figure 56.5.

(5) $\lambda_i = 0$ and $\mu_i = 0$. By $(*)$, $\xi_i = 0$ and $\xi'_i = 0$, so we have

$$\begin{aligned} w^\top x_i + b - y_i &\leq \epsilon \\ -w^\top x_i - b + y_i &\leq \epsilon, \end{aligned}$$

that is

$$-\epsilon \leq w^\top x_i + b - y_i \leq \epsilon.$$

If $w^\top x_i + b - y_i = \epsilon$, then x_i is on the blue margin, and if $w^\top x_i + b - y_i = -\epsilon$, then x_i is on the red margin. If $-\epsilon < w^\top x_i + b - y_i < \epsilon$, then x_i is strictly inside of the ϵ-slab (bounded by the blue margin and the red margin). See Figure 56.6.

The above classification shows that the point x_i is an error iff $\lambda_i = C/m$ and $\xi_i > 0$ or or $\mu_i = C/m$ and $\xi'_i > 0$.

As in the case of SVM (see Section 50.6) we define support vectors as follows.

Definition 56.1. A vector x_i such that either $w^\top x_i + b - y_i = \epsilon$ (which implies $\xi_i = 0$) or $-w^\top x_i - b + y_i = \epsilon$ (which implies $\xi'_i = 0$) is called a *support vector*. Support vectors x_i such that $0 < \lambda_i < C/m$ and support vectors x_j such that $0 < \mu_j < C/m$ are *support vectors of type 1*. Support vectors of type 1 play a special role so we denote the sets of indices associated with them by

$$\begin{aligned} I_\lambda &= \{i \in \{1, \ldots, m\} \mid 0 < \lambda_i < C/m\} \\ I_\mu &= \{j \in \{1, \ldots, m\} \mid 0 < \mu_j < C/m\}. \end{aligned}$$

We denote their cardinalities by $numsvl_1 = |I_\lambda|$ and $numsvm_1 = |I_\mu|$. Support vectors x_i such that $\lambda_i = C/m$ and support vectors x_j such that $\mu_j = C/m$ are *support vectors of type 2*. Support vectors for which $\lambda_i = \mu_i = 0$ are called *exceptional support vectors*.

The following definition also gives a useful classification criterion.

Definition 56.2. A point x_i such that either $\lambda_i = C/m$ or $\mu_i = C/m$ is said to *fail the margin*. The sets of indices associated with the vectors failing the margin are denoted by

$$\begin{aligned} K_\lambda &= \{i \in \{1, \ldots, m\} \mid \lambda_i = C/m\} \\ K_\mu &= \{j \in \{1, \ldots, m\} \mid \mu_j = C/m\}. \end{aligned}$$

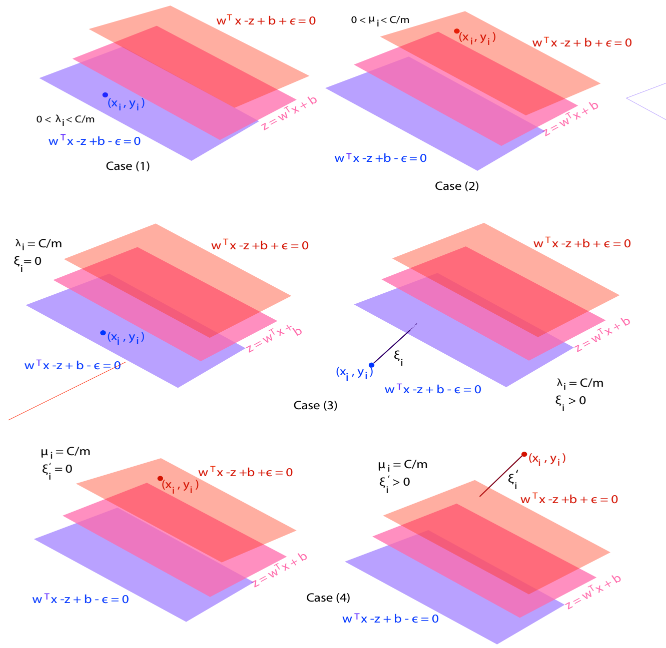


Figure 56.5: Classifying x_i in terms of nonzero λ and μ.

We denote their cardinalities by $p_f = |K_\lambda|$ and $q_f = |K_\mu|$.

Vectors u_i such that $\lambda_i > 0$ and vectors v_j such that $\mu_j > 0$ are said to *have margin at most* ϵ. A point x_i such that either $\lambda_i > 0$ or $\mu_i > 0$ is said to have *margin at most* ϵ. The sets of indices associated with these vectors are denoted by

$$\begin{aligned} I_{\lambda>0} &= \{i \in \{1, \ldots, m\} \mid \lambda_i > 0\} \\ I_{\mu>0} &= \{j \in \{1, \ldots, m\} \mid \mu_j > 0\}. \end{aligned}$$

We denote their cardinalities by $p_m = |I_{\lambda>0}|$ and $q_m = |I_{\mu>0}|$.

Points that fail the margin and are not on the boundary of the ϵ-slab lie outside the closed ϵ-slab, so they are errors, also called *outliers*; they correspond to $\xi_i > 0$ or $\xi_i' > 0$.

Observe that we have the equations $I_\lambda \cup K_\lambda = I_{\lambda>0}$ and $I_\mu \cup K_\mu = I_{\mu>0}$, and the inequalities $p_f \leq p_m$ and $q_f \leq q_m$.

We also have the following results showing that p_f, q_f, p_m and q_m have a direct influence on the choice of ν.

Proposition 56.2. *(1) Let p_f be the number of points x_i such that $\lambda_i = C/m$, and let q_f be the number of points x_i such that $\mu_i = C/m$. We have $p_f, q_f \leq (m\nu)/2$.*

(2) Let p_m be the number of points x_i such that $\lambda_i > 0$, and let q_m be the number of points x_i such that $\mu_i > 0$. We have $p_m, q_m \geq (m\nu)/2$.

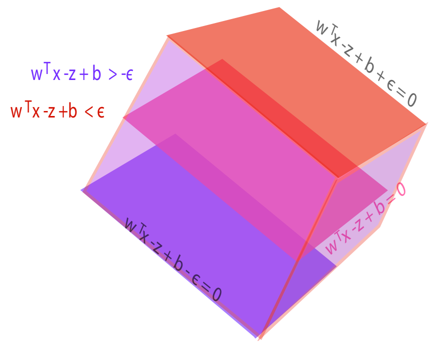


Figure 56.6: The closed ϵ- tube associated with zero multiplier classification, namely $\lambda_i = 0$ and $\mu_i = 0$.

(3) If $p_f \geq 1$ or $q_f \geq 1$, then $\nu \geq 2/m$.

Proof. (1) Recall that for an optimal solution with $w \neq 0$ and $\epsilon > 0$ we have $\gamma = 0$, so we have the equations

$$\sum_{i=1}^{m} \lambda_i = \frac{C\nu}{2} \quad \text{and} \quad \sum_{j=1}^{m} \mu_j = \frac{C\nu}{2}.$$

If there are p_f points such that $\lambda_i = C/m$, then

$$\frac{C\nu}{2} = \sum_{i=1}^{m} \lambda_i \geq p_f \frac{C}{m},$$

so

$$p_f \leq \frac{m\nu}{2}.$$

A similar reasoning applies if there are q_f points such that $\mu_i = C/m$, and we get

$$q_f \leq \frac{m\nu}{2}.$$

(2) If $I_{\lambda>0} = \{i \in \{1, \ldots, m\} \mid \lambda_i > 0\}$ and $p_m = |I_{\lambda>0}|$, then

$$\frac{C\nu}{2} = \sum_{i=1}^{m} \lambda_i = \sum_{i \in I_{\lambda>0}} \lambda_i,$$

and since $\lambda_i \leq C/m$, we have

$$\frac{C\nu}{2} = \sum_{i \in I_{\lambda>0}} \lambda_i \leq p_m \frac{C}{m},$$

which yields
$$p_m \geq \frac{\nu m}{2}.$$
A similar reasoning applies if $\mu_i > 0$.

(3) This follows immediately from (1). $\square$

Proposition 56.2 yields bounds on ν, namely
$$\max\left\{\frac{2p_f}{m}, \frac{2q_f}{m}\right\} \leq \nu \leq \min\left\{\frac{2p_m}{m}, \frac{2q_m}{m}\right\},$$
with $p_f \leq p_m$, $q_f \leq q_m$, $p_f + q_f \leq m$ and $p_m + q_m \leq m$. Also, $p_f = q_f = 0$ means that the ϵ-slab is wide enough so that there are no errors (no points strictly outside the slab).

Observe that a small value of ν keeps p_f and q_f small, which is achieved if the ϵ-slab is wide. A large value of ν allows p_m and q_m to be fairly large, which is achieved if the ϵ-slab is narrow. Thus the smaller ν is, the wider the ϵ-slab is, and the larger ν is, the narrower the ϵ-slab is.

56.2 Existence of Support Vectors

We now consider the issue of the existence of support vectors. We will show that in the generic case, for any optimal solution for which $\epsilon > 0$, there is some support vector on the blue margin *and* some support vector on the red margin. Here generic means that there is an optimal solution for some $\nu < (m-1)/m$.

If the data set (X, y) is well fit by some affine function $f(x) = w^\top x + b$, in the sense that for many pairs (x_i, y_i) we have $y_i = w^\top x_i + b$ and the ℓ^1-error
$$\sum_{i=1}^{m} |w^\top x_i + b - y_i|$$
is small, then an optimal solution may have $\epsilon = 0$. Geometrically, many points (x_i, y_i) belong to the hyperplane $H_{w,b}$. The situation in which $\epsilon = 0$ corresponds to minimizing the ℓ^1-error with a quadratic penalization of w. This is a sort of dual of lasso. The fact that the affine function $f(x) = w^\top x + b$ fits perfectly many points corresponds to the fact that an ℓ^1-minimization tends to encourage sparsity. In this case, if C is chosen too small, it is possible that all points are errors (although "small") and there are no support vectors. But if C is large enough, the solution will be sparse and there will be *many* support vectors on the hyperplane $H_{w,b}$.

Let $E_\lambda = \{i \in \{1, \ldots, m\} \mid \xi_i > 0\}$, $E_\mu = \{j \in \{1, \ldots, m\} \mid \xi'_j > 0\}$, $p_{sf} = |E_\lambda|$ and $q_{sf} = |E_\mu|$. Obviously, E_λ and E_μ are disjoint.

Given any real numbers u, v, x, y, if $\max\{u, v\} < \min\{x, y\}$, then $u < x$ and $v < y$. This is because $u, v \leq \max\{u, v\} < \min\{x, y\} \leq x, y$.

Proposition 56.3. *If $\nu < (m-1)/m$, then $p_f < \lfloor m/2 \rfloor$ and $q_f < \lfloor m/2 \rfloor$.*

Proof. By Proposition 56.2, $\max\{2p_f/m, 2q_f/m\} \leq \nu$. If m is even, say $m = 2k$, then

$$2p_f/m = 2p_f/(2k) \leq \nu < (m-1)/m = (2k-1)/2k,$$

so $2p_f < 2k-1$, which implies $p_f < k = \lfloor m/2 \rfloor$. A similar argument shows that $q_f < k = \lfloor m/2 \rfloor$.

If m is odd, say $m = 2k+1$, then

$$2p_f/m = 2p_f/(2k+1) \leq \nu < (m-1)/m = 2k/(2k+1),$$

so $2p_f < 2k$, which implies $p_f < k = \lfloor m/2 \rfloor$. A similar argument shows that $q_f < k = \lfloor m/2 \rfloor$. ∎

Since $p_{sf} \leq p_f$ and $q_{sf} \leq q_f$, we also have $p_{sf} < \lfloor m/2 \rfloor$ and $q_{sf} < \lfloor m/2 \rfloor$. This implies that $\{1, \ldots, m\} - (E_\lambda \cup E_\mu)$ contains at least two elements and there are constraints corresponding to at least two $i \notin (E_\lambda \cup E_\mu)$ (in which case $\xi_i = \xi'_i = 0$), of the form

$$\begin{aligned} w^\top x_i + b - y_i &\leq \epsilon & \quad & i \notin (E_\lambda \cup E_\mu) \\ -w^\top x_i - b + y_i &\leq \epsilon & & i \notin (E_\lambda \cup E_\mu). \end{aligned}$$

If $w^\top x_i + b - y_i = \epsilon$ for some $i \notin (E_\lambda \cup E_\mu)$ and $-w^\top x_j - b + y_j = \epsilon$ for some $j \notin (E_\lambda \cup E_\mu)$ with $i \neq j$, then we have a blue support vector and a red support vector. Otherwise, we show how to modify b and ϵ to obtain an optimal solution with a blue support vector and a red support vector.

Proposition 56.4. *For every optimal solution $(w, b, \epsilon, \xi, \xi')$ with $w \neq 0$ and $\epsilon > 0$, if*

$$\nu < (m-1)/m$$

and if either no x_i is a blue support vector or no x_i is a red support vector, then there is another optimal solution (for the same w) with some i_0 such that $\xi_{i_0} = 0$ and $w^\top x_{i_0} + b - y_{i_0} = \epsilon$, and there is some j_0 such that $\xi'_{j_0} = 0$ and $-w^\top x_{j_0} - b + y_{j_0} = \epsilon$; in other words, some x_{i_0} is a blue support vector and some x_{j_0} is a red support vector (with $i_0 \neq j_0$). If all points (x_i, y_i) that are not errors lie on one of the margin hyperplanes, then there is an optimal solution for which $\epsilon = 0$.

Proof. By Proposition 56.3 if $\nu < (m-1)/m$, then $p_f < \lfloor m/2 \rfloor$ and $q_f < \lfloor m/2 \rfloor$, so the following constraints hold:

$$\begin{aligned} w^\top x_i + b - y_i &= \epsilon + \xi_i & \quad \xi_i &> 0 & \quad & i \in E_\lambda \\ -w^\top x_j - b + y_j &= \epsilon + \xi'_j & \xi'_j &> 0 & & j \in E_\mu \\ w^\top x_i + b - y_i &\leq \epsilon & & & & i \notin (E_\lambda \cup E_\mu) \\ -w^\top x_i - b + y_i &\leq \epsilon & & & & i \notin (E_\lambda \cup E_\mu), \end{aligned}$$

where $|\{1, \ldots, m\} - (E_\lambda \cup E_\mu)| \geq 2$.

If our optimal solution does not have a blue support vector and a red support vector, then either $w^\top x_i + b - y_i < \epsilon$ for all $i \notin (E_\lambda \cup E_\mu)$ or $-w^\top x_i - b + y_i < \epsilon$ for all $i \notin (E_\lambda \cup E_\mu)$.

Case 1. We have

$$\begin{aligned} w^\top x_i + b - y_i &< \epsilon & & i \notin (E_\lambda \cup E_\mu) \\ -w^\top x_i - b + y_i &\leq \epsilon & & i \notin (E_\lambda \cup E_\mu). \end{aligned}$$

There are two subcases.

Case 1a. Assume that there is some $j \notin (E_\lambda \cup E_\mu)$ such that $-w^\top x_j - b + y_j = \epsilon$. Our strategy is to decrease ϵ and increase b by a small amount θ in such a way that some inequality $w^\top x_i + b - y_i < \epsilon$ becomes an equation for some $i \notin (E_\lambda \cup E_\mu)$. Geometrically, this amounts to raising the separating hyperplane $H_{w,b}$ and decreasing the width of the slab, keeping the red margin hyperplane unchanged. See Figure 56.7.

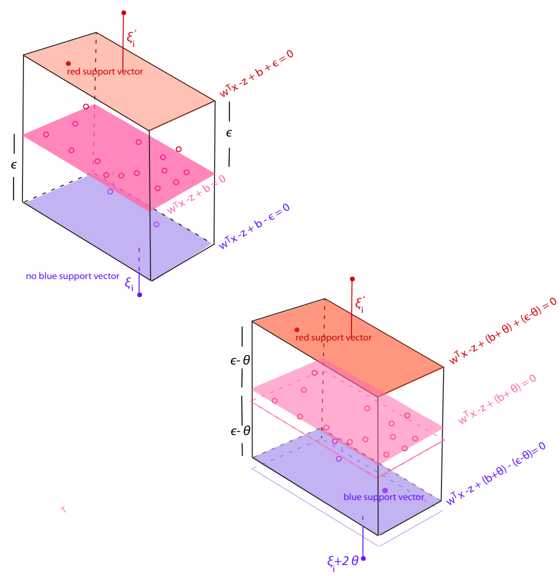


Figure 56.7: In this illustration points within the ϵ-tube are denoted by open circles. In the original, upper left configuration, there is no blue support vector. By raising the pink separating hyperplane and decreasing the width of the slab, we end up with a blue support vector.

The inequalities imply that

$$-\epsilon \leq w^\top x_i + b - y_i < \epsilon.$$

Let us pick θ such that

$$\theta = (1/2) \min\{\epsilon - w^\top x_i - b + y_i \mid i \notin (E_\lambda \cup E_\mu)\}.$$

Our hypotheses imply that $\theta > 0$, and we have $\theta \leq \epsilon$, because $(1/2)(\epsilon - w^\top x_i - b + y_i) \leq \epsilon$ is equivalent to $\epsilon - w^\top x_i - b + y_i \leq 2\epsilon$ which is equivalent to $-w^\top x_i - b + y_i \leq \epsilon$, which holds for all $i \notin (E_\lambda \cup E_\mu)$ by hypothesis.

We can write

$$\begin{aligned}
w^\top x_i + b + \theta - y_i &= \epsilon - \theta + \xi_i + 2\theta & \xi_i &> 0 & i &\in E_\lambda \\
-w^\top x_j - (b + \theta) + y_j &= \epsilon - \theta + \xi'_j & \xi'_j &> 0 & j &\in E_\mu \\
w^\top x_i + b + \theta - y_i &\leq \epsilon - \theta & & & i &\notin (E_\lambda \cup E_\mu) \\
-w^\top x_i - (b + \theta) + y_i &\leq \epsilon - \theta & & & i &\notin (E_\lambda \cup E_\mu).
\end{aligned}$$

By hypothesis

$$-w^\top x_j - (b + \theta) + y_j = \epsilon - \theta \quad \text{for some } j \notin (E_\lambda \cup E_\mu)$$

and by the choice of θ,

$$w^\top x_i + b + \theta - y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu).$$

The value of $C > 0$ is irrelevant in the following argument so we may assume that $C = 1$. The new value of the objective function is

$$\begin{aligned}
\omega(\theta) &= \frac{1}{2} w^\top w + \nu(\epsilon - \theta) + \frac{1}{m}\left(\sum_{i \in E_\lambda} (\xi_i + 2\theta) + \sum_{j \in E_\mu} \xi'_j\right) \\
&= \frac{1}{2} w^\top w + \nu\epsilon + \frac{1}{m}\left(\sum_{i \in E_\lambda} \xi_i + \sum_{j \in E_\mu} \xi'_j\right) - \left(\nu - \frac{2p_{sf}}{m}\right)\theta.
\end{aligned}$$

By Proposition 56.2 we have

$$\max\left\{\frac{2p_f}{m}, \frac{2q_f}{m}\right\} \leq \nu$$

and $p_{sf} \leq p_f$ and $q_{sf} \leq q_f$, which implies that

$$\nu - \frac{2p_{sf}}{m} \geq 0, \tag{$*_1$}$$

and so $\omega(\theta) \leq \omega(0)$. If inequality $(*_1)$ is strict, then this contradicts the optimality of the original solution. Therefore, $\nu = 2p_{sf}/m$, $\omega(\theta) = \omega(0)$ and $(w, b + \theta, \epsilon - \theta, \xi + 2\theta, \xi')$ is an optimal solution such that

$$\begin{aligned}
w^\top x_i + b + \theta - y_i &= \epsilon - \theta \\
-w^\top x_j - (b + \theta) + y_j &= \epsilon - \theta
\end{aligned}$$

for some $i, j \notin (E_\lambda \cup E_\mu)$ with $i \neq j$.

Observe that the exceptional case in which $\theta = \epsilon$ may arise. In this case all points (x_i, y_i) that are not errors (strictly outside the ϵ-slab) are on the red margin hyperplane. This case can only arise if $\nu = 2p_{sf}/m$.

Case 1b. We have $-w^\top x_i - b + y_i < \epsilon$ for all $i \notin (E_\lambda \cup E_\mu)$. Our strategy is to decrease ϵ and increase the errors by a small θ in such a way that some inequality becomes an equation for some $i \notin (E_\lambda \cup E_\mu)$. Geometrically, this corresponds to decreasing the width of the slab, keeping the separating hyperplane unchanged. See Figures 56.8 and 56.9. Then we are reduced to *Case 1a* or *Case 2a*.

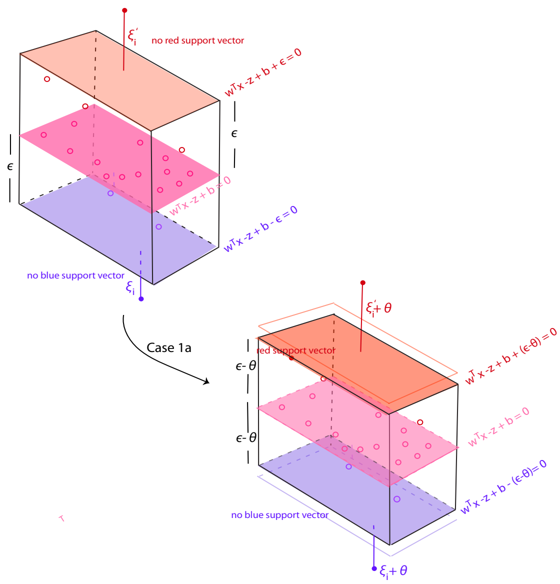


Figure 56.8: In this illustration points within the ϵ-tube are denoted by open circles. In the original, upper left configuration, there is no blue support vector and no red support vector. By decreasing the width of the slab, we end up with a red support vector and reduce to Case 1a.

We have

$$
\begin{aligned}
w^\top x_i + b - y_i &= \epsilon + \xi_i & \xi_i &> 0 & & i \in E_\lambda \\
-w^\top x_j - b + y_j &= \epsilon + \xi_j' & \xi_j' &> 0 & & j \in E_\mu \\
w^\top x_i + b - y_i &< \epsilon & & & & i \notin (E_\lambda \cup E_\mu) \\
-w^\top x_i - b + y_i &< \epsilon & & & & i \notin (E_\lambda \cup E_\mu).
\end{aligned}
$$

Let us pick θ such that

$$
\theta = \min\{\epsilon - (w^\top x_i + b - y_i),\ \epsilon + w^\top x_i + b - y_i \mid i \notin (E_\lambda \cup E_\mu)\},
$$

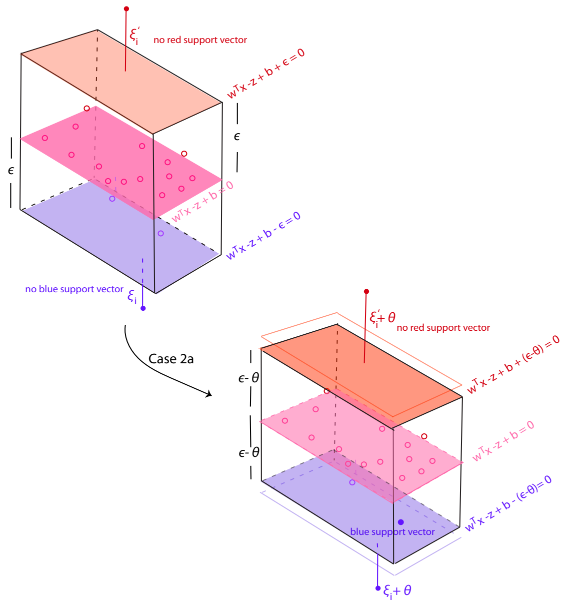


Figure 56.9: In this illustration points within ϵ-tube are denoted by open circles. In the original, upper left configuration, there is no blue support vector and no red support vector. By decreasing the width of the slab, we end up with a blue support vector and reduce to Case 2a.

Our hypotheses imply that $0 < \theta < \epsilon$. We can write

$$\begin{aligned}
w^\top x_i + b - y_i &= \epsilon - \theta + \xi_i + \theta & \xi_i &> 0 & i &\in E_\lambda \\
-w^\top x_j - b + y_j &= \epsilon - \theta + \xi_j' + \theta & \xi_j' &> 0 & j &\in E_\mu \\
w^\top x_i + b - y_i &\leq \epsilon - \theta & & & i &\notin (E_\lambda \cup E_\mu) \\
-w^\top x_i - b + y_i &\leq \epsilon - \theta & & & i &\notin (E_\lambda \cup E_\mu),
\end{aligned}$$

and by the choice of θ, either

$$w^\top x_i + b - y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu)$$

or

$$-w^\top x_i - b + y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu).$$

The new value of the objective function is

$$\begin{aligned}
\omega(\theta) &= \frac{1}{2} w^\top w + \nu(\epsilon - \theta) + \frac{1}{m}\left(\sum_{i \in E_\lambda} (\xi_i + \theta) + \sum_{j \in E_\mu} (\xi_j' + \theta)\right) \\
&= \frac{1}{2} w^\top w + \nu\epsilon + \frac{1}{m}\left(\sum_{i \in E_\lambda} \xi_i + \sum_{j \in E_\mu} \xi_j'\right) - \left(\nu - \frac{p_{sf} + q_{sf}}{m}\right)\theta.
\end{aligned}$$

Since $\max\{2p_f/m, 2q_f/m\} \leq \nu$ implies that $(p_f + q_f)/m \leq \nu$ and $p_{sf} \leq p_f$, $q_{sf} \leq q_f$, we have

$$\nu - \frac{p_{sf} + q_{sf}}{m} \geq 0, \tag{$*_2$}$$

and so $\omega(\theta) \leq \omega(0)$. If inequality ($*_2$) is strict, then this contradicts the optimality of the original solution. Therefore, $\nu = (p_{sf} + q_{sf})/m$, $\omega(\theta) = \omega(0)$ and $(w, b, \epsilon - \theta, \xi + \theta, \xi' + \theta)$ is an optimal solution such that either

$$w^\top x_i + b - y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu)$$

or

$$-w^\top x_i - b + y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu).$$

We are now reduced to *Case 1a* or or *Case 2a*.

Case 2. We have

$$\begin{aligned} w^\top x_i + b - y_i &\leq \epsilon & & i \notin (E_\lambda \cup E_\mu) \\ -w^\top x_i - b + y_i &< \epsilon & & i \notin (E_\lambda \cup E_\mu). \end{aligned}$$

Again there are two subcases.

Case 2a. Assume that there is some $i \notin (E_\lambda \cup E_\mu)$ such that $w^\top x_i + b - y_i = \epsilon$. Our strategy is to decrease ϵ and decrease b by a small amount θ in such a way that some inequality $-w^\top x_j - b + y_j < \epsilon$ becomes an equation for some $j \notin (E_\lambda \cup E_\mu)$. Geometrically, this amounts to lowering the separating hyperplane $H_{w,b}$ and decreasing the width of the slab, keeping the blue margin hyperplane unchanged. See Figure 56.10.

The inequalities imply that

$$-\epsilon < w^\top x_i + b - y_i \leq \epsilon.$$

Let us pick θ such that

$$\theta = (1/2) \min\{\epsilon - (-w^\top x_i - b + y_i) \mid i \notin (E_\lambda \cup E_\mu)\}.$$

Our hypotheses imply that $\theta > 0$, and we have $\theta \leq \epsilon$, because $(1/2)(\epsilon - (-w^\top x_i - b + y_i)) \leq \epsilon$ is equivalent to $\epsilon - (-w^\top x_i - b + y_i) \leq 2\epsilon$ which is equivalent to $w^\top x_i + b - y_i \leq \epsilon$ which holds for all $i \notin (E_\lambda \cup E_\mu)$ by hypothesis.

We can write

$$\begin{aligned} w^\top x_i + b - \theta - y_i &= \epsilon - \theta + \xi_i & \xi_i &> 0 & & i \in E_\lambda \\ -w^\top x_j - (b - \theta) + y_j &= \epsilon - \theta + \xi'_j + 2\theta & \xi'_j &> 0 & & j \in E_\mu \\ w^\top x_i + b - \theta - y_i &\leq \epsilon - \theta & & & & i \notin (E_\lambda \cup E_\mu) \\ -w^\top x_i - (b - \theta) + y_i &\leq \epsilon - \theta & & & & i \notin (E_\lambda \cup E_\mu). \end{aligned}$$

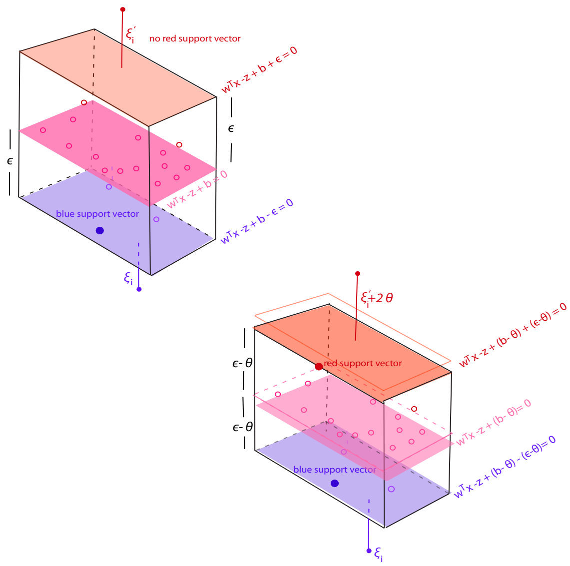


Figure 56.10: In this illustration points within the ϵ-tube are denoted by open circles. In the original, upper left configuration, there is no red support vector. By lowering the pink separating hyperplane and decreasing the width of the slab, we end up with a red support vector.

By hypothesis

$$w^\top x_i + (b - \theta) - y_i = \epsilon - \theta \quad \text{for some } i \notin (E_\lambda \cup E_\mu),$$

and by the choice of θ,

$$-w^\top x_j - (b - \theta) + y_j = \epsilon - \theta \quad \text{for some } j \notin (E_\lambda \cup E_\mu).$$

The new value of the objective function is

$$\begin{aligned}
\omega(\theta) &= \frac{1}{2} w^\top w + \nu(\epsilon - \theta) + \frac{1}{m}\left(\sum_{i \in E_\lambda} \xi_i + \sum_{j \in E_\mu} (\xi'_j + 2\theta)\right) \\
&= \frac{1}{2} w^\top w + \nu\epsilon + \frac{1}{m}\left(\sum_{i \in E_\lambda} \xi_i + \sum_{j \in E_\mu} \xi'_j\right) - \left(\nu - \frac{2q_{sf}}{m}\right)\theta.
\end{aligned}$$

The rest of the proof is similar except that $2p_{sf}/m$ is replaced by $2q_{sf}/m$. Observe that the exceptional case in which $\theta = \epsilon$ may arise. In this case all points (x_i, y_i) that are not errors (strictly outside the ϵ-slab) are on the blue margin hyperplane. This case can only arise if $\nu = 2q_{sf}/m$.

Case 2b. We have $w^\top x_i + b - y_i < \epsilon$ for all $i \notin (E_\lambda \cup E_\mu)$. Since we also assumed that $-w^\top x_i - b + y_i < \epsilon$ for all $i \notin (E_\lambda \cup E_\mu)$, Case 2b is identical to Case 1b and we are done. $\square$

The proof of Proposition 56.4 reveals that there are three critical values for ν:

$$\frac{2p_{sf}}{m}, \quad \frac{2q_{sf}}{m}, \quad \frac{p_{sf}+q_{sf}}{m}.$$

These values can be avoided by requiring the strict inequality

$$\max\left\{\frac{2p_{sf}}{m}, \frac{2q_{sf}}{m}\right\} < \nu.$$

Then the following corollary holds.

Theorem 56.5. *For every optimal solution $(w, b, \epsilon, \xi, \xi')$ with $w \neq 0$ and $\epsilon > 0$, if*

$$\max\left\{\frac{2p_{sf}}{m}, \frac{2q_{sf}}{m}\right\} < \nu < (m-1)/m,$$

then some x_{i_0} is a blue support vector and some x_{j_0} is a red support vector (with $i_0 \neq j_0$).

Proof. We proceed by contradiction. Suppose that for every optimal solution with $w \neq 0$ and $\epsilon > 0$ no x_i is a blue support vector or no x_i is a red support vector. Since $\nu < (m-1)/m$, Proposition 56.4 holds, so there is another optimal solution. But since the critical values of ν are avoided, the proof of Proposition 56.4 shows that the value of the objective function for this new optimal solution is strictly smaller than the original optimal value, a contradiction. $\square$

Remark: If an optimal solution has $\epsilon = 0$, then depending on the value of C there may not be any support vectors, or many.

If the primal has an optimal solution with $w \neq 0$ and $\epsilon > 0$, then by $(*_w)$ and since

$$\sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i = 0 \quad \text{and} \quad \lambda_i \mu_i = 0,$$

there is i_0 such that $\lambda_{i_0} > 0$ and some $j_0 \neq i_0$ such that $\mu_{j_0} > 0$.

Under the mild hypothesis called the **Standard Margin Hypothesis** that there is some i_0 such that $0 < \alpha_{i_0} < \frac{C}{m}$ *and* there is some $j_0 \neq i_0$ such that $0 < \mu_{j_0} < \frac{C}{m}$, in other words there is a blue support vector of type 1 and there is a red support vector of type 1, then by $(*)$ we have $\xi_{i_0} = 0, \xi'_{j_0} = 0$, and we have the two equations

$$\begin{aligned} w^\top x_{i_0} + b - y_{i_0} &= \epsilon \\ -w^\top x_{j_0} - b + y_{j_0} &= \epsilon, \end{aligned}$$

so b and ϵ can be computed. In particular,

$$\begin{aligned} b &= \frac{1}{2}\left(y_{i_0} + y_{j_0} - w^\top(x_{i_0} + x_{j_0})\right) \\ \epsilon &= \frac{1}{2}\left(y_{j_0} - y_{i_0} + w^\top(x_{i_0} - x_{j_0})\right). \end{aligned}$$

The function $f(x) = w^\top x + b$ (often called *regression estimate*) is given by

$$f(x) = \sum_{i=1}^m (\mu_i - \lambda_i) x_i^\top x + b.$$

In practice, due to numerical inaccurracy, it is complicated to write a computer program that will select two *distinct indices* as above. It is preferable to compute the list I_λ of indices i such that $0 < \lambda_i < C/m$ and the list I_μ of indices j such that $0 < \mu_j < C/m$. Then it is easy to see that

$$\begin{aligned} b &= \left(\left(\sum_{i_0 \in I_\lambda} y_{i_0}\right)/|I_\lambda| + \left(\sum_{j_0 \in I_\mu} y_{j_0}\right)/|I_\mu| - w^\top\left(\left(\sum_{i_0 \in I_\lambda} x_{i_0}\right)/|I_\lambda| + \left(\sum_{j_0 \in I_\mu} x_{j_0}\right)/|I_\mu|\right)\right)/2 \\ \epsilon &= \left(\left(\sum_{j_0 \in I_\mu} y_{j_0}\right)/|I_\mu| - \left(\sum_{i_0 \in I_\lambda} y_{i_0}\right)/|I_\lambda| + w^\top\left(\left(\sum_{i_0 \in I_\lambda} x_{i_0}\right)/|I_\lambda| - \left(\sum_{j_0 \in I_\mu} x_{j_0}\right)/|I_\mu|\right)\right)/2. \end{aligned}$$

These formulae are numerically a lot more stable, but we still have to be cautious to set suitable tolerance factors to decide whether $\lambda_i > 0$ and $\lambda_i < C/m$ (and similarly for μ_i).

The following result gives sufficient conditions for expressing ϵ in terms of a single support vector.

Proposition 56.6. *For every optimal solution $(w, b, \epsilon, \xi, \xi')$ with $w \neq 0$ and $\epsilon > 0$, if*

$$\max\left\{\frac{2p_{sf}}{m}, \frac{2q_{sf}}{m}\right\} < \nu < (m-1)/m,$$

then ϵ and b are determined from a solution (λ, μ) of the dual in terms of a single support vector.

Proof sketch. If we express that the duality gap is zero we obtain the following equation expressing ϵ in terms of b:

$$\begin{aligned} C\left(\nu - \frac{p_f + q_f}{m}\right)\epsilon &= -\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}y^\top & -y^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ &\quad - \frac{C}{m}\left(w^\top\left(\sum_{i \in K_\lambda} x_i - \sum_{j \in K_\mu} x_j\right) - \sum_{i \in K_\lambda} y_i + \sum_{j \in K_\mu} y_j + (p_f - q_f)b\right). \end{aligned}$$

The proof is very similar to the proof of the corresponding formula in Section 56.5. By Theorem 56.5, there is some suppor vector x_i, say

$$w^\top x_{i_0} + b - y_{i_0} = \epsilon \quad \text{or} \quad -w^\top x_{j_0} - b + y_{j_0} = \epsilon.$$

Then we find an equation expressing ϵ in terms of λ, μ and w, provided that $\nu \neq 2p_f/m$ and $\nu \neq 2q_f/m$. The proof is analogous to the proof of Proposition 54.4 and is left as an exercise. $\square$

56.3 Solving ν-Regression Using ADMM

The quadratic functional $F(\lambda, \mu)$ occurring in the dual program given by

$$F(\lambda, \mu) = \frac{1}{2} \sum_{i,j=1}^m (\lambda_i - \mu_i)(\lambda_j - \mu_j) x_i^\top x_j + \sum_{i=1}^m (\lambda_i - \mu_i) y_i$$

is not of the form $\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} P \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + q^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$, but it can be converted in such a form using a trick. First, if we let $\mathbf{K}$ be the $m \times m$ symmetric matrix $\mathbf{K} = XX^\top = (x_i^\top x_j)$, then we have

$$F(\lambda, \mu) = \frac{1}{2}(\lambda^\top - \mu^\top)\mathbf{K}(\lambda - \mu) + y^\top \lambda - y^\top \mu.$$

Consequently, if we define the $2m \times 2m$ symmetric matrix P by

$$P = \begin{pmatrix} XX^\top & -XX^\top \\ -XX^\top & XX^\top \end{pmatrix} = \begin{pmatrix} \mathbf{K} & -\mathbf{K} \\ -\mathbf{K} & \mathbf{K} \end{pmatrix}$$

and the $2m \times 1$ matrix q by

$$q = \begin{pmatrix} y \\ -y \end{pmatrix},$$

it is easy to check that

$$F(\lambda, \mu) = \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} P \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + q^\top \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \frac{1}{2}\lambda^\top \mathbf{K}\lambda + \frac{1}{2}\mu^\top \mathbf{K}\mu - \lambda^\top \mathbf{K}\mu + y^\top \lambda - y^\top \mu. \quad (*_q)$$

Since

$$\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} P \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \frac{1}{2}(\lambda^\top - \mu^\top)\mathbf{K}(\lambda - \mu)$$

and the matrix $\mathbf{K} = XX^\top$ is symmetric positive semidefinite, the matrix P is also symmetric positive semidefinite. Thus we are in a position to apply ADMM since the constraints are

$$\sum_{i=1}^{m} \lambda_i - \sum_{i=1}^{m} \mu_i = 0$$
$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i + \gamma = C\nu$$
$$\lambda + \alpha = \frac{C}{m}, \quad \mu + \beta = \frac{C}{m},$$

namely affine. We need to check that the $(2m+2) \times (4m+1)$ matrix A corresponding to this system has rank $2m+2$. Let us clarify this point. The matrix A corresponding to the above equations is

$$A = \begin{pmatrix} \mathbf{1}_m^\top & -\mathbf{1}_m^\top & 0_m^\top & 0_m^\top & 0 \\ \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top & 1 \\ I_m & 0_{m,m} & I_m & 0_{m,m} & 0_m \\ 0_{m,m} & I_m & 0_{m,m} & I_m & 0_m \end{pmatrix}.$$

For example, for $m = 3$ we have the 8×13 matrix

$$\begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We leave it as an exercise to show that A has rank $2m+2$. Recall that

$$q = \begin{pmatrix} y \\ -y \end{pmatrix}$$

and we also define the vector c (of dimension $2m+2$) as

$$c = \begin{pmatrix} 0 \\ C\nu \\ \frac{C}{m}\mathbf{1}_{2m} \end{pmatrix}.$$

The constraints are given by the system of affine equations $Ax = c$, where

$$x = \begin{pmatrix} \lambda^\top & \mu^\top & \alpha^\top & \beta^\top & \gamma \end{pmatrix}^\top.$$

Since there are $4m+1$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta, \gamma)$, we need to pad the $2m \times 2m$ matrix P with zeros to make it into a $(4m+1) \times (4m+1)$ matrix

$$P_a = \begin{pmatrix} P & 0_{2m,2m+1} \\ 0_{2m+1,2m} & 0_{2m+1,2m+1} \end{pmatrix}.$$

Similarly, we pad q with zeros to make it a vector q_a of dimension $4m+1$,

$$q_a = \begin{pmatrix} q \\ 0_{2m+1} \end{pmatrix}.$$

In order to solve our dual program, we apply ADMM to the quadractic functional

$$\frac{1}{2} x^\top P_a x + q_a^\top x,$$

subject to the constraints

$$Ax = c, \quad x \geq 0,$$

with P_a, q_a, A, b and x, as above.

Since for an optimal solution with $\epsilon > 0$ we must have $\gamma = 0$ (from the KKT conditions), we can solve the dual problem with the following set of constraints only involving the Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$,

$$\begin{aligned} \sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i &= 0 \\ \sum_{i=1}^m \lambda_i + \sum_{i=1}^m \mu_i &= C\nu \\ \lambda + \alpha = \frac{C}{m}, \quad \mu + \beta &= \frac{C}{m}, \end{aligned}$$

which corresponds to the $(2m+2) \times 4m$ A_2 given by

$$A_2 = \begin{pmatrix} \mathbf{1}_m^\top & -\mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ I_m & 0_{m,m} & I_m & 0_{m,m} \\ 0_{m,m} & I_m & 0_{m,m} & I_m \end{pmatrix}.$$

We leave it as an exercise to show that A_2 has rank $2m+2$. We define the vector c_2 (of dimension $2m+2$) as

$$c_2 = c = \begin{pmatrix} 0 \\ C\nu \\ \frac{C}{m} \mathbf{1}_{2m} \end{pmatrix}.$$

Since there are $4m$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$, we need to pad the $2m \times 2m$ matrix P with zeros to make it into a $4m \times 4m$ matrix

$$P_{2a} = \begin{pmatrix} P & 0_{2m,2m} \\ 0_{2m,2m} & 0_{2m,2m} \end{pmatrix}.$$

Similarly, we pad q with zeros to make it a vector q_{2a} of dimension $4m$,

$$q_{2a} = \begin{pmatrix} q \\ 0_{2m} \end{pmatrix}.$$

We implemented the above methods in `Matlab`; see Appendix B, Section B.4. Choosing $C = m$ is typically a good choice because then the values of λ_i and μ_j are not too small ($C/m = 1$). If C is chosen too small, we found that numerical instability increases drastically and very poor results are obtained. Increasing C tends to encourage sparsity.

We ran our `Matlab` implementation of the above method on the set of 50 points generated at random by the program shown below with $C = 50$ and various values of ν starting with $\nu = 0.03$:

```
X13 = 15*randn(50,1);
ww13 = 1;
y13 = X13*ww13 + 10*randn(50,1) + 20;
[~,~,~,~,~,~,~,~,~,w1] = runuregb(rho,0.03,X13,y13,50)
```

Figure 56.11 shows the result of running the program with $\nu = 0.03$. We have $p_f = 0, q_f = 0, p_m = 2$ and $q_m = 1$. There are 47 points strictly inside the slab. The slab is large enough to contain all the data points, so none of them is considered an error.

The next value of ν is $\nu = 0.21$, see Figure 56.12. We have $p_f = 4, q_f = 5, p_m = 6$ and $q_m = 6$. There are 38 points strictly inside the slab.

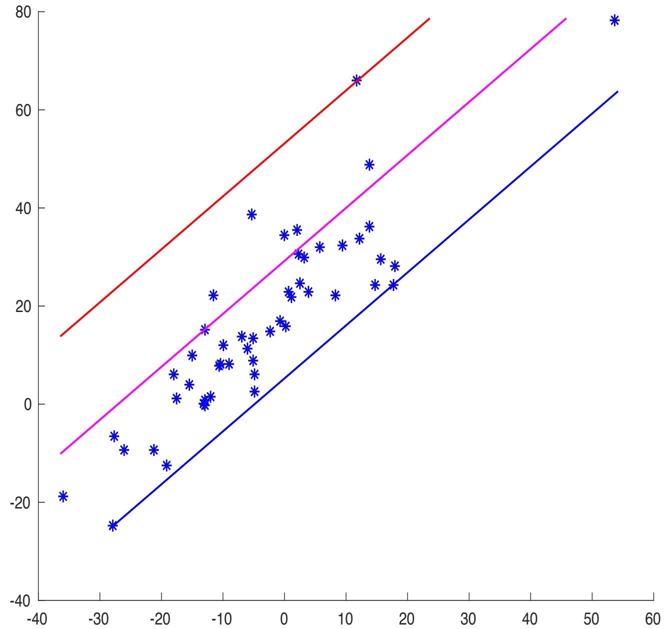


Figure 56.11: Running ν-SV regression on a set of 50 points; $\nu = 0.03$.

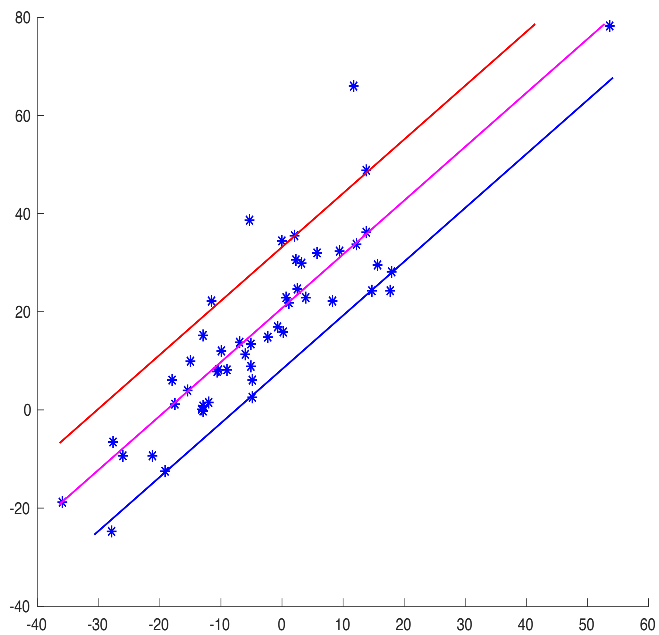


Figure 56.12: Running ν-SV regression on a set of 50 points; $\nu = 0.21$.

The next value of ν is $\nu = 0.5$, see Figure 56.13. We have $p_f = 12, q_f = 12, p_m = 13$ and

$q_m = 14$. There are 23 points strictly inside the slab.

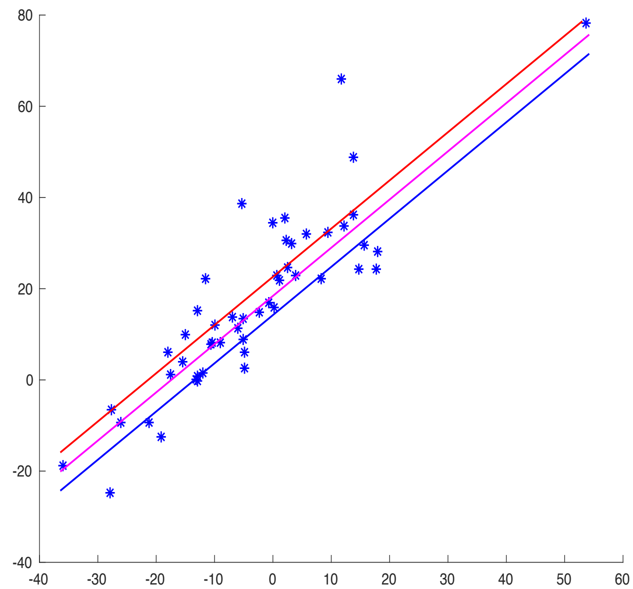


Figure 56.13: Running ν-SV regression on a set of 50 points; $\nu = 0.5$.

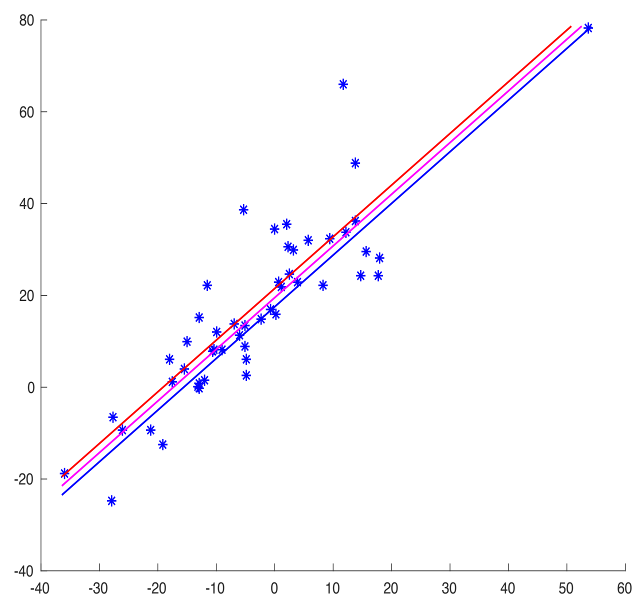


Figure 56.14: Running ν-SV regression on a set of 50 points; $\nu = 0.7$.

The next value of ν is $\nu = 0.7$, see Figure 56.14. We have $p_f = 17, q_f = 17, p_m = 18$ and

$q_m = 19$. There are 13 points strictly inside the slab.

The last value of ν is $\nu = 0.97$, see Figure 56.15. We have $p_f = 23, q_f = 24, p_m = 25$ and $q_m = 25$. There are 0 points strictly inside the slab. The slab is so narrow that it does not contain any of the points x_i in it. Running the program with any value $\nu > 0.97$ yields $\epsilon = 0$.

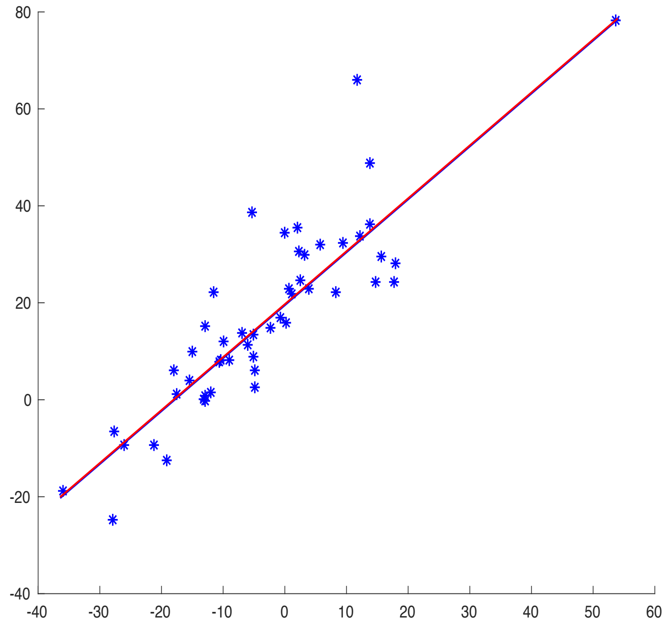


Figure 56.15: Running ν-SV regression on a set of 50 points; $\nu = 0.97$.

56.4 Kernel ν-SV Regression

Since the formulae for w, b, and $f(x)$,

$$\begin{aligned} w &= \sum_{i=1}^{m} (\mu_i - \lambda_i) x_i \\ b &= \frac{1}{2} \left(y_{i_0} + y_{j_0} - w^\top (x_{i_0} + x_{j_0}) \right) \\ f(x) &= \sum_{i=1}^{m} (\mu_i - \lambda_i) x_i^\top x + b, \end{aligned}$$

only involve inner products among the data points x_i and x, and since the objective function $-G(\alpha, \mu)$ of the dual program also only involves inner products among the data points x_i, we can kernelize the ν-SV regression method.

As in the previous section, we assume that our data points $\{x_1, \ldots, x_m\}$ belong to a set $\mathcal{X}$ and we pretend that we have feature space $(F, \langle -, - \rangle)$ and a feature embedding map $\varphi\colon \mathcal{X} \to F$, but we only have access to the kernel function $\kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$. We wish to perform ν-SV regression in the feature space F on the data set $\{(\varphi(x_1), y_1), \ldots, (\varphi(x_m), y_m)\}$. Going over the previous computation, we see that the primal program is given by

Program kernel ν-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\langle w, w\rangle + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^m (\xi_i + \xi_i')\right)\\
&\text{subject to}\\
&\qquad \langle w, \varphi(x_i)\rangle + b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m\\
&\quad - \langle w, \varphi(x_i)\rangle - b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m\\
&\quad \epsilon \geq 0,
\end{aligned}$$

minimizing over the variables w, ϵ, b, ξ, and ξ'.

The Lagrangian is given by

$$\begin{aligned}
L(w, b, \lambda, \mu, \gamma, \xi, \xi', \epsilon, \alpha, \beta) &= \frac{1}{2}\langle w, w\rangle + \left\langle w, \sum_{i=1}^m (\lambda_i - \mu_i)\varphi(x_i)\right\rangle\\
&\quad + \epsilon\left(C\nu - \gamma - \sum_{i=1}^m (\lambda_i + \mu_i)\right) + \sum_{i=1}^m \xi_i\left(\frac{C}{m} - \lambda_i - \alpha_i\right)\\
&\quad + \sum_{i=1}^m \xi_i'\left(\frac{C}{m} - \mu_i - \beta_i\right) + b\left(\sum_{i=1}^m (\lambda_i - \mu_i)\right) - \sum_{i=1}^m (\lambda_i - \mu_i)y_i.
\end{aligned}$$

Setting the gradient $\nabla L_{w,\epsilon,b,\xi,\xi'}$ of the Lagrangian to zero, we also obtain the equations

$$w = \sum_{i=1}^m (\mu_i - \lambda_i)\varphi(x_i), \tag{$*_w$}$$

$$\begin{aligned}
\sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i &= 0\\
\sum_{i=1}^m \lambda_i + \sum_{i=1}^m \mu_i + \gamma &= C\nu\\
\lambda + \alpha = \frac{C}{m}, \quad \mu + \beta &= \frac{C}{m}.
\end{aligned}$$

Using the above equations, we find that the dual function G is independent of the variables β, α, β, and we obtain the following dual program:

Dual Program kernel ν-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m}(\lambda_i - \mu_i)(\lambda_j - \mu_j)\kappa(x_i, x_j) + \sum_{i=1}^{m}(\lambda_i - \mu_i)y_i \\
&\text{subject to} \\
&\qquad \sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\mu_i = 0 \\
&\qquad \sum_{i=1}^{m}\lambda_i + \sum_{i=1}^{m}\mu_i \leq C\nu \\
&\qquad 0 \leq \lambda_i \leq \frac{C}{m}, \quad 0 \leq \mu_i \leq \frac{C}{m}, \quad i = 1, \ldots, m,
\end{aligned}$$

minimizing over α and μ.

Everything we said before also applies to the kernel ν-SV regression method, except that x_i is replaced by $\varphi(x_i)$ and that the inner product $\langle -, - \rangle$ must be used, and we have the formulae

$$\begin{aligned}
w &= \sum_{i=1}^{m}(\mu_i - \lambda_i)\varphi(x_i) \\
b &= \frac{1}{2}\left(y_{i_0} + y_{j_0} - \sum_{i=1}^{m}(\mu_i - \lambda_i)(\kappa(x_i, x_{i_0}) + \kappa(x_i, x_{j_0}))\right) \\
f(x) &= \sum_{i=1}^{m}(\mu_i - \lambda_i)\kappa(x_i, x) + b,
\end{aligned}$$

expressions that only involve κ.

Remark: There is a variant of ν-SV regression obtained by setting $\nu = 0$ and holding $\epsilon > 0$ fixed. This method is called *ϵ-SV regression* or *(linear) ϵ-insensitive SV regression*. The corresponding optimization program is

Program ϵ-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i') \\
&\text{subject to} \\
&\qquad w^\top x_i + b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m \\
&\qquad - w^\top x_i - b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m,
\end{aligned}$$

minimizing over the variables w, b, ξ, and ξ', holding ϵ fixed.

It is easy to see that the dual program is

Dual Program ϵ-SV Regression:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m}(\lambda_i-\mu_i)(\lambda_j-\mu_j)x_i^\top x_j+\sum_{i=1}^{m}(\lambda_i-\mu_i)y_i+\epsilon\sum_{i=1}^{m}(\lambda_i+\mu_i)\\
&\text{subject to}\\
&\qquad \sum_{i=1}^{m}\lambda_i-\sum_{i=1}^{m}\mu_i=0\\
&\qquad 0\leq\lambda_i\leq\frac{C}{m},\quad 0\leq\mu_i\leq\frac{C}{m},\quad i=1,\ldots,m,
\end{aligned}$$

minimizing over α and μ.

The constraint

$$\sum_{i=1}^{m}\lambda_i+\sum_{i=1}^{m}\mu_i\leq C\nu$$

is gone but the extra term $\epsilon\sum_{i=1}^{m}(\lambda_i+\mu_i)$ has been added to the dual function, to prevent λ_i and μ_i from blowing up.

There is an obvious kernelized version of ϵ-SV regression. It is easy to show that ν-SV regression subsumes ϵ-SV regression, in the sense that if ν-SV regression succeeds and yields $w, b, \epsilon > 0$, then ϵ-SV regression with the same C and the same value of ϵ also succeeds and returns the same pair (w, b). For more details on these methods, see Schölkopf, Smola, Williamson, and Bartlett [147].

Remark: The linear penalty function $\sum_{i=1}^{m}(\xi_i+\xi'_i)$ can be replaced by the quadratic penalty function $\sum_{i=1}^{m}(\xi_i^2+\xi_i'^2)$; see Shawe–Taylor and Christianini [159] (Chapter 7). In this case, it is easy to see that for an optimal solution we must have $\xi_i \geq 0$ and $\xi'_i \geq 0$, so we may omit the constraints $\xi_i \geq 0$ and $\xi'_i \geq 0$. We must also have $\gamma = 0$ so we omit the variable γ as well. It can be shown that $\xi = (m/2C)\lambda$ and $\xi' = (m/2C)\mu$. This problem is very similar to the Soft Margin SVM (SVM_{s4}) discussed in Section 54.13.

56.5 ν-Regression Version 2; Penalizing b

Yet another variant of ν-SV regression is to add the term $\frac{1}{2}b^2$ to the objective function. We will see that solving the dual not only determines w but also b and ϵ (provided a mild condition on ν). We wish to solve the following program:

Program ν-SV Regression Version 2

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}w^\top w + \frac{1}{2}b^2 + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^m(\xi_i + \xi_i')\right) \\
&\text{subject to} \\
&\qquad w^\top x_i + b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0 \qquad i = 1, \ldots, m \\
&\qquad - w^\top x_i - b + y_i \leq \epsilon + \xi_i', \quad \xi_i' \geq 0 \qquad i = 1, \ldots, m,
\end{aligned}
$$

minimizing over the variables w, b, ϵ, ξ, and ξ'. The constraint $\epsilon \geq 0$ is omitted since the problem has no solution if $\epsilon < 0$.

We leave it as an exercise to show that the new Lagrangian is

$$
\begin{aligned}
L(w, b, \lambda, \mu, \xi, \xi', \epsilon, \alpha, \beta) &= \frac{1}{2}w^\top w + w^\top\left(\sum_{i=1}^m(\lambda_i - \mu_i)x_i\right) \\
&\quad + \epsilon\left(C\nu - \sum_{i=1}^m(\lambda_i + \mu_i)\right) + \sum_{i=1}^m \xi_i\left(\frac{C}{m} - \lambda_i - \alpha_i\right) \\
&\quad + \sum_{i=1}^m \xi_i'\left(\frac{C}{m} - \mu_i - \beta_i\right) + \frac{1}{2}b^2 + b\left(\sum_{i=1}^m(\lambda_i - \mu_i)\right) - \sum_{i=1}^m(\lambda_i - \mu_i)y_i.
\end{aligned}
$$

If we set the Laplacian $\nabla L_{w,\epsilon,b,\xi,\xi'}$ to zero we obtain the equations

$$
\begin{aligned}
w &= \sum_{i=1}^m(\mu_i - \lambda_i)x_i = X^\top(\mu - \lambda) \qquad (*_w) \\
C\nu - \sum_{i=1}^m(\lambda_i + \mu_i) &= 0 \\
b + \sum_{i=1}^m(\lambda_i - \mu_i) &= 0 \\
\frac{C}{m} - \lambda - \alpha = 0, &\quad \frac{C}{m} - \mu - \beta = 0.
\end{aligned}
$$

We obtain the new equation

$$
b = -\sum_{i=1}^m(\lambda_i - \mu_i) = -(\mathbf{1}_m^\top\lambda - \mathbf{1}_m^\top\mu) \qquad (*_b)
$$

determining b, which replaces the equation

$$
\sum_{i=1}^m \lambda_i - \sum_{i=1}^m \mu_i = 0.
$$

Plugging back w from $(*_w)$ and b from $(*_b)$ into the Lagrangian we get

$$\begin{aligned}
G(\lambda, \mu, \alpha, \beta) &= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top \begin{pmatrix}\lambda \\ \mu\end{pmatrix} + \frac{1}{2}b^2 - b^2 \\
&= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \frac{1}{2}b^2 \\
&= -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right) \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top \begin{pmatrix}\lambda \\ \mu\end{pmatrix},
\end{aligned}$$

with

$$P = \begin{pmatrix} XX^\top & -XX^\top \\ -XX^\top & XX^\top \end{pmatrix} = \begin{pmatrix} \mathbf{K} & -\mathbf{K} \\ -\mathbf{K} & \mathbf{K} \end{pmatrix}$$

and

$$q = \begin{pmatrix} y \\ -y \end{pmatrix}.$$

The new dual program is

Dual Program ν-SV Regression Version 2

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} \left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right) \begin{pmatrix}\lambda \\ \mu\end{pmatrix} + q^\top \begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\text{subject to} \\
&\qquad \sum_{i=1}^m \lambda_i + \sum_{i=1}^m \mu_i = C\nu \\
&\qquad 0 \leq \lambda_i \leq \frac{C}{m}, \quad 0 \leq \mu_i \leq \frac{C}{m}, \quad i = 1, \ldots, m.
\end{aligned}$$

Definition 56.1 and Definition 56.2 are unchanged. We have the following version of Proposition 56.2 showing that p_f, q_f, p_m an q_m have direct influence on the choice of ν.

Proposition 56.7. *(1) Let p_f be the number of points x_i such that $\lambda_i = C/m$, and let q_f be the number of points x_i such that $\mu_i = C/m$. We have $p_f + q_f \leq m\nu$.*

(2) Let p_m be the number of points x_i such that $\lambda_i > 0$, and let q_m be the number of points x_i such that $\mu_i > 0$. We have $p_m + q_m \geq m\nu$.

(3) If $p_f \geq 1$ or $q_f \geq 1$, then $\nu \geq 1/m$.

Proof. (1) Let K_λ and K_μ be the sets of indices corresponding to points failing the margin,

$$\begin{aligned}
K_\lambda &= \{i \in \{1, \ldots, m\} \mid \lambda_i = C/m\} \\
K_\mu &= \{i \in \{1, \ldots, m\} \mid \mu_i = C/m\}.
\end{aligned}$$

By definition $p_f = |K_\lambda|, q_f = |K_\mu|$. Since the equation

$$\sum_{i=1}^m \lambda_i + \sum_{j=1}^m \mu_j = C\nu$$

holds, by definition of K_λ and K_μ we have

$$(p_f + q_f)\frac{C}{m} = \sum_{i \in K_\lambda} \lambda_i + \sum_{j \in K_\mu} \mu_j \leq \sum_{i=1}^m \lambda_i + \sum_{j=1}^m \mu_j = C\nu,$$

which implies that

$$p_f + q_f \leq m\nu.$$

(2) Let $I_{\lambda>0}$ and $I_{\mu>0}$ be the sets of indices

$$\begin{aligned} I_{\lambda>0} &= \{i \in \{1, \ldots, m\} \mid \lambda_i > 0\} \\ I_{\mu>0} &= \{i \in \{1, \ldots, m\} \mid \mu_i > 0\}. \end{aligned}$$

By definition $p_m = |I_{\lambda>0}|, q_m = |I_{\mu>0}|$. We have

$$\sum_{i=1}^m \lambda_i + \sum_{j=1}^m \mu_j = \sum_{i \in I_{\lambda>0}} \lambda_i + \sum_{j \in I_{\mu>0}} \mu_j = C\nu.$$

Since $\lambda_i \leq C/m$ and $\mu_j \leq C/m$, we obtain

$$C\nu \leq (p_m + q_m)\frac{C}{m},$$

that is, $p_m + q_m \geq m\nu$.

(3) follows immediately from (1). ☐

Proposition 56.7 yields the following bounds on ν:

$$\frac{p_f + q_f}{m} \leq \nu \leq \frac{p_m + q_m}{m}.$$

Again, the smaller ν is, the wider the ϵ-slab is, and the larger ν is, the narrower the ϵ-slab is.

Remark: It can be shown that for any optimal solution with $w \neq 0$ and $\epsilon > 0$, if the inequalities $(p_f + q_f)/m < \nu < 1$ hold, then some point x_i is a support vector. The proof is essentially Case 1b in the proof of Proposition 56.4. We leave the details as an exercise.

The new dual program is solved using ADMM. The $(2m+1)\times 4m$ matrix A_3 corresponding to the equational constraints

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i = C\nu$$
$$\lambda + \alpha = \frac{C}{m}, \quad \mu + \beta = \frac{C}{m},$$

is given by

$$A_3 = \begin{pmatrix} \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ I_m & 0_{m,m} & I_m & 0_{m,m} \\ 0_{m,m} & I_m & 0_{m,m} & I_m \end{pmatrix}.$$

We leave it as an exercise to show that A_3 has rank $2m+1$. We define the vector c_3 (of dimension $2m+1$) as

$$c_3 = \begin{pmatrix} C\nu \\ \frac{C}{m}\mathbf{1}_{2m} \end{pmatrix}.$$

Since there are $4m$ Lagrange multipliers $(\lambda, \mu, \alpha, \beta)$, we need to pad the $2m \times 2m$ matrix $P_3 = P + \begin{pmatrix} \mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top \end{pmatrix}$ with zeros to make it into a $4m \times 4m$ matrix

$$P_{3a} = \begin{pmatrix} P_3 & 0_{2m,2m} \\ 0_{2m,2m} & 0_{2m,2m} \end{pmatrix}.$$

Similarly, we pad q with zeros to make it a vector q_{3a} of dimension $4m$,

$$q_{3a} = \begin{pmatrix} q \\ 0_{2m} \end{pmatrix}.$$

It remains to compute ϵ. Ther are two methods to do this.

The first method assumes the **Standard Margin Hypothesis**, which is that there is some i_0 such that $0 < \lambda_{i_0} < C/m$ *or* there is some j_0 such that $0 < \mu_{j_0} < C/m$; in other words, there is some support vector of type 1. By the complementary slackness conditions, $\xi_{i_0} = 0$ or $\xi'_{j_0} = 0$, so we have either $w^\top x_{i_0} + b - y_{i_0} = \epsilon$ or $-w^\top x_{j_0} - b + y_{j_0} = \epsilon$, which determines ϵ.

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$I_\lambda = \{i \in \{1, \ldots, m\} \mid 0 < \lambda_i < C/m\}$$
$$I_\mu = \{j \in \{1, \ldots, m\} \mid 0 < \mu_j < C/m\}.$$

Then it is easy to see that we can compute ϵ using the following averaging formulae: if $I_\lambda \neq \emptyset$, then

$$\epsilon = w^\top \left(\sum_{i\in I_\lambda} x_i\right)/|I_\lambda| + b - \left(\sum_{i\in I_\lambda} y_i\right)/|I_\lambda|,$$

and if $I_\mu \neq \emptyset$, then

$$\epsilon = -w^\top \left(\sum_{j\in I_\mu} x_j\right)/|I_\mu| - b + \left(\sum_{i\in I_\mu} y_i\right)/|I_\mu|.$$

The second method uses duality. Under a mild condition, expressing that the duality gap is zero, we can determine ϵ in terms of λ, μ and b. This is because points x_i that fail the margin, which means that $\lambda_i = C/m$ or $\mu_i = C/m$, are the only points for which $\xi_i > 0$ or $\xi_i' > 0$. But in this case we have an active constraint

$$w^\top x_i + b - y_i = \epsilon + \xi_i \tag{$*_\xi$}$$

or

$$-w^\top x_i - b + y_i = \epsilon + \xi_i', \tag{$*_{\xi'}$}$$

so ξ_i and ξ_i' can be expressed in terms of w and b. Since the duality gap is zero for an optimal solution, the optimal value of the primal is equal to the optimal value of the dual. Using the fact that

$$\begin{aligned} w &= X^\top(\mu - \lambda) \\ b &= -(\mathbf{1}_m^\top\lambda - \mathbf{1}_m^\top\mu) = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\begin{pmatrix}-\mathbf{1}_m \\ \mathbf{1}_m\end{pmatrix}\end{aligned}$$

we obtain an expression for the optimal value of the primal. First we have

$$\begin{aligned} \frac{1}{2}w^\top w + \frac{1}{2}b^2 &= \frac{1}{2}(\lambda^\top - \mu^\top)XX^\top(\lambda - \mu) + \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix}, \end{aligned}$$

with

$$P = \begin{pmatrix} XX^\top & -XX^\top \\ -XX^\top & XX^\top\end{pmatrix}.$$

Let K_λ and K_μ be the sets of indices corresponding to points failing the margin,

$$\begin{aligned} K_\lambda &= \{i \in \{1, \ldots, m\} \mid \lambda_i = C/m\} \\ K_\mu &= \{i \in \{1, \ldots, m\} \mid \mu_i = C/m\}. \end{aligned}$$

Because $\lambda_i\mu_i = 0$, the sets K_λ and K_μ are disjoint. Observe that from Definition 56.2 we have $p_f = |K_\lambda|$ and $q_f = |K_\mu|$. Then by $(*_\xi)$ and $(*_{\xi'})$, we have

$$\begin{aligned}
\sum_{i=1}^m (\xi_i + \xi_i') &= \sum_{i\in K_\lambda} \xi_i + \sum_{j\in K_\mu} \xi_j' \\
&= \sum_{i\in K_\lambda} (w^\top x_i + b - y_i - \epsilon) + \sum_{j\in K_\mu} (-w^\top x_j - b + y_j - \epsilon) \\
&= w^\top\Bigg(\sum_{i\in K_\lambda} x_i - \sum_{j\in K_\mu} x_j\Bigg) - \sum_{i\in K_\lambda} y_i + \sum_{j\in K_\mu} y_j + (p_f - q_f)b - (p_f + q_f)\epsilon.
\end{aligned}$$

The optimal value of the dual is given by

$$-\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top\begin{pmatrix}\lambda \\ \mu\end{pmatrix},$$

with

$$q = \begin{pmatrix} y \\ -y\end{pmatrix}.$$

Expressing that the duality gap is zero we have

$$\begin{aligned}
&\frac{1}{2}w^\top w + \frac{1}{2}b^2 + C\nu\epsilon + \frac{C}{m}\sum_{i=1}^m(\xi_i + \xi_i') \\
&\qquad = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top\begin{pmatrix}\lambda \\ \mu\end{pmatrix},
\end{aligned}$$

that is,

$$\begin{aligned}
&\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} + C\nu\epsilon \\
&\quad + \frac{C}{m}\Bigg(w^\top\Bigg(\sum_{i\in K_\lambda} x_i - \sum_{j\in K_\mu} x_j\Bigg) - \sum_{i\in K_\lambda} y_i + \sum_{j\in K_\mu} y_j + (p_f - q_f)b - (p_f + q_f)\epsilon\Bigg) \\
&\qquad = -\frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - q^\top\begin{pmatrix}\lambda \\ \mu\end{pmatrix}.
\end{aligned}$$

Solving for ϵ we get

$$\begin{aligned}
C\left(\nu - \frac{p_f + q_f}{m}\right)\epsilon &= -\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\left(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\right)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}y^\top & -y^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\
&\quad - \frac{C}{m}\Bigg(w^\top\Bigg(\sum_{i\in K_\lambda} x_i - \sum_{j\in K_\mu} x_j\Bigg) - \sum_{i\in K_\lambda} y_i + \sum_{j\in K_\mu} y_j + (p_f - q_f)b\Bigg),
\end{aligned}$$

so we get

$$\epsilon = -\Bigg(\frac{m}{C}\bigg(\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\bigg(P + \begin{pmatrix}\mathbf{1}_m\mathbf{1}_m^\top & -\mathbf{1}_m\mathbf{1}_m^\top \\ -\mathbf{1}_m\mathbf{1}_m^\top & \mathbf{1}_m\mathbf{1}_m^\top\end{pmatrix}\bigg)\begin{pmatrix}\lambda \\ \mu\end{pmatrix} + \begin{pmatrix}y^\top & -y^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu\end{pmatrix}\bigg)$$
$$+ w^\top\bigg(\sum_{i\in K_\lambda} x_i - \sum_{j\in K_\mu} x_j\bigg) - \sum_{i\in K_\lambda} y_i + \sum_{j\in K_\mu} y_j + (p_f - q_f)b\Bigg)\Big/(m\nu - p_f - q_f).$$

Using the equations

$$\begin{aligned} w &= X^\top(\mu - \lambda) \\ b &= -(\mathbf{1}_m^\top\lambda - \mathbf{1}_m^\top\mu) = \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix}\begin{pmatrix}-\mathbf{1}_m \\ \mathbf{1}_m\end{pmatrix}, \end{aligned}$$

we see that ϵ is determined by λ and μ provided that $(p_f + q_f)/m < \nu$.

By Proposition 56.7(1),

$$\frac{p_f + q_f}{m} \leq \nu,$$

therefore the condition $(p_f + q_f)/m < \nu$ is very natural.

We have implemented this method in `Matlab`, and we have observed that for some examples the choice of ν caused the equation $\nu(p_f + q_f) = m$ to hold. In such cases, running the program again with a slightly perturbed value of ν always succeeded.

The other observation we made is that b tends to be smaller and ϵ tends to be bigger in ν-SV Regression Version 2, so the fit is actually not as good as in ν-SV Regression without penalizing b. Figure 56.16 shows the result of running our program on the data set of Section 56.3. Compare with Figure 56.13.

56.6 Summary

The main concepts and results of this chapter are listed below:

- ν-support vector regression (ν-SV regression).
- Regression estimate.
- Kernel ν-SV regression.
- ϵ-SV regression, ϵ-insensitive SV regression,
- ν-SV regression Version 2; penalizing b.

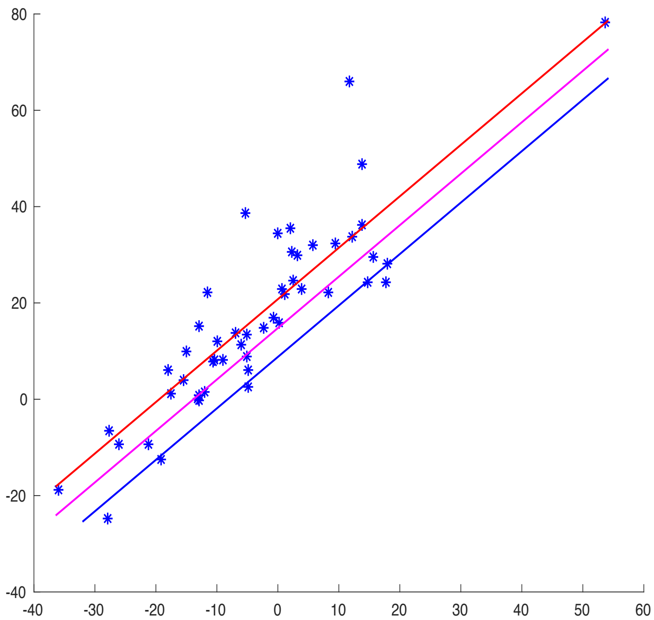


Figure 56.16: Running ν-SV regression version 2 on a set of 50 points; $\nu = 0.5$.

56.7 Problems

Problem 56.1. Prove that if ν-SV regression succeeds and yields $w, b, \epsilon > 0$, then ϵ-SV regression with the same C and the same value of ϵ also succeeds and returns the same pair (w, b).

Problem 56.2. Prove the formulae

$$b = \left(\left(\sum_{i_0 \in I_\lambda} y_{i_0}\right)/|I_\lambda| + \left(\sum_{j_0 \in I_\mu} y_{j_0}\right)/|I_\mu| - w^\top\left(\left(\sum_{i_0 \in I_\lambda} x_{i_0}\right)/|I_\lambda| + \left(\sum_{j_0 \in I_\mu} x_{j_0}\right)/|I_\mu|\right)\right)/2$$

$$\epsilon = \left(\left(\sum_{j_0 \in I_\mu} y_{j_0}\right)/|I_\mu| - \left(\sum_{i_0 \in I_\lambda} y_{i_0}\right)/|I_\lambda| + w^\top\left(\left(\sum_{i_0 \in I_\lambda} x_{i_0}\right)/|I_\lambda| - \left(\sum_{j_0 \in I_\mu} x_{j_0}\right)/|I_\mu|\right)\right)/2$$

stated just before Proposition 56.6.

Problem 56.3. Give the details of the proof of Proposition 56.6. In particular, prove that

$$\begin{aligned} C\left(\nu - \frac{p_f + q_f}{m}\right)\epsilon = &- \begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} P \begin{pmatrix}\lambda \\ \mu\end{pmatrix} - \begin{pmatrix}y^\top & -y^\top\end{pmatrix}\begin{pmatrix}\lambda \\ \mu\end{pmatrix} \\ &- \frac{C}{m}\left(w^\top\left(\sum_{i \in K_\lambda} x_i - \sum_{j \in K_\mu} x_j\right) - \sum_{i \in K_\lambda} y_i + \sum_{j \in K_\mu} y_j + (p_f - q_f)b\right). \end{aligned}$$

Problem 56.4. Prove that the matrices

$$A = \begin{pmatrix} \mathbf{1}_m^\top & -\mathbf{1}_m^\top & 0_m^\top & 0_m^\top & 0 \\ \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top & 1 \\ I_m & 0_{m,m} & I_m & 0_{m,m} & 0_m \\ 0_{m,m} & I_m & 0_{m,m} & I_m & 0_m \end{pmatrix}, \quad A_2 = \begin{pmatrix} \mathbf{1}_m^\top & -\mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ I_m & 0_{m,m} & I_m & 0_{m,m} \\ 0_{m,m} & I_m & 0_{m,m} & I_m \end{pmatrix}$$

have rank $2m+2$.

Problem 56.5. Derive the version of ν-SV regression in which the linear penalty function $\sum_{i=1}^m (\xi_i + \xi_i')$ is replaced by the quadratic penalty function $\sum_{i=1}^m (\xi_i^2 + \xi_i'^2)$. Derive the dual program.

Problem 56.6. The linear penalty function $\sum_{i=1}^m (\xi_i + \xi_i')$ can be replaced by the quadratic penalty function $\sum_{i=1}^m (\xi_i^2 + \xi_i'^2)$. Prove that for an optimal solution we must have $\xi_i \geq 0$ and $\xi_i' \geq 0$, so we may omit the constraints $\xi_i \geq 0$ and $\xi_i' \geq 0$. We must also have $\gamma = 0$ so we may omit the variable γ as well. Prove that $\xi = (m/2C)\lambda$ and $\xi' = (m/2C)\mu$. This problem is very similar to the Soft Margin SVM (SVM_{s4}) discussed in Section 54.13.

Problem 56.7. Consider the version of ν-SV regression in Section 56.5. Prove that for any optimal solution with $w \neq 0$ and $\epsilon > 0$, if the inequalities $(p_f + q_f)/m < \nu < 1$ hold, then some point x_i is a support vector.

Problem 56.8. Prove that the matrix

$$A_3 = \begin{pmatrix} \mathbf{1}_m^\top & \mathbf{1}_m^\top & 0_m^\top & 0_m^\top \\ I_m & 0_{m,m} & I_m & 0_{m,m} \\ 0_{m,m} & I_m & 0_{m,m} & I_m \end{pmatrix}$$

has rank $2m+1$.

Problem 56.9. Consider the version of ν-SV regression in Section 56.5. Prove the following formulae: If $I_\lambda \neq \emptyset$, then

$$\epsilon = w^\top \left(\sum_{i \in I_\lambda} x_i \right) / |I_\lambda| + b - \left(\sum_{i \in I_\lambda} y_i \right) / |I_\lambda|,$$

and if $I_\mu \neq \emptyset$, then

$$\epsilon = -w^\top \left(\sum_{j \in I_\mu} x_j \right) / |I_\mu| - b + \left(\sum_{i \in I_\mu} y_i \right) / |I_\mu|.$$

Problem 56.10. Implement ν-Regression Version 2 described in Section 56.5. Run examples using both the original version and version 2 and compare the results.

Part X

Appendices

Appendix A

Total Orthogonal Families in Hilbert Spaces

A.1 Total Orthogonal Families (Hilbert Bases), Fourier Coefficients

We conclude our quick tour of Hilbert spaces by showing that the notion of orthogonal basis can be generalized to Hilbert spaces. However, the useful notion is not the usual notion of a basis, but a notion which is an abstraction of the concept of Fourier series. Every element of a Hilbert space is the "sum" of its Fourier series.

Definition A.1. Given a Hilbert space E, a family $(u_k)_{k\in K}$ of nonnull vectors is an *orthogonal family* iff the u_k are pairwise orthogonal, i.e., $\langle u_i, u_j\rangle = 0$ for all $i \neq j$ $(i, j \in K)$, and an *orthonormal family* iff $\langle u_i, u_j\rangle = \delta_{i,j}$, for all $i, j \in K$. A *total orthogonal family (or system)* or *Hilbert basis* is an orthogonal family that is dense in E. This means that for every $v \in E$, for every $\epsilon > 0$, there is some finite subset $I \subseteq K$ and some family $(\lambda_i)_{i\in I}$ of complex numbers, such that

$$\left\| v - \sum_{i\in I} \lambda_i u_i \right\| < \epsilon.$$

Given an orthogonal family $(u_k)_{k\in K}$, for every $v \in E$, for every $k \in K$, the scalar $c_k = \langle v, u_k\rangle / \|u_k\|^2$ is called the *k-th Fourier coefficient of v over $(u_k)_{k\in K}$*.

Remark: The terminology Hilbert basis is misleading because a Hilbert basis $(u_k)_{k\in K}$ is not necessarily a basis in the algebraic sense. Indeed, in general, $(u_k)_{k\in K}$ does not span E. Intuitively, it takes linear combinations of the u_k's with infinitely many nonnull coefficients to span E. Technically, this is achieved in terms of limits. In order to avoid the confusion between bases in the algebraic sense and Hilbert bases, some authors refer to algebraic bases as *Hamel bases* and to total orthogonal families (or Hilbert bases) as *Schauder bases*.

Given an orthogonal family $(u_k)_{k\in K}$, for any finite subset I of K, we often call sums of the form $\sum_{i\in I} \lambda_i u_i$ *partial sums of Fourier series*, and if these partial sums converge to a limit denoted as $\sum_{k\in K} c_k u_k$, we call $\sum_{k\in K} c_k u_k$ a *Fourier series*.

However, we have to make sense of such sums! Indeed, when K is unordered or uncountable, the notion of limit or sum has not been defined. This can be done as follows (for more details, see Dixmier [51]):

Definition A.2. Given a normed vector space E (say, a Hilbert space), for any nonempty index set K, we say that a family $(u_k)_{k\in K}$ of vectors in E is *summable with sum* $v \in E$ iff for every $\epsilon > 0$, there is some finite subset I of K, such that,

$$\left\| v - \sum_{j\in J} u_j \right\| < \epsilon$$

for every finite subset J with $I \subseteq J \subseteq K$. We say that the family $(u_k)_{k\in K}$ is *summable* iff there is some $v \in E$ such that $(u_k)_{k\in K}$ is summable with sum v. A family $(u_k)_{k\in K}$ is a *Cauchy family* iff for every $\epsilon > 0$, there is a finite subset I of K, such that,

$$\left\| \sum_{j\in J} u_j \right\| < \epsilon$$

for every finite subset J of K with $I \cap J = \emptyset$,

If $(u_k)_{k\in K}$ is summable with sum v, we usually denote v as $\sum_{k\in K} u_k$. The following technical proposition will be needed:

Proposition A.1. *Let E be a complete normed vector space (say, a Hilbert space).*

(1) For any nonempty index set K, a family $(u_k)_{k\in K}$ is summable iff it is a Cauchy family.

(2) Given a family $(r_k)_{k\in K}$ of nonnegative reals $r_k \geq 0$, if there is some real number $B > 0$ such that $\sum_{i\in I} r_i < B$ for every finite subset I of K, then $(r_k)_{k\in K}$ is summable and $\sum_{k\in K} r_k = r$, where r is least upper bound of the set of finite sums $\sum_{i\in I} r_i$ ($I \subseteq K$).

Proof. (1) If $(u_k)_{k\in K}$ is summable, for every finite subset I of K, let

$$u_I = \sum_{i\in I} u_i \quad \text{and} \quad u = \sum_{k\in K} u_k$$

For every $\epsilon > 0$, there is some finite subset I of K such that

$$\|u - u_L\| < \epsilon/2$$

for all finite subsets L such that $I \subseteq L \subseteq K$. For every finite subset J of K such that $I \cap J = \emptyset$, since $I \subseteq I \cup J \subseteq K$ and $I \cup J$ is finite, we have

$$\|u - u_{I\cup J}\| < \epsilon/2 \quad \text{and} \quad \|u - u_I\| < \epsilon/2,$$

and since
$$\|u_{I\cup J} - u_I\| \leq \|u_{I\cup J} - u\| + \|u - u_I\|$$
and $u_{I\cup J} - u_I = u_J$ since $I \cap J = \emptyset$, we get

$$\|u_J\| = \|u_{I\cup J} - u_I\| < \epsilon,$$

which is the condition for $(u_k)_{k\in K}$ to be a Cauchy family.

Conversely, assume that $(u_k)_{k\in K}$ is a Cauchy family. We define inductively a decreasing sequence (X_n) of subsets of E, each of diameter at most $1/n$, as follows: For $n = 1$, since $(u_k)_{k\in K}$ is a Cauchy family, there is some finite subset J_1 of K such that

$$\|u_J\| < 1/2$$

for every finite subset J of K with $J_1 \cap J = \emptyset$. We pick some finite subset J_1 with the above property, and we let $I_1 = J_1$ and

$$X_1 = \{u_I \mid I_1 \subseteq I \subseteq K,\ I \text{ finite}\}.$$

For $n \geq 1$, there is some finite subset J_{n+1} of K such that

$$\|u_J\| < 1/(2n+2)$$

for every finite subset J of K with $J_{n+1} \cap J = \emptyset$. We pick some finite subset J_{n+1} with the above property, and we let $I_{n+1} = I_n \cup J_{n+1}$ and

$$X_{n+1} = \{u_I \mid I_{n+1} \subseteq I \subseteq K,\ I \text{ finite}\}.$$

Since $I_n \subseteq I_{n+1}$, it is obvious that $X_{n+1} \subseteq X_n$ for all $n \geq 1$. We need to prove that each X_n has diameter at most $1/n$. Since J_n was chosen such that

$$\|u_J\| < 1/(2n)$$

for every finite subset J of K with $J_n \cap J = \emptyset$, and since $J_n \subseteq I_n$, it is also true that

$$\|u_J\| < 1/(2n)$$

for every finite subset J of K with $I_n \cap J = \emptyset$ (since $I_n \cap J = \emptyset$ and $J_n \subseteq I_n$ implies that $J_n \cap J = \emptyset$). Then for every two finite subsets J, L such that $I_n \subseteq J, L \subseteq K$, we have

$$\|u_{J-I_n}\| < 1/(2n) \quad \text{and} \quad \|u_{L-I_n}\| < 1/(2n),$$

and since
$$\|u_J - u_L\| \leq \|u_J - u_{I_n}\| + \|u_{I_n} - u_L\| = \|u_{J-I_n}\| + \|u_{L-I_n}\|,$$
we get
$$\|u_J - u_L\| < 1/n,$$

which proves that $\delta(X_n) \leq 1/n$. Now if we consider the sequence of closed sets $(\overline{X_n})$, we still have $\overline{X_{n+1}} \subseteq \overline{X_n}$, and by Proposition 48.4, $\delta(\overline{X_n}) = \delta(X_n) \leq 1/n$, which means that $\lim_{n\to\infty} \delta(\overline{X_n}) = 0$, and by Proposition 48.4, $\bigcap_{n=1}^{\infty} \overline{X_n}$ consists of a single element u. We claim that u is the sum of the family $(u_k)_{k\in K}$.

For every $\epsilon > 0$, there is some $n \geq 1$ such that $n > 2/\epsilon$, and since $u \in \overline{X_m}$ for all $m \geq 1$, there is some finite subset J_0 of K such that $I_n \subseteq J_0$ and

$$\|u - u_{J_0}\| < \epsilon/2,$$

where I_n is the finite subset of K involved in the definition of X_n. However, since $\delta(X_n) \leq 1/n$, for every finite subset J of K such that $I_n \subseteq J$, we have

$$\|u_J - u_{J_0}\| \leq 1/n < \epsilon/2,$$

and since

$$\|u - u_J\| \leq \|u - u_{J_0}\| + \|u_{J_0} - u_J\|,$$

we get

$$\|u - u_J\| < \epsilon$$

for every finite subset J of K with $I_n \subseteq J$, which proves that u is the sum of the family $(u_k)_{k\in K}$.

(2) Since every finite sum $\sum_{i\in I} r_i$ is bounded by the uniform bound B, the set of these finite sums has a least upper bound $r \leq B$. For every $\epsilon > 0$, since r is the least upper bound of the finite sums $\sum_{i\in I} r_i$ (where I finite, $I \subseteq K$), there is some finite $I \subseteq K$ such that

$$\left| r - \sum_{i\in I} r_i \right| < \epsilon,$$

and since $r_k \geq 0$ for all $k \in K$, we have

$$\sum_{i\in I} r_i \leq \sum_{j\in J} r_j$$

whenever $I \subseteq J$, which shows that

$$\left| r - \sum_{j\in J} r_j \right| \leq \left| r - \sum_{i\in I} r_i \right| < \epsilon$$

for every finite subset J such that $I \subseteq J \subseteq K$, proving that $(r_k)_{k\in K}$ is summable with sum $\sum_{k\in K} r_k = r$. $\square$

Remark: The notion of summability implies that the sum of a family $(u_k)_{k\in K}$ is independent of any order on K. In this sense it is a kind of "commutative summability." More precisely, it is easy to show that for every bijection $\varphi\colon K \to K$ (intuitively, a reordering of K), the family $(u_k)_{k\in K}$ is summable iff the family $(u_l)_{l\in\varphi(K)}$ is summable, and if so, they have the same sum.

The following proposition gives some of the main properties of Fourier coefficients. Among other things, at most countably many of the Fourier coefficient may be nonnull, and the partial sums of a Fourier series converge. Given an orthogonal family $(u_k)_{k\in K}$, we let $U_k = \mathbb{C}u_k$, and $p_{U_k}\colon E \to U_k$ is the projection of E onto U_k.

Proposition A.2. *Let E be a Hilbert space, $(u_k)_{k\in K}$ an orthogonal family in E, and V the closure of the subspace generated by $(u_k)_{k\in K}$. The following properties hold:*

(1) For every $v \in E$, for every finite subset $I \subseteq K$, we have

$$\sum_{i\in I} |c_i|^2 \leq \|v\|^2,$$

where the c_k are the Fourier coefficients of v.

(2) For every vector $v \in E$, if $(c_k)_{k\in K}$ are the Fourier coefficients of v, the following conditions are equivalent:

(2a) $v \in V$

(2b) The family $(c_k u_k)_{k\in K}$ is summable and $v = \sum_{k\in K} c_k u_k$.

(2c) The family $(|c_k|^2)_{k\in K}$ is summable and $\|v\|^2 = \sum_{k\in K} |c_k|^2$;

(3) The family $(|c_k|^2)_{k\in K}$ is summable, and we have the Bessel inequality:

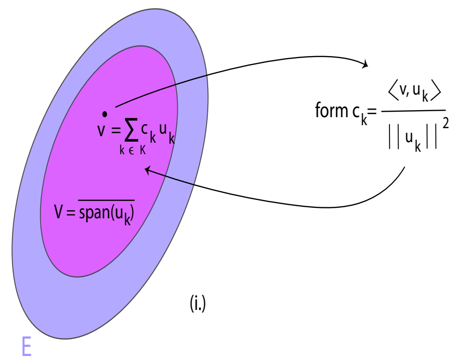
$$\sum_{k\in K} |c_k|^2 \leq \|v\|^2.$$

As a consequence, at most countably many of the c_k may be nonzero. The family $(c_k u_k)_{k\in K}$ forms a Cauchy family, and thus, the Fourier series $\sum_{k\in K} c_k u_k$ converges in E to some vector $u = \sum_{k\in K} c_k u_k$. Furthermore, $u = p_V(v)$.

See Figure A.1.

Proof. (1) Let

$$u_I = \sum_{i\in I} c_i u_i$$



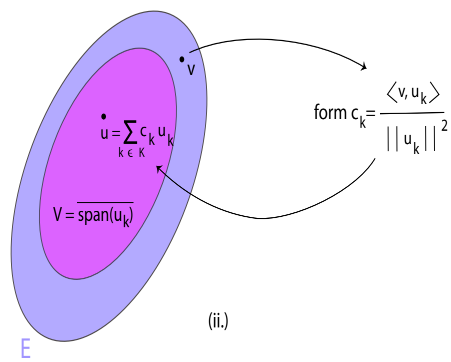


Figure A.1: A schematic illustration of Proposition A.2. Figure (i.) illustrates Condition (2b), while Figure (ii.) illustrates Condition (3). Note E is the purple oval and V is the magenta oval. In both cases, take a vector of E, form the Fourier coefficients c_k, then form the Fourier series $\sum_{k\in K} c_k u_k$. Condition (2b) ensures v equals its Fourier series since $v \in V$. However, if $v \notin V$, the Fourier series does not equal v. Eventually, we will discover that $V = E$, which implies that that Fourier series converges to its vector v.

for any finite subset I of K. We claim that $v - u_I$ is orthogonal to u_i for every $i \in I$. Indeed,

$$\begin{aligned}
\langle v - u_I, u_i \rangle &= \left\langle v - \sum_{j\in I} c_j u_j, u_i \right\rangle \\
&= \langle v, u_i \rangle - \sum_{j\in I} c_j \langle u_j, u_i \rangle \\
&= \langle v, u_i \rangle - c_i \|u_i\|^2 \\
&= \langle v, u_i \rangle - \langle v, u_i \rangle = 0,
\end{aligned}$$

since $\langle u_j, u_i \rangle = 0$ for all $i \neq j$ and $c_i = \langle v, u_i \rangle / \|u_i\|^2$. As a consequence, we have

$$\begin{aligned}
\|v\|^2 &= \Big\| v - \sum_{i \in I} c_i u_i + \sum_{i \in I} c_i u_i \Big\|^2 \\
&= \Big\| v - \sum_{i \in I} c_i u_i \Big\|^2 + \Big\| \sum_{i \in I} c_i u_i \Big\|^2 \\
&= \Big\| v - \sum_{i \in I} c_i u_i \Big\|^2 + \sum_{i \in I} |c_i|^2,
\end{aligned}$$

since the u_i are pairwise orthogonal, that is,

$$\|v\|^2 = \Big\| v - \sum_{i \in I} c_i u_i \Big\|^2 + \sum_{i \in I} |c_i|^2.$$

Thus,

$$\sum_{i \in I} |c_i|^2 \leq \|v\|^2,$$

as claimed.

(2) We prove the chain of implications $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)$.

$(a) \Rightarrow (b)$: If $v \in V$, since V is the closure of the subspace spanned by $(u_k)_{k \in K}$, for every $\epsilon > 0$, there is some finite subset I of K and some family $(\lambda_i)_{i \in I}$ of complex numbers, such that

$$\Big\| v - \sum_{i \in I} \lambda_i u_i \Big\| < \epsilon.$$

Now for every finite subset J of K such that $I \subseteq J$, we have

$$\begin{aligned}
\Big\| v - \sum_{i \in I} \lambda_i u_i \Big\|^2 &= \Big\| v - \sum_{j \in J} c_j u_j + \sum_{j \in J} c_j u_j - \sum_{i \in I} \lambda_i u_i \Big\|^2 \\
&= \Big\| v - \sum_{j \in J} c_j u_j \Big\|^2 + \Big\| \sum_{j \in J} c_j u_j - \sum_{i \in I} \lambda_i u_i \Big\|^2,
\end{aligned}$$

since $I \subseteq J$ and the u_j (with $j \in J$) are orthogonal to $v - \sum_{j \in J} c_j u_j$ by the argument in (1), which shows that

$$\Big\| v - \sum_{j \in J} c_j u_j \Big\| \leq \Big\| v - \sum_{i \in I} \lambda_i u_i \Big\| < \epsilon,$$

and thus, that the family $(c_k u_k)_{k \in K}$ is summable with sum v, so that

$$v = \sum_{k \in K} c_k u_k.$$

$(b) \Rightarrow (c)$: If $v = \sum_{k\in K} c_k u_k$, then for every $\epsilon > 0$, there some finite subset I of K, such that

$$\left\| v - \sum_{j\in J} c_j u_j \right\| < \sqrt{\epsilon},$$

for every finite subset J of K such that $I \subseteq J$, and since we proved in (1) that

$$\|v\|^2 = \left\| v - \sum_{j\in J} c_j u_j \right\|^2 + \sum_{j\in J} |c_j|^2,$$

we get

$$\|v\|^2 - \sum_{j\in J} |c_j|^2 < \epsilon,$$

which proves that $(|c_k|^2)_{k\in K}$ is summable with sum $\|v\|^2$.

$(c) \Rightarrow (a)$: Finally, if $(|c_k|^2)_{k\in K}$ is summable with sum $\|v\|^2$, for every $\epsilon > 0$, there is some finite subset I of K such that

$$\|v\|^2 - \sum_{j\in J} |c_j|^2 < \epsilon^2$$

for every finite subset J of K such that $I \subseteq J$, and again, using the fact that

$$\|v\|^2 = \left\| v - \sum_{j\in J} c_j u_j \right\|^2 + \sum_{j\in J} |c_j|^2,$$

we get

$$\left\| v - \sum_{j\in J} c_j u_j \right\| < \epsilon,$$

which proves that $(c_k u_k)_{k\in K}$ is summable with sum $\sum_{k\in K} c_k u_k = v$, and $v \in V$.

(3) Since $\sum_{i\in I} |c_i|^2 \leq \|v\|^2$ for every finite subset I of K, by Proposition A.1(2), the family $(|c_k|^2)_{k\in K}$ is summable. The Bessel inequality

$$\sum_{k\in K} |c_k|^2 \leq \|v\|^2$$

is an obvious consequence of the inequality $\sum_{i\in I} |c_i|^2 \leq \|v\|^2$ (for every finite $I \subseteq K$). Now for every natural number $n \geq 1$, if K_n is the subset of K consisting of all c_k such that $|c_k| \geq 1/n$, the number of elements in K_n is at most

$$\sum_{k\in K_n} |nc_k|^2 \leq n^2 \sum_{k\in K} |c_k|^2 \leq n^2 \|v\|^2,$$

which is finite, and thus, at most a countable number of the c_k may be nonzero.

Since $(|c_k|^2)_{k\in K}$ is summable with sum c, by Proposition A.1(1) we know that for every $\epsilon > 0$, there is some finite subset I of K such that

$$\sum_{j\in J} |c_j|^2 < \epsilon^2$$

for every finite subset J of K such that $I \cap J = \emptyset$. Since

$$\Big\| \sum_{j\in J} c_j u_j \Big\|^2 = \sum_{j\in J} |c_j|^2,$$

we get

$$\Big\| \sum_{j\in J} c_j u_j \Big\| < \epsilon.$$

This proves that $(c_k u_k)_{k\in K}$ is a Cauchy family, which, by Proposition A.1(1), implies that $(c_k u_k)_{k\in K}$ is summable since E is complete. Thus, the Fourier series $\sum_{k\in K} c_k u_k$ is summable, with its sum denoted $u \in V$.

Since $\sum_{k\in K} c_k u_k$ is summable with sum u, for every $\epsilon > 0$, there is some finite subset I_1 of K such that

$$\Big\| u - \sum_{j\in J} c_j u_j \Big\| < \epsilon$$

for every finite subset J of K such that $I_1 \subseteq J$. By the triangle inequality, for every finite subset I of K,

$$\Big\| u - v \Big\| \leq \Big\| u - \sum_{i\in I} c_i u_i \Big\| + \Big\| \sum_{i\in I} c_i u_i - v \Big\|.$$

By (2), every $w \in V$ is the sum of its Fourier series $\sum_{k\in K} \lambda_k u_k$, and for every $\epsilon > 0$, there is some finite subset I_2 of K such that

$$\Big\| w - \sum_{j\in J} \lambda_j u_j \Big\| < \epsilon$$

for every finite subset J of K such that $I_2 \subseteq J$. By the triangle inequality, for every finite subset I of K,

$$\Big\| v - \sum_{i\in I} \lambda_i u_i \Big\| \leq \| v - w \| + \Big\| w - \sum_{i\in I} \lambda_i u_i \Big\|.$$

Letting $I = I_1 \cup I_2$, since we showed in (2) that

$$\Big\| v - \sum_{i\in I} c_i u_i \Big\| \leq \Big\| v - \sum_{i\in I} \lambda_i u_i \Big\|$$

for every finite subset I of K, we get

$$\begin{aligned}\|u-v\| &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \Big\|\sum_{i\in I} c_i u_i - v\Big\| \\ &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \Big\|\sum_{i\in I} \lambda_i u_i - v\Big\| \\ &\leq \Big\|u-\sum_{i\in I} c_i u_i\Big\| + \|v-w\| + \Big\|w-\sum_{i\in I} \lambda_i u_i\Big\|,\end{aligned}$$

and thus

$$\|u-v\| \leq \|v-w\| + 2\epsilon.$$

Since this holds for every $\epsilon > 0$, we have

$$\|u-v\| \leq \|v-w\|$$

for all $w \in V$, i.e. $\|v-u\| = d(v, V)$, with $u \in V$, which proves that $u = p_V(v)$. □

A.2 The Hilbert Space $\ell^2(K)$ and the Riesz–Fischer Theorem

Proposition A.2 suggests looking at the space of sequences $(z_k)_{k\in K}$ (where $z_k \in \mathbb{C}$) such that $(|z_k|^2)_{k\in K}$ is summable. Indeed, such spaces are Hilbert spaces, and it turns out that every Hilbert space is isomorphic to one of those. Such spaces are the infinite-dimensional version of the spaces $\mathbb{C}^n$ under the usual Euclidean norm.

Definition A.3. Given any nonempty index set K, the space $\ell^2(K)$ is the set of all sequences $(z_k)_{k\in K}$, where $z_k \in \mathbb{C}$, such that $(|z_k|^2)_{k\in K}$ is summable, i.e., $\sum_{k\in K} |z_k|^2 < \infty$.

Remarks:

(1) When K is a finite set of cardinality n, $\ell^2(K)$ is isomorphic to $\mathbb{C}^n$.

(2) When $K = \mathbb{N}$, the space $\ell^2(\mathbb{N})$ corresponds to the space ℓ^2 of Example 2 in Section 14.1 . In that example, we claimed that ℓ^2 was a Hermitian space, and in fact, a Hilbert space. We now prove this fact for any index set K.

Proposition A.3. *Given any nonempty index set K, the space $\ell^2(K)$ is a Hilbert space under the Hermitian product*

$$\langle (x_k)_{k\in K}, (y_k)_{k\in K}\rangle = \sum_{k\in K} x_k \overline{y_k}.$$

The subspace consisting of sequences $(z_k)_{k\in K}$ such that $z_k = 0$, except perhaps for finitely many k, is a dense subspace of $\ell^2(K)$.

Proof. First we need to prove that $\ell^2(K)$ is a vector space. Assume that $(x_k)_{k\in K}$ and $(y_k)_{k\in K}$ are in $\ell^2(K)$. This means that $(|x_k|^2)_{k\in K}$ and $(|y_k|^2)_{k\in K}$ are summable, which, in view of Proposition A.1(2), is equivalent to the existence of some positive bounds A and B such that $\sum_{i\in I}|x_i|^2 < A$ and $\sum_{i\in I}|y_i|^2 < B$, for every finite subset I of K. To prove that $(|x_k + y_k|^2)_{k\in K}$ is summable, it is sufficient to prove that there is some $C > 0$ such that $\sum_{i\in I}|x_i + y_i|^2 < C$ for every finite subset I of K. However, the parallelogram inequality implies that

$$\sum_{i\in I}|x_i + y_i|^2 \leq \sum_{i\in I} 2(|x_i|^2 + |y_i|^2) \leq 2(A + B),$$

for every finite subset I of K, and we conclude by Proposition A.1(2). Similarly, for every $\lambda \in \mathbb{C}$,

$$\sum_{i\in I}|\lambda x_i|^2 \leq \sum_{i\in I}|\lambda|^2|x_i|^2 \leq |\lambda|^2 A,$$

and $(\lambda_k x_k)_{k\in K}$ is summable. Therefore, $\ell^2(K)$ is a vector space.

By the Cauchy-Schwarz inequality,

$$\sum_{i\in I}|x_i\overline{y_i}| \leq \sum_{i\in I}|x_i||y_i| \leq \Big(\sum_{i\in I}|x_i|^2\Big)^{1/2}\Big(\sum_{i\in I}|xy_i|^2\Big)^{1/2} \leq \sum_{i\in I}(|x_i|^2 + |y_i|^2)/2 \leq (A+B)/2,$$

for every finite subset I of K. For the third inequality we used the fact that

$$4CD \leq (C + D)^2,$$

(with $C = \sum_{i\in I}|x_i|^2$ and $D = \sum_{i\in I}|y_i|^2$) which is equivalent to

$$(C - D)^2 \geq 0.$$

By Proposition A.1(2), $(|x_k\overline{y_k}|)_{k\in K}$ is summable. The customary language is that $(x_k\overline{y_k})_{k\in K}$ is absolutely summable. However, it is a standard fact that this implies that $(x_k\overline{y_k})_{k\in K}$ is summable (For every $\epsilon > 0$, there is some finite subset I of K such that

$$\sum_{j\in J}|x_j\overline{y_j}| < \epsilon$$

for every finite subset J of K such that $I \cap J = \emptyset$, and thus

$$|\sum_{j\in J} x_j\overline{y_j}| \leq \sum_{i\in J}|x_j\overline{y_j}| < \epsilon,$$

proving that $(x_k\overline{y_k})_{k\in K}$ is a Cauchy family, and thus summable). We still have to prove that $\ell^2(K)$ is complete.

Consider a sequence $((\lambda_k^n)_{k\in K})_{n\geq 1}$ of sequences $(\lambda_k^n)_{k\in K} \in \ell^2(K)$, and assume that it is a Cauchy sequence. This means that for every $\epsilon > 0$, there is some $N \geq 1$ such that

$$\sum_{k\in K}|\lambda_k^m - \lambda_k^n|^2 < \epsilon^2$$

for all $m, n \geq N$. For every fixed $k \in K$, this implies that

$$|\lambda_k^m - \lambda_k^n| < \epsilon$$

for all $m, n \geq N$, which shows that $(\lambda_k^n)_{n\geq 1}$ is a Cauchy sequence in $\mathbb{C}$. Since $\mathbb{C}$ is complete, the sequence $(\lambda_k^n)_{n\geq 1}$ has a limit $\lambda_k \in \mathbb{C}$. We claim that $(\lambda_k)_{k\in K} \in \ell^2(K)$ and that this is the limit of $((\lambda_k^n)_{k\in K})_{n\geq 1}$.

Given any $\epsilon > 0$, the fact that $((\lambda_k^n)_{k\in K})_{n\geq 1}$ is a Cauchy sequence implies that there is some $N \geq 1$ such that for every finite subset I of K, we have

$$\sum_{i\in I} |\lambda_i^m - \lambda_i^n|^2 < \epsilon/4$$

for all $m, n \geq N$. Let $p = |I|$. Then

$$|\lambda_i^m - \lambda_i^n| < \frac{\sqrt{\epsilon}}{2\sqrt{p}}$$

for every $i \in I$. Since λ_i is the limit of $(\lambda_i^n)_{n\geq 1}$, we can find some n large enough so that

$$|\lambda_i^n - \lambda_i| < \frac{\sqrt{\epsilon}}{2\sqrt{p}}$$

for every $i \in I$. Since

$$|\lambda_i^m - \lambda_i| \leq |\lambda_i^m - \lambda_i^n| + |\lambda_i^n - \lambda_i|,$$

we get

$$|\lambda_i^m - \lambda_i| < \frac{\sqrt{\epsilon}}{\sqrt{p}},$$

and thus,

$$\sum_{i\in I} |\lambda_i^m - \lambda_i|^2 < \epsilon,$$

for all $m \geq N$. Since the above holds for every finite subset I of K, by Proposition A.1(2), we get

$$\sum_{k\in K} |\lambda_k^m - \lambda_k|^2 < \epsilon,$$

for all $m \geq N$. This proves that $(\lambda_k^m - \lambda_k)_{k\in K} \in \ell^2(K)$ for all $m \geq N$, and since $\ell^2(K)$ is a vector space and $(\lambda_k^m)_{k\in K} \in \ell^2(K)$ for all $m \geq 1$, we get $(\lambda_k)_{k\in K} \in \ell^2(K)$. However,

$$\sum_{k\in K} |\lambda_k^m - \lambda_k|^2 < \epsilon$$

for all $m \geq N$, means that the sequence $(\lambda_k^m)_{k\in K}$ converges to $(\lambda_k)_{k\in K} \in \ell^2(K)$. The fact that the subspace consisting of sequences $(z_k)_{k\in K}$ such that $z_k = 0$ except perhaps for finitely many k is a dense subspace of $\ell^2(K)$ is left as an easy exercise. $\square$

Remark: The subspace consisting of all sequences $(z_k)_{k\in K}$ such that $z_k = 0$, except perhaps for finitely many k, provides an example of a subspace which is not closed in $\ell^2(K)$. Indeed, this space is strictly contained in $\ell^2(K)$, since there are countable sequences of nonnull elements in $\ell^2(K)$ (why?).

We just need two more propositions before being able to prove that every Hilbert space is isomorphic to some $\ell^2(K)$.

Proposition A.4. *Let E be a Hilbert space, and $(u_k)_{k\in K}$ an orthogonal family in E. The following properties hold:*

(1) For every family $(\lambda_k)_{k\in K} \in \ell^2(K)$, the family $(\lambda_k u_k)_{k\in K}$ is summable. Furthermore, $v = \sum_{k\in K} \lambda_k u_k$ is the only vector such that $c_k = \lambda_k$ for all $k \in K$, where the c_k are the Fourier coefficients of v.

(2) For any two families $(\lambda_k)_{k\in K} \in \ell^2(K)$ and $(\mu_k)_{k\in K} \in \ell^2(K)$, if $v = \sum_{k\in K} \lambda_k u_k$ and $w = \sum_{k\in K} \mu_k u_k$, we have the following equation, also called Parseval identity:

$$\langle v, w\rangle = \sum_{k\in K} \lambda_k \overline{\mu_k}.$$

Proof. (1) The fact that $(\lambda_k)_{k\in K} \in \ell^2(K)$ means that $(|\lambda_k|^2)_{k\in K}$ is summable. The proof given in Proposition A.2 (3) applies to the family $(|\lambda_k|^2)_{k\in K}$ (instead of $(|c_k|^2)_{k\in K}$), and yields the fact that $(\lambda_k u_k)_{k\in K}$ is summable. Letting $v = \sum_{k\in K} \lambda_k u_k$, recall that $c_k = \langle v, u_k\rangle / \|u_k\|^2$. Pick some $k \in K$. Since $\langle -, -\rangle$ is continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that

$$|\langle v, u_k\rangle - \langle w, u_k\rangle| < \epsilon \|u_k\|^2$$

whenever

$$\|v - w\| < \eta.$$

However, since for every $\eta > 0$, there is some finite subset I of K such that

$$\left\| v - \sum_{j\in J} \lambda_j u_j \right\| < \eta$$

for every finite subset J of K such that $I \subseteq J$, we can pick $J = I \cup \{k\}$ and letting $w = \sum_{j\in J} \lambda_j u_j$ we get

$$\left| \langle v, u_k\rangle - \left\langle \sum_{j\in J} \lambda_j u_j, u_k \right\rangle \right| < \epsilon \|u_k\|^2.$$

However,

$$\langle v, u_k\rangle = c_k \|u_k\|^2 \quad \text{and} \quad \left\langle \sum_{j\in J} \lambda_j u_j, u_k \right\rangle = \lambda_k \|u_k\|^2,$$

and thus, the above proves that $|c_k - \lambda_k| < \epsilon$ for every $\epsilon > 0$, and thus, that $c_k = \lambda_k$.

(2) Since $\langle -, - \rangle$ is continuous, for every $\epsilon > 0$, there are some $\eta_1 > 0$ and $\eta_2 > 0$, such that

$$|\langle x, y\rangle| < \epsilon$$

whenever $\|x\| < \eta_1$ and $\|y\| < \eta_2$. Since $v = \sum_{k\in K} \lambda_k u_k$ and $w = \sum_{k\in K} \mu_k u_k$, there is some finite subset I_1 of K such that

$$\left\| v - \sum_{j\in J} \lambda_j u_j \right\| < \eta_1$$

for every finite subset J of K such that $I_1 \subseteq J$, and there is some finite subset I_2 of K such that

$$\left\| w - \sum_{j\in J} \mu_j u_j \right\| < \eta_2$$

for every finite subset J of K such that $I_2 \subseteq J$. Letting $I = I_1 \cup I_2$, we get

$$\left| \left\langle v - \sum_{i\in I} \lambda_i u_i, w - \sum_{i\in I} \mu_i u_i \right\rangle \right| < \epsilon.$$

Furthermore,

$$\begin{aligned}
\langle v, w\rangle &= \left\langle v - \sum_{i\in I} \lambda_i u_i + \sum_{i\in I} \lambda_i u_i, w - \sum_{i\in I} \mu_i u_i + \sum_{i\in I} \mu_i u_i \right\rangle \\
&= \left\langle v - \sum_{i\in I} \lambda_i u_i, w - \sum_{i\in I} \mu_i u_i \right\rangle + \sum_{i\in I} \lambda_i \overline{\mu_i},
\end{aligned}$$

since the u_i are orthogonal to $v - \sum_{i\in I} \lambda_i u_i$ and $w - \sum_{i\in I} \mu_i u_i$ for all $i \in I$. This proves that for every $\epsilon > 0$, there is some finite subset I of K such that

$$\left| \langle v, w\rangle - \sum_{i\in I} \lambda_i \overline{\mu_i} \right| < \epsilon.$$

We already know from Proposition A.3 that $(\lambda_k \overline{\mu_k})_{k\in K}$ is summable, and since $\epsilon > 0$ is arbitrary we get

$$\langle v, w\rangle = \sum_{k\in K} \lambda_k \overline{\mu_k}.$$

□

The next proposition states properties characterizing Hilbert bases (total orthogonal families).

Proposition A.5. *Let E be a Hilbert space, and let $(u_k)_{k\in K}$ be an orthogonal family in E. The following properties are equivalent:*

(1) The family $(u_k)_{k\in K}$ is a total orthogonal family.

(2) For every vector $v \in E$, if $(c_k)_{k\in K}$ are the Fourier coefficients of v, then the family $(c_k u_k)_{k\in K}$ is summable and $v = \sum_{k\in K} c_k u_k$.

(3) For every vector $v \in E$, we have the Parseval identity:

$$\|v\|^2 = \sum_{k\in K} |c_k|^2.$$

(4) For every vector $u \in E$, if $\langle u, u_k\rangle = 0$ for all $k \in K$, then $u = 0$.

See Figure A.2.

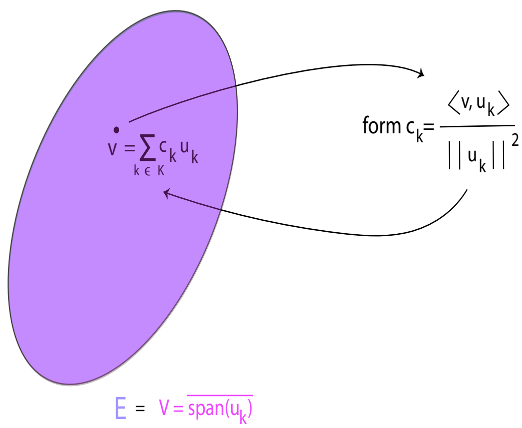


Figure A.2: A schematic illustration of Proposition A.5. Since $(u_k)_{k\in K}$ is a Hilbert basis, $V = E$. Then given a vector of E, if we form the Fourier coefficients c_k, then form the Fourier series $\sum_{k\in K} c_k u_k$, we are ensured that v is equal to its Fourier series.

Proof. The equivalence of (1), (2), and (3) is an immediate consequence of Proposition A.2 and Proposition A.4.

(4) If $(u_k)_{k\in K}$ is a total orthogonal family and $\langle u, u_k\rangle = 0$ for all $k \in K$, since $u = \sum_{k\in K} c_k u_k$ where $c_k = \langle u, u_k\rangle/\|u_k\|^2$, we have $c_k = 0$ for all $k \in K$, and $u = 0$.

Conversely, assume that the closure V of $(u_k)_{k\in K}$ is different from E. Then by Proposition 48.7, we have $E = V \oplus V^\perp$, where $V^\perp$ is the orthogonal complement of V, and $V^\perp$ is nontrivial since $V \neq E$. As a consequence, there is some nonnull vector $u \in V^\perp$. But then u is orthogonal to every vector in V, and in particular,

$$\langle u, u_k\rangle = 0$$

for all $k \in K$, which, by assumption, implies that $u = 0$, contradicting the fact that $u \neq 0$. ∎

Remarks:

(1) If E is a Hilbert space and $(u_k)_{k\in K}$ is a total orthogonal family in E, there is a simpler argument to prove that $u = 0$ if $\langle u, u_k\rangle = 0$ for all $k \in K$ based on the continuity of $\langle -, -\rangle$. The argument is to prove that the assumption implies that $\langle v, u\rangle = 0$ for all $v \in E$. Since $\langle -, -\rangle$ is positive definite, this implies that $u = 0$. By continuity of $\langle -, -\rangle$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for every finite subset I of K, for every family $(\lambda_i)_{i\in I}$, for every $v \in E$,

$$\left|\langle v, u\rangle - \left\langle \sum_{i\in I} \lambda_i u_i, u\right\rangle\right| < \epsilon$$

whenever

$$\left\|v - \sum_{i\in I} \lambda_i u_i\right\| < \eta.$$

Since $(u_k)_{k\in K}$ is dense in E, for every $v \in E$, there is some finite subset I of K and some family $(\lambda_i)_{i\in I}$ such that

$$\left\|v - \sum_{i\in I} \lambda_i u_i\right\| < \eta,$$

and since by assumption, $\left\langle \sum_{i\in I} \lambda_i u_i, u\right\rangle = 0$, we get

$$|\langle v, u\rangle| < \epsilon.$$

Since this holds for every $\epsilon > 0$, we must have $\langle v, u\rangle = 0$

(2) If V is any nonempty subset of E, the kind of argument used in the previous remark can be used to prove that $V^\perp$ is closed (even if V is not), and that $V^{\perp\perp}$ is the closure of V.

We will now prove that every Hilbert space has some Hilbert basis. This requires using a fundamental theorem from set theory known as *Zorn's lemma*, which we quickly review.

Given any set X with a partial ordering $\leq$, recall that a nonempty subset C of X is a *chain* if it is totally ordered (i.e., for all $x, y \in C$, either $x \leq y$ or $y \leq x$). A nonempty subset Y of X is *bounded* iff there is some $b \in X$ such that $y \leq b$ for all $y \in Y$. Some $m \in X$ is *maximal* iff for every $x \in X$, $m \leq x$ implies that $x = m$. We can now state Zorn's lemma. For more details, see Rudin [140], Lang [109], or Artin [7].

Proposition A.6. *(Zorn's lemma) Given any nonempty partially ordered set X, if every (nonempty) chain in X is bounded, then X has some maximal element.*

We can now prove the existence of Hilbert bases. We define a partial order on families $(u_k)_{k\in K}$ as follows: for any two families $(u_k)_{k\in K_1}$ and $(v_k)_{k\in K_2}$, we say that

$$(u_k)_{k\in K_1} \leq (v_k)_{k\in K_2}$$

iff $K_1 \subseteq K_2$ and $u_k = v_k$ for all $k \in K_1$. This is clearly a partial order.

Proposition A.7. *Let E be a Hilbert space. Given any orthogonal family $(u_k)_{k\in K}$ in E, there is a total orthogonal family $(u_l)_{l\in L}$ containing $(u_k)_{k\in K}$.*

Proof. Consider the set $\mathcal{S}$ of all orthogonal families greater than or equal to the family $B = (u_k)_{k\in K}$. We claim that every chain in $\mathcal{S}$ is bounded. Indeed, if $C = (C_l)_{l\in L}$ is a chain in $\mathcal{S}$, where $C_l = (u_{k,l})_{k\in K_l}$, the union family

$$(u_k)_{k\in \bigcup_{l\in L} K_l}, \text{ where } u_k = u_{k,l} \text{ whenever } k \in K_l,$$

is clearly an upper bound for C, and it is immediately verified that it is an orthogonal family. By Zorn's Lemma A.6, there is a maximal family $(u_l)_{l\in L}$ containing $(u_k)_{k\in K}$. If $(u_l)_{l\in L}$ is not dense in E, then its closure V is strictly contained in E, and by Proposition 48.7, the orthogonal complement $V^\perp$ of V is nontrivial since $V \neq E$. As a consequence, there is some nonnull vector $u \in V^\perp$. But then u is orthogonal to every vector in $(u_l)_{l\in L}$, and we can form an orthogonal family strictly greater than $(u_l)_{l\in L}$ by adding u to this family, contradicting the maximality of $(u_l)_{l\in L}$. Therefore, $(u_l)_{l\in L}$ is dense in E, and thus it is a Hilbert basis. $\square$

Remark: It is possible to prove that all Hilbert bases for a Hilbert space E have index sets K of the same cardinality. For a proof, see Bourbaki [27].

Definition A.4. A Hilbert space E is *separable* if its Hilbert bases are countable.

At last, we can prove that every Hilbert space is isomorphic to some Hilbert space $\ell^2(K)$ for some suitable K.

Theorem A.8. *(Riesz–Fischer) For every Hilbert space E, there is some nonempty set K such that E is isomorphic to the Hilbert space $\ell^2(K)$. More specifically, for any Hilbert basis $(u_k)_{k\in K}$ of E, the maps $f\colon \ell^2(K) \to E$ and $g\colon E \to \ell^2(K)$ defined such that*

$$f\left((\lambda_k)_{k\in K}\right) = \sum_{k\in K} \lambda_k u_k \quad \textit{and} \quad g(u) = \left(\langle u, u_k\rangle / \|u_k\|^2\right)_{k\in K} = (c_k)_{k\in K},$$

are bijective linear isometries such that $g \circ f = \mathrm{id}$ and $f \circ g = \mathrm{id}$.

Proof. By Proposition A.4 (1), the map f is well defined, and it is clearly linear. By Proposition A.2 (3), the map g is well defined, and it is also clearly linear. By Proposition A.2 (2b), we have

$$f(g(u)) = u = \sum_{k\in K} c_k u_k,$$

and by Proposition A.4 (1), we have

$$g(f((\lambda_k)_{k\in K})) = (\lambda_k)_{k\in K},$$

and thus $g \circ f = \mathrm{id}$ and $f \circ g = \mathrm{id}$. By Proposition A.4 (2), the linear map g is an isometry. Therefore, f is a linear bijection and an isometry between $\ell^2(K)$ and E, with inverse g. □

Remark: The surjectivity of the map $g\colon E \to \ell^2(K)$ is known as the *Riesz–Fischer* theorem.

Having done all this hard work, we sketch how these results apply to Fourier series. Again we refer the readers to Rudin [140] or Lang [111, 112] for a comprehensive exposition.

Let $\mathcal{C}(T)$ denote the set of all periodic continuous functions $f\colon [-\pi,\pi] \to \mathbb{C}$ with period 2π. There is a Hilbert space $\mathrm{L}^2(T)$ containing $\mathcal{C}(T)$ and such that $\mathcal{C}(T)$ is dense in $\mathrm{L}^2(T)$, whose inner product is given by

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx.$$

The Hilbert space $\mathrm{L}^2(T)$ is the space of *Lebesgue square-integrable periodic functions* (of period 2π).

It turns out that the family $(e^{ikx})_{k\in\mathbb{Z}}$ is a total orthogonal family in $\mathrm{L}^2(T)$, because it is already dense in $\mathcal{C}(T)$ (for instance, see Rudin [140]). Then the Riesz–Fischer theorem says that for every family $(c_k)_{k\in\mathbb{Z}}$ of complex numbers such that

$$\sum_{k\in\mathbb{Z}} |c_k|^2 < \infty,$$

there is a unique function $f \in \mathrm{L}^2(T)$ such that f is equal to its Fourier series

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx},$$

where the Fourier coefficients c_k of f are given by the formula

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}dt.$$

The Parseval theorem says that

$$\sum_{k=-\infty}^{+\infty} c_k\overline{d_k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\overline{g(t)}dt$$

for all $f, g \in \mathrm{L}^2(T)$, where c_k and d_k are the Fourier coefficients of f and g.

Thus, there is an isomorphism between the two Hilbert spaces $\mathrm{L}^2(T)$ and $\ell^2(\mathbb{Z})$, which is the deep reason why the Fourier coefficients "work." Theorem A.8 implies that the Fourier series $\sum_{k\in\mathbb{Z}} c_k e^{ikx}$ of a function $f \in \mathrm{L}^2(T)$ converges to f in the L^2-sense, i.e., in the mean-square sense. This does not necessarily imply that the Fourier series converges to f pointwise! This is a subtle issue, and for more on this subject, the reader is referred to Lang [111, 112] or Schwartz [152, 153].

We can also consider the set $\mathcal{C}([-1,1])$ of continuous functions $f\colon [-1,1] \to \mathbb{C}$. There is a Hilbert space $\mathrm{L}^2([-1,1])$ containing $\mathcal{C}([-1,1])$ and such that $\mathcal{C}([-1,1])$ is dense in $\mathrm{L}^2([-1,1])$, whose inner product is given by

$$\langle f, g\rangle = \int_{-1}^{1} f(x)\overline{g(x)}dx.$$

The Hilbert space $\mathrm{L}^2([-1,1])$ is the space of *Lebesgue square-integrable functions* over $[-1,1]$. The Legendre polynomials $P_n(x)$ defined in Example 5 of Section 12.2 (Chapter 12). form a Hilbert basis of $\mathrm{L}^2([-1,1])$. Recall that if we let f_n be the function

$$f_n(x) = (x^2-1)^n,$$

$P_n(x)$ is defined as follows:

$$P_0(x) = 1, \quad \text{and} \quad P_n(x) = \frac{1}{2^n n!}\, f_n^{(n)}(x),$$

where $f_n^{(n)}$ is the nth derivative of f_n. The reason for the leading coefficient is to get $P_n(1) = 1$. It can be shown with much efforts that

$$P_n(x) = \sum_{0\leq k\leq n/2} (-1)^k\, \frac{(2(n-k))!}{2^n(n-k)!k!(n-2k)!}\, x^{n-2k}.$$

A.3 Summary

The main concepts and results of this chapter are listed below:

- Hilbert space
- Orthogonal family, total orthogonal family.
- Hilbert basis.
- Fourier coefficients.
- Hamel bases, Schauder bases.
- Fourier series.

- Cauchy family, summable family.
- Bessel inequality.
- The Hilbert space $\ell^2(K)$.
- Parseval identity.
- Zorn's lemma.
- Riesz–Fischer theorem.
- Legendre polynomials.

A.4 Problems

Problem A.1. Prove that the subspace consisting of sequences $(z_k)_{k\in K}$ such that $z_k = 0$ except perhaps for finitely many k is a dense suspace of $\ell^2(K)$.

Problem A.2. If V is any nonempty subset of E, prove that $V^{\perp}$ is closed (even if V is not) and that $V^{\perp\perp}$ is the closure of V (see the remarks following Proposition A.5).

Appendix B

Matlab Programs

B.1 Hard Margin (SVM_{h2})

The following `Matlab` programs implement the method described in Section 52.7.

The first program is the heart of the method; it implements ADMM for quadratic programming.

```
function [x,u,nr,ns,k] = qsolve1(P, q, A, b, rho, tolr, tols, iternum)
%  Solve a quadratic programming problem
%  min (1/2) x^T P x + x^T q + r
%  subject to Ax = b, x >= 0  using ADMM
%  P n x n,   q, r, in R^n, A m x n,  b in R^m
%  A of rank m

m = size(A,1); fprintf('m =  %d ',m)
n = size(P,1); fprintf('    n =  %d \n',n)
u = ones(n,1); u(1,1) = 0; % to initialize u
z = ones(n,1);             % to initialize z
% iternum = maximum number of iterations;
% iternum = 80000 works well
k = 0; nr= 1; ns = 1;
%  typically tolr = 10^(-10); tols = 10^(-10);
%  Convergence is controlled by the norm nr of the primal residual r
%  and the norm ns of the dual residual s

while (k <= iternum) && (ns > tols || nr > tolr)
    z0 = z;
    k = k+1;
    %  Makes KKT matrix
    KK = [P + rho* eye(n) A'; A zeros(m,m)];
```

```
    % Makes right hand side of KKT equation
    bb = [-q + rho*(z - u); b];
    % Solves KKT equation
    xx = KK\bb;
    %  update x, z, u (ADMM update steps)
    x = xx(1:n);
    z = poslin(x + u);
    u = u + x - z;
    %  to test stopping criterion
    r = x - z;                      % primal residual
    nr = sqrt(r'*r);                % norm of primal residual
    s = rho*(z - z0);               % dual residual
    ns = sqrt(s'*s);                % norm of dual residual
end
end
```

The second program SBVMhard2 implements hard margin SVM (version 2).

```
function [lamb,mu,w] = SVMhard2(rho,u,v)
%
%   Runs hard margin SVM version 2
%
%   p green vectors u_1, ..., u_p in n x p array u
%   q red   vectors v_1, ..., v_q in n x q array v
%
%   First builds the matrices for the dual program
%
p = size(u,2); q = size(v,2); n = size(u,1);
[A,c,X,Pa,qa] = buildhardSVM2(u,v);
%
%  Runs quadratic solver
%
tolr = 10^(-10); tols = 10^(-10); iternum = 80000;
[lam,U,nr,ns,kk] = qsolve1(Pa, qa, A, c, rho, tolr, tols, iternum);
fprintf('nr =  %d ',nr)
fprintf('   ns =  %d \n',ns)
fprintf('kk =  %d \n',kk)
if kk > iternum
   fprintf('** qsolve did not converge. Problem not solvable ** \n')
end
w = -X*lam;
nw = sqrt(w'*w);   % norm of w
fprintf('nw =  %.15f \n',nw)
```

```
delta = 1/nw;
fprintf('delta =  %.15f \n',delta)
if   delta < 10^(-9)
     fprintf('** Warning, delta too small, program does not converge ** \n')
end
%
lamb = lam(1:p,1);
mu = lam(p+1:p+q,1);
b = 0;
tols = 10^(-10);
% tols < lambda_i; finds the nonzero lambda_i
[lambnz,numsvl1] = countmlu2(lamb,tols);
% tols < mu_i; finds the nonzero mu_j
[munz,numsvm1] = countmlv2(mu,tols);
fprintf('numsvl1 = %d ',numsvl1)
fprintf('  numsvm1 = %d \n',numsvm1)

if numsvl1 > 0 && numsvm1 > 0
   sx1 = zeros(n,1);  num1 = 0;
   sx2 = zeros(n,1);  num2 = 0;
   for i = 1:p
       if lambnz(i) > 0
          sx1 = sx1 + u(:,i);
          num1 = num1 + 1;
       end
   end
   for j = 1:q
       if munz(j) > 0
          sx2 = sx2 + v(:,j);
          num2 = num2 + 1;
       end
   end
   b =  (w'*(sx1/num1 + sx2/num2))/2;
   fprintf('b =  %.15f \n',b)
else
   fprintf('** Not enough support vectors ** \n')
end

if n == 2
   [ll,mm] = showdata(u,v);
   if numsvl1 > 0 && numsvm1 > 0
      showSVMs2(w,b,1,ll,mm,nw)
```

```
   end
end
end
```

The function buildhardSVM2 builds the constraint matrix and the matrices defining the quadratic functional.

```
function [A,c,X,Xa,q] = buildhardSVM2(u,v)
%   builds the matrix of constraints A for
%   hard SVM h2, and the right hand side c
%   Aso builds X and Xa = X'*X, and the vector q = -1_{p+q}
%   for the linear part of the quadratic function
%   The right-hand side is c = 0 (Ax = 0).
p = size(u,2); q = size(v,2);
A = [ones(1,p) -ones(1,q)];
c = 0;
X = [-u v];
Xa = X'*X;
q = -ones(p+q,1);
end
```

The function countmlu2 returns a vector consisting of those λ_i such that $\lambda_i > 0$, and the number of such λ_i.

```
function [lambnz, mlu] = countmlu2(lambda,tols)
%  Counts the number of points u_i (in u)
%  such that lambda_i > 0 and returns a vector
%  of these lambda_i

% tols = 10^(-11);
p = size(lambda,1); lambnz = zeros(p,1);
mlu = 0;
for i = 1:p
    if lambda(i) > tols
        mlu = mlu + 1;
        lambnz(i) = lambda(i);
    end
end
end
```

The function countmlv2 returns a vector consisting of those μ_j such that $\mu_j > 0$, and the number of such μ_j. It is similar to countmlu2. Here a judicious choice of tols is crucial and one has to experiment with various values.

The function `showdata` displays the data points (the u_i and the v_j) and the function `showSVMs2` displays the separating line and the two margin lines.

```
function  showSVMs2(w,b,eta,ll,mm,nw)
%
%  Function to display the result of running SVM
%  on p blue points u_1, ..., u_p in u
%  and q red points v_1, ..., v_q in v

l = makeline(w,b,ll,mm,nw);        %  makes separating line
lm1 = makeline(w,b+eta,ll,mm,nw);  %  makes blue margin line
lm2 = makeline(w,b-eta,ll,mm,nw);  %  makes red margin line

plot(l(1,:),l(2,:),'-m','LineWidth',1.2)            %  plots separating line
plot(lm1(1,:),lm1(2,:),'-b','LineWidth',1.2)        %  plots blue margin line
plot(lm2(1,:),lm2(2,:),'-r','LineWidth',1.2)        %  plots red margin line
hold off
end
```

Actually, implementing the above function is not entirely trivial. It is necessary to write a function `makeline` to plot the line segment which is part of the line of equation $w_1x+w_2y = b$ inside a box containing the data points. We leave the details an exercises.

B.2 Soft Margin SVM ($\mathrm{SVM}_{s2'}$)

The following `Matlab` programs implement the method described in Section 54.8.

The function `doSVMs2pbv3` calls the function `solve1` given in Section 52.7.

```
function [lamb,mu,alpha,beta,lambnz,munz,numsvl1,numsvm1,badnu,w,nw,b,eta]
    = doSVMs2pbv3(nu,rho,u,v,K)
%
%   Best version
%   Uses the duality gap to compute eta
%   In principle, needs a single support vector of type 1
%
%   Soft margin nu-SVM version s2'
%   with the constraint
%   \sum_{i = 1}^p + \sum_{j = 1}^q mu_j  = K_m
%   (without the variable gamma)
%
%   p green vectors u_1, ..., u_p in n x p array u
%   q red   vectors v_1, ..., v_q in n x q array v
```

```
%
%   First builds the matrices for the dual program
%    K is a scale factor
%
p = size(u,2); q = size(v,2); n = size(u,1);
[A,c,X,Pa,qa] = buildSVMs2pb(nu,u,v,K);
%
%  Runs quadratic solver
%
tolr = 10^(-10); tols = 10^(-10); iternum = 80000;
[x,U,nr,ns,kk] = qsolve1(Pa, qa, A, c, rho, tolr, tols, iternum);
fprintf('nr =  %d ',nr)
fprintf('   ns =  %d \n',ns)
fprintf('kk =  %d \n',kk)
noconv = 0;
if kk > iternum
   noconv = 1;
   fprintf('** qsolve did not converge. Problem not solvable ** \n')
end
lam = x(1:(p+q),1);
alpha = x((p+q+1):2*p+q,1);
beta = x(2*p+q+1:2*(p+q),1);
w = -X*lam;
nw = sqrt(w'*w);   % norm of w
fprintf('nw =  %d \n',nw)
%
lamb = x(1:p,1);
mu = x(p+1:p+q,1);
tols = 10^(-10); tolh = 10^(-9);
% tols < lambda_i < K - tolh
[lambnz,numsvl1] = findpsv2(lamb,K,tols,tolh);
% tols < mu_i < K - tolh
[munz,numsvm1] = findpsv2(mu,K,tols,tolh);
fprintf('numsvl1 = %d ',numsvl1)
fprintf('   numsvm1 = %d \n',numsvm1)
%  lambda_i >= K - tolh
[lamK,pf] = countumf2(lamb,K,tolh);  % number of blue margin failures
%  mu_j >= K - tolh
[muK,qf] = countvmf2(mu,K,tolh);  % number of red margin failures
fprintf('pf =  %d ',pf)
fprintf('   qf =  %d \n',qf)
[~,pm] = countmlu2(lamb,tols);  % number of  points such that lambda_i > tols
```

```
[~,qm] = countmlv2(mu,tols);   % number of  points such that mu_i > 0
fprintf('pm =  %d ',pm)
fprintf('   qm =  %d \n',qm)
fprintf('p - pm =  %d ',p - pm)
fprintf('   q - qm =  %d \n',q - qm)
lnu = max(2*pf/(p+q),2*qf/(p+q)); unu = min(2*pm/(p+q),2*qm/(p+q));
fprintf('lnu =  %d ',lnu)
fprintf('    unu =  %d \n',unu)
if nu < lnu
      fprintf('** Warning; nu is too small ** \n')
else
      if nu > unu
           fprintf('** Warning; nu is too big ** \n')
      end
end

sx1 = zeros(n,1);  num1 = 0;
sKu = zeros(n,1);  Knum1 = 0;
for i = 1:p
       if lambnz(i) > 0
          sx1 = sx1 + u(:,i);
          num1 = num1 + 1;
       end
       if lamK(i) > 0
          sKu = sKu + u(:,i);
          Knum1 = Knum1 + 1;
       end
end
% Knum1
sx2 = zeros(n,1);  num2 = 0;
sKv = zeros(n,1);  Knum2 = 0;
for j = 1:q
       if munz(j) > 0
          sx2 = sx2 + v(:,j);
          num2 = num2 + 1;
       end
       if muK(j) > 0
          sKv = sKv + v(:,j);
          Knum2 = Knum2 + 1;
       end
end
% Knum2
```

```
b = 0; eta = 0;
epsilon = 0; xi = 0;
P2 = X'*X;
badnu = 0;
if numsvl1 > 0
   if numsvm1 > 0
      b =  (w'*(sx1/num1 + sx2/num2))/2;
      fprintf('b =  %.15f \n',b)
      eta =  (w'*(sx1/num1 - sx2/num2))/2;
      fprintf('eta =  %.15f \n',eta)
   else
       errterm = w'*(sKv - sKu) + (pf - qf)*w'*(sx1/num1);
       Pterm = (1/K)*(lam'*P2*lam);
       denomqf = (p+q)*nu  -2*qf;
       fprintf('denomqf =  %.15f \n',denomqf)
       if denomqf > 0
          eta = (errterm + Pterm)/denomqf;
          fprintf('eta =  %.15f \n',eta)
          b = -eta + w'*sx1/num1;
       else
           badnu = 1;
           fprintf('** Warning: numsvl1 > 0, numsvm1 = 0 and nu = 2*qf/(p+q) **  \n')
       end
   end
else
    if numsvm1 > 0
       errterm = w'*(sKv - sKu) + (pf - qf)*w'*(sx2/num2);
       Pterm = (1/K)*(lam'*P2*lam);
       denompf = (p+q)*nu  -2*pf;
       fprintf('denompf =  %.15f \n',denompf)
       if denompf > 0
          eta = (errterm + Pterm)/denompf;
          fprintf('eta =  %.15f \n',eta)
          b = eta + w'*sx2/num2;
       else
           badnu = 1;
           fprintf('** Warning: numsvm1 > 0, numsvl1 = 0 and nu = 2*pf/(p+q) **  \n')
       end
    else
       fprintf('** Not enough support vectors ** \n')
    end
```

```
end

Km = (p+q)*nu*K;
fprintf('K  =  %.15f ',K)
fprintf('    (p+q)*nu*Ks/2 =  %.15f \n',Km/2)
fprintf('sum(lambda) =  %.15f ',sum(lamb))
fprintf('    sum(mu) =  %.15f \n',sum(mu))

if (numsvl1 > 0 || numsvm1 > 0) && badnu == 0
   if eta < 10^(-9)
         fprintf('** Warning, eta too small or negative ** \n')
         eta = 0;
   end
   delta = eta/nw;
   fprintf('delta =  %.15f \n',delta)
   tolxi = 10^(-10);
   % tols < lambda_i < K - tolh or K - tolh <= lambda_i and epsilon_i < tolxi
   [lamsv,psf,epsilon] = findsvl2(lamb,w,b,u,eta,K,tols,tolh,tolxi);
    % tols < mu_i < K - tolh or K - tolh <= mu_i and xi_i < tolxi
   [musv,qsf,xi] = findsvm2(mu,w,b,v,eta,K,tols,tolh,tolxi);
   fprintf('psf =  %d ',psf)
   fprintf('   qsf =  %d \n',qsf)
   fprintf('pf - psf =  %d ',pf - psf)
   fprintf('   qf - qsf =  %d \n',qf - qsf)

   % computes eta from the duality gap
   errterm = w'*(sKv - sKu) + (pf - qf)*b;
   Pterm = (1/K)*(lam'*P2*lam);
   denom = (p+q)*nu - pf -qf;
   fprintf('denom =  %.15f \n',denom)
   if denom > 0
      eta1 = (errterm + Pterm)/denom;
      fprintf('eta1 =  %.15f \n',eta1)
   end
end
end
```

The constraint matrix and the matrices defining the quadratic program are constructed by the function `buildSVMs2pb`.

```
function [A,c,X,Pa,q] = buildSVMs2pb(nu,u,v,K)
%  builds the matrix of constraints A for
%  soft margin nu-SVM s2'
```

```
%  with the constraint
%  \sum_{i = 1}^p + \sum_{j = 1}^q mu_j  = K_m
%  (without the  variable gamma) and the right-hand side c
%  u: vector of p blue points (each an n-dim vector)
%  v: vector of q red points (each an n-dim vector)
%  builds the matrix X = [-u_1 ... -u_p v1 .... v_q]
%  and the matrix Pa as 2(p+q) matrix obtained
%  by augmenting X'*X with zeros
%  K is a scale factor (K = Ks)

p = size(u,2); q = size(v,2);
% Ks = 1/(p+q);
Ks = K; Km = (p+q)*K*nu;
A = [ones(1,p) -ones(1,q) zeros(1,p+q);
     ones(1,p) ones(1,q)  zeros(1,p+q) ;
     eye(p) zeros(p,q) eye(p) zeros(p,q);
     zeros(q,p) eye(q)  zeros(q,p) eye(q) ];
c = [0; Km; Ks*ones(p+q,1)];
X = [-u v];
XX = X'*X;
Pa = [XX zeros(p+q,p+q); zeros(p+q, 2*(p+q))];
q = zeros(2*(p+q),1);
end
```

The function findpsv2 makes a vector of λ_i (and μ_j) corresponding to support vectors of type 1.

```
function [lampsv,num] = findpsv2(lambda,K,tols,tolh)
%
%   This function find the vector of
%   lambda_i's such that 0 < lambda_i < K
%   and the number of such  lambda_i.
%
% tols = 10^(-11);  %  the smaller this is, the larger the number of
                    %  points on the margin
% tolh = 10^(-9);  %
m = size(lambda,1); lampsv = zeros(m,1);
num = 0;
for i = 1:m
    if lambda(i) > tols && lambda(i) < K - tolh
       lampsv(i) = lambda(i);
       num = num + 1;
    end
```

```
end
end
```

The function countumf2 finds those λ_i such that $\lambda_i = K$.

```
function [lamK,mf] = countumf2(lambda,K,tolh)
%  Counts the number of margin failures, that is,
%  points u_i (in u) such that lambda_i = K

p = size(lambda,1);
mf = 0; lamK = zeros(p,1);
for i = 1:p
    if lambda(i) >= K - tolh
       mf = mf + 1;
       lamK(i) = lambda(i);
    end
end
end
```

Similarly, the function countvmf2 finds those μ_j such that $\mu_j = K$.

The function countmlu2 finds those λ_i such that $\lambda_i > 0$.

```
function [lambnz, mlu] = countmlu2(lambda,tols)
%  Counts the number of points u_i (in u)
%  such that lambda_i > 0 and returns a vector
%  of these lambda_i

% tols = 10^(-11);
p = size(lambda,1); lambnz = zeros(p,1);
mlu = 0;
for i = 1:p
    if lambda(i) > tols
        mlu = mlu + 1;
        lambnz(i) = lambda(i);
    end
end
end
```

Similarly, the function countmlv2 finds those μ_j such that $\mu_j > 0$. The function findsvl2 finds the λ_i corresponding to blue support vectors of type 1 and 2 and the error vector ϵ. The number of blue errors is *psf* (the u_i for which $\epsilon_i > 0$). Similarly the function findsvm2 finds the μ_j corresponding to red support vectors of type 1 and 2 and the error vector ξ. The number of red errors is *qsf* (the v_j for which $\xi_j > 0$).

The main function runSVMs2pbv3 calls doSVMs2pbv3 and displays the separating line (or plane) and the two margin lines (or planes).

```
function [lamb,mu,alpha,beta,lambnz,munz,w] = runSVMs2pbv3(nu,rho,u,v,K)
%
%   Best version
%   Uses the duality gap to compute eta
%   In principle, needs a single support vector of type 1
%
%   Runs soft margin nu-SVM version s2'
%   with the constraint
%   \sum_{i = 1}^p + \sum_{j = 1}^q mu_j  = K_m
%   (without the variable gamma)
%
%   p green vectors u_1, ..., u_p in n x p array u
%   q red   vectors v_1, ..., v_q in n x q array v
%
%   First builds the matrices for the dual program
%    K is a scale factor
%
p = size(u,2); q = size(v,2); n = size(u,1);

[lamb,mu,alpha,beta,lambnz,munz,numsvl1,numsvm1,badnu,w,nw,b,eta]
      = doSVMs2pbv3(nu,rho,u,v,K);

   if n == 2
       [ll,mm] = showdata(u,v);
       if (numsvl1 > 0 || numsvm1 > 0) && badnu == 0
          showSVMs2(w,b,eta,ll,mm,nw)
       end
   else
      if n == 3
         showpointsSVM(u,v)
         if (numsvl1 > 0 || numsvm1 > 0) && badnu == 0
            offset = 10;
            C1 = [1 0 1];  % magenta
            plotplaneSVM(u,v,w,b,offset,C1)
            C2 = [0 0 1];  % blue
            plotplaneSVM(u,v,w,b+eta,offset,C2)
            C3 = [1,0,0];  % red
            plotplaneSVM(u,v,w,b-eta,offset,C3)
         end
         axis equal
```

```
        view([-1 -1 1]);
        xlabel('X','fontsize',14);ylabel('Y','fontsize',14);
              zlabel('Z','fontsize',14);
        hold off
     end
   end
end
```

B.3 Soft Margin SVM (SVM_{s3})

The following `Matlab` programs implement the method described in Section 54.12. The main function `doSVMs3b` is given below.

```
function [lamb,mu,alpha,beta,lambnz,munz,lamK,muK,w,b,eta,nw,fail]
= doSVMs3b (nu,rho,u,v,K)
%
%  Soft margin nu-SVM version s3
%
%  Computes eta using the duality gap
%  Needs a single support vector of type 1
%
%   p green vectors u_1, ..., u_p in n x p array u
%   q red   vectors v_1, ..., v_q in n x q array v
%
%   First builds the matrices for the dual program
%    K is a scale factor
%
p = size(u,2); q = size(v,2); n = size(u,1);
[A,c,X,P2,Pa,qa] = buildSVMs3b (nu,u,v,K);
%
%  Runs quadratic solver
%
tolr = 10^(-10); tols = 10^(-10); iternum = 80000;
[x,U,nr,ns,kk] = qsolve1(Pa, qa, A, c, rho, tolr, tols, iternum);
fprintf('nr =  %d ',nr)
fprintf('    ns =  %d ',ns)
fprintf('    kk =  %d \n',kk)
noconv = 0;
if kk > iternum
   noconv = 1;
   fprintf('** qsolve did not converge. Problem not solvable ** \n')
end
```

```
lam = x(1:(p+q),1);
alpha = x((p+q+1):2*p+q,1);
beta = x(2*p+q+1:2*(p+q),1);
w = -X*lam;
nw = sqrt(w'*w);   % norm of w
fprintf('nw =  %d \n',nw)
lamb = x(1:p,1);
mu = x(p+1:p+q,1);
b = -(sum(lamb) - sum(mu));
fprintf('b =  %.15f \n',b)
%

tols = 10^(-10); tolh = 10^(-9);
% tols < lambda_i < K - tolh
[lambnz,numsvl1] = findpsv2(lamb,K,tols,tolh);
% tols < mu_i < K - tolh
[munz,numsvm1] = findpsv2(mu,K,tols,tolh);
fprintf('numsvl1 = %d ',numsvl1)
fprintf('    numsvm1 = %d \n',numsvm1)
%  lambda_i >= K - tolh
[lamK,pf] = countumf2(lamb,K,tolh);  % number of blue margin failures
%  mu_j >= K - tolh
[muK,qf] = countvmf2(mu,K,tolh);  % number of red margin failures
fprintf('pf =  %d ',pf)
fprintf('   qf =  %d \n',qf)
[~,pm] = countmlu2(lamb,tols);  % number of  points such that lambda_i > tols
[~,qm] = countmlv2(mu,tols);   % number of  points such that mu_i > 0
fprintf('pm =  %d ',pm)
fprintf('   qm =  %d \n',qm)
fprintf('p - pm =  %d ',p - pm)
fprintf('   q - qm =  %d \n',q - qm)
lnu =  (pf + qf)/(p+q); unu = (pm + qm)/(p+q);
fprintf('lnu =  %d ',lnu)
   fprintf('    unu =  %d \n',unu)
if nu < lnu
      fprintf('** Warning; nu is too small ** \n')
else
      if nu > unu
           fprintf('** Warning; nu is too big ** \n')
      end
end
```

```
sx1 = zeros(n,1);  num1 = 0;
sKu = zeros(n,1);  Knum1 = 0;
for i = 1:p
       if lambnz(i) > 0
          sx1 = sx1 + u(:,i);
          num1 = num1 + 1;
       end
       if lamK(i) > 0
          sKu = sKu + u(:,i);
          Knum1 = Knum1 + 1;
       end
end
% Knum1
sx2 = zeros(n,1);  num2 = 0;
sKv = zeros(n,1);  Knum2 = 0;
for j = 1:q
       if munz(j) > 0
          sx2 = sx2 + v(:,j);
          num2 = num2 + 1;
       end
       if muK(j) > 0
          sKv = sKv + v(:,j);
          Knum2 = Knum2 + 1;
       end
end
% Knum2
% computes eta from the duality gap
errterm = w'*(sKv - sKu) + (pf - qf)*b;
Pterm = (1/K)*(lam'*P2*lam);
denom = (p+q)*nu - pf -qf;
fprintf('denom =  %.15f \n',denom)
epsilon = 0; xi = 0;
if denom > 0
   eta = (errterm + Pterm)/denom;
   fprintf('eta =  %.15f \n',eta)
   if eta < 10^(-10)
      fprintf('** Warning; eta is too small or negative ** \n')
   end

   tolxi = 10^(-10);
   % tols < lambda_i < K - tolh or K - tolh <= lambda_i and epsilon_i < tolxi
   [lamsv,psf,epsilon] = findsvl2(lamb,w,b,u,eta,K,tols,tolh,tolxi);
```

```
    % tols < mu_i < K - tolh or K - tolh <= mu_i and xi_i < tolxi
   [musv,qsf,xi] = findsvm2(mu,w,b,v,eta,K,tols,tolh,tolxi);
   fprintf('psf =  %d ',psf)
   fprintf('   qsf =  %d \n',qsf)
   fprintf('pf - psf =  %d ',pf - psf)
   fprintf('   qf - qsf =  %d \n',qf - qsf)
else
    eta = 0;
    denom = 0;
    fprintf('** Warning, nu = (pf + qf)/(p+q) **  \n')
end

Km = (p+q)*nu*K;
fprintf('K  =  %.15f ',K)
fprintf('    (p+q)*nu*Ks =  %.15f \n',Km)
fprintf('sum(lambda) + sum(mu)=  %.15f \n',sum(lamb) + sum(mu))

eta1 = 0;
if numsvl1 > 0 || numsvm1 > 0
   if numsvl1 > numsvm1
      eta1 =  w'*sx1/num1 - b;
   else
      eta1 =  b - w'*sx2/num2;
   end
   fprintf('eta1 =  %.15f \n',eta1)
else
      fprintf('** Warning: not enough support vectors ** \n')
end
if denom == 0
   if numsvl1 > 0 || numsvm1 > 0
      eta = eta1;
      fail = 0;
   else
      fail = 1;
      fprintf('** Warning, denom = 0 and not enough support vectors **  \n')
   end
else
   fail = 0;
end
end
```

The main function doSVMs3b is executed by the following function:

```
function [lamb,mu,alpha,beta,lambnz,munz,w] = runSVMs3b(nu,rho,u,v,K)
%
%   Runs soft margin nu-SVM version s3
%
%  Computes eta using the duality gap
%  Needs a single support vector of type 1
%
%   p green vectors u_1, ..., u_p in n x p array u
%   q red   vectors v_1, ..., v_q in n x q array v
%
%   First builds the matrices for the dual program
%    K is a scale factor
%
p = size(u,2); q = size(v,2); n = size(u,1);

[lamb,mu,alpha,beta,lambnz,munz,lamK,muK,w,b,eta,nw,fail]
   = doSVMs3b(nu,rho,u,v,K);

   if n == 2
       [ll,mm] = showdata(u,v);
       if fail == 0
          showSVMs2(w,b,eta,ll,mm,nw)
       end
   else
      if n == 3
         showpointsSVM(u,v)
         if fail == 0
            offset = 10;
            C1 = [1 0 1];  % magenta
            plotplaneSVM(u,v,w,b,offset,C1)
            C2 = [0 0 1];  % blue
            plotplaneSVM(u,v,w,b+eta,offset,C2)
            C3 = [1,0,0];  % red
            plotplaneSVM(u,v,w,b-eta,offset,C3)
         end
         axis equal
  %      axis([ll(1) mm(1) ll(2) mm(2)]);
         view([-1 -1 1]);
         xlabel('X','fontsize',14);ylabel('Y','fontsize',14);zlabel('Z',
                'fontsize',14);
         hold off
      end
```

```
   end
end
```

The function buildSVMs3b builds the constraint matrix and the matrices defining the quadratic program.

```
function [A,c,X,P2,Pa,q] = buildSVMs3b(nu,u,v,K)
%  builds the matrix of constraints A for
%  soft margin nu-SVM s3 and the right-hand side c
%  u: vector of p blue points (each an n-dim vector)
%  v: vector of q red points (each an n-dim vector)
%  builds the matrix X = [-u_1 ... -u_p v1 .... v_q]
%  and the matrix Xa as 2(p+q)  matrix obtained
%  by augmenting X'*X with zeros
%  K is a scale factor (K = Ks)

p = size(u,2); q = size(v,2);
% Ks = 1/(p+q);
Ks = K; Km = (p+q)*K*nu;
A = [ones(1,p) ones(1,q)  zeros(1,p+q) ;
     eye(p) zeros(p,q) eye(p) zeros(p,q);
     zeros(q,p) eye(q)  zeros(q,p) eye(q) ];
c = [Km; Ks*ones(p+q,1)];
X = [-u v];
XX1 = X'*X;
XX2 = [ones(p,1)*ones(p,1)' -ones(p,1)*ones(q,1)';
       -ones(q,1)*ones(p,1)' ones(q,1)*ones(q,1)'];
P2 = XX1 + XX2;
Pa = [P2 zeros(p+q,p+q); zeros(p+q, 2*(p+q))];
q = zeros(2*(p+q),1);
end
```

B.4 ν-SV Regression

g The main function donuregb is given below.

```
function
[lamb,mu,alpha,beta,lambnz,munz,lamK,muK,numsvl1,numsvm1,w,epsilon,b]
            = donuregb (rho,nu,X,y,C)
%
%   Soft margin nu-regression
%   with the constraint
%   \sum_{i = 1}^m + \sum_{j = 1}^m mu_j = C nu
```

```
%   (Without the variable gamma)
%
%  Input: an m x n matrix of data points represented as
%  as the rows of X, and y a vector in R^n
%
%   First builds the matrices for the dual program
%    C is a scale factor
%
m = size(X,1); n = size(X,2);
[A,c,P,Pa,qa] = buildnuregb(nu,X,y,C);
%
%  Runs quadratic solver
%
tolr = 10^(-10); tols = 10^(-10); iternum = 80000;
[x,U,nr,ns,kk] = qsolve1(Pa, qa, A, c, rho, tolr, tols, iternum);
% fprintf('nr =  %d ',nr)
% fprintf('   ns =  %d \n',ns)
fprintf('nr =  %d',nr)
fprintf('  ns =  %d',ns)
fprintf('  kk =  %d \n',kk)
noconv = 0;
if kk > iternum
   noconv = 1;
   fprintf('** qsolve did not converge. Problem not solvable ** \n')
end
lamb = x(1:m,1);
mu = x(m+1:2*m,1);
alpha = x((2*m+1):3*m,1);
beta = x(3*m+1:4*m,1);
w = X'*(mu - lamb);
%
b = 0; epsilon = 0;
tols = 10^(-10); tolh = 10^(-9);
% tols < lambda_i < C/m - tolh
[lambnz,numsvl1] = findpsv2(lamb,C/m,tols,tolh);
% tols < mu_i < C/m - tolh
[munz,numsvm1] = findpsv2(mu,C/m,tols,tolh);
fprintf('numsvl1 = %d',numsvl1)
fprintf('   numsvm1 = %d \n',numsvm1)
%  lambda_i >= C/m - tolh
[lamK,pf] = countumf2(lamb,C/m,tolh);  % number of blue margin failures
%  mu_j >= C/m - tolh
```

```
[muK,qf] = countvmf2(mu,C/m,tolh);  % number of red margin failures
fprintf('pf =  %d',pf)
fprintf('    qf =  %d \n',qf)
[~,pm] = countmlu2(lamb,tols);  % number of  points such that lambda_i > tols
[~,qm] = countmlv2(mu,tols);   % number of  points such that mu_i > 0
fprintf('pm =  %d',pm)
fprintf('    qm =  %d \n',qm)
%  lambda_i <= tols
[lmz,nz] = countLzero(lamb,mu,tols);
pm2 = numsvl1 + pf; qm2 = numsvm1 + qf;
fprintf('pm2 =  %d',pm2)
fprintf('   qm2 =  %d \n',qm2)
lnu = max(2*pf/m,2*qf/m); unu = min(2*pm/m,2*qm/m);
fprintf('lnu =  %d',lnu)
fprintf('  unu =  %d \n',unu)
fprintf('nz =  %d \n',nz)
if nu < lnu
      fprintf('** Warning; nu is too small ** \n')
else
      if nu > unu
           fprintf('** Warning; nu is too big ** \n')
      end
end

fprintf('C/m  =  %.15f ',C/m)
fprintf('    (C nu)/2 =  %.15f \n',(C*nu)/2)
fprintf('sum(lambda) =  %.15f ',sum(lamb))
fprintf('    sum(mu) =  %.15f \n',sum(mu))
lamsv = 0; musv = 0; xi = 0; xip = 0;
if numsvl1 > 0 && numsvm1 > 0
   sx1 = zeros(n,1); sy1 = 0; num1 = 0;
   sx2 = zeros(n,1); sy2 = 0; num2 = 0;
   for i = 1:m
       if lambnz(i) > 0
          sx1 = sx1 + X(i,:)'; sy1 = sy1 + y(i);
          num1 = num1 + 1;
       end
       if munz(i) > 0
          sx2 = sx2 + X(i,:)'; sy2 = sy2 + y(i);
          num2 = num2 + 1;
       end
   end
```

```
   % num1
   % num2
   b =  (sy1/num1 + sy2/num2 - w'*(sx1/num1 + sx2/num2))/2;
   fprintf('b =  %.15f \n',b)
   epsilon =  (w'*(sx1/num1 - sx2/num2)  + sy2/num2 - sy1/num1)/2;
   fprintf('epsilon =  %.15f \n',epsilon)
   if epsilon < 10^(-10)
      fprintf('** Warning; epsilon is too small or negative ** \n')
   end
   nw = sqrt(w'*w);   % norm of w
   fprintf('nw =  %.15f \n',nw)
   %
   tolxi = 10^(-10);
   % tols < lambda_i < C/m - tolh or C/m - tolh <= lambda_i and xi_i < tolxi
   [lamsv,psf,xi] = findnuregsvl2(lamb,w,b,X,y,epsilon,C/m,tols,tolh,tolxi);
    % tols < mu_i < C/m - tolh or C/m - tolh <= mu_i and xi_i' < tolxi
   [musv,qsf,xip] = findnuregsvm2(mu,w,b,X,y,epsilon,C/m,tols,tolh,tolxi);
   fprintf('psf =  %d ',psf)
   fprintf('   qsf =  %d \n',qsf)
else
   fprintf('** Not enough support vectors ** \n')
end
end
```

To run `donuregb` use the function `runuregb` listed below.

```
function [lamb,mu,alpha,beta,lambnz,munz,lamK,muK,w] = runuregb (rho,nu,X,y,C)
%
%   Runs soft margin nu-regression
%   with the constraint
%   \sum_{i = 1}^m + \sum_{j = 1}^m mu_j = C nu
%  (Without the variable gamma)
%
%  Input: an m x n matrix of data points represented as
%  as the rows of X, and y a vector in R^n
%
%   First builds the matrices for the dual program
%    C is a scale factor
%

m = size(X,1); n = size(X,2);
[lamb,mu,alpha,beta,lambnz,munz,lamK,muK,numsvl1,numsvm1,w,epsilon,b]
      = donuregb(rho,nu,X,y,C);
```

```
   if n == 1
      [ll,mm] = showgraph(X,y);
      ww = [w;-1]; n1 = sqrt(ww'*ww);
      if numsvl1 > 0 && numsvm1 > 0
         showSVMs2(ww,-b,epsilon,ll,mm,n1)
      end
   else
       if n == 2
          offset = 10;
          [ll,mm] = showpoints(X,y,offset);
          if numsvl1 > 0 && numsvm1 > 0
             showplanes(w,b,ll,mm,epsilon)
          end
          axis equal
          axis([ll(1) mm(1) ll(2) mm(2)]);
          view([-1 -1 1]);
          xlabel('X','fontsize',14);ylabel('Y','fontsize',14);
          zlabel('Z','fontsize',14);
       end
   end
end
```

The function buildnuregb creates the constraint matrix and the matrices defining the quadratic functional.

```
function [A,c,P,Pa,qa] = buildnuregb (nu,X,y,C)
%  builds the matrix of constraints A for
%  soft margin nu-regression
%   with the constraint
%   \sum_{i = 1}^m + \sum_{j = 1}^m mu_j = C nu
%  (without the  variable gamma)
%  and the right-hand side c.
%  Input: an m x n matrix X of data points represented as
%  as the rows of X, and y a vector in R^n.
%  builds the m x m matrix X*X^T, the 2m x 2m matrix
%  P = [X*X^T -X*X^T;  -X*X^T X*X^T],
%  and the matrix Pa as the 4m  x 4m   matrix obtained
%  by augmenting  with zeros.
%  Also builds the vector q_a (q augmented with zeros).
%  C is a scale factor.

m = size(X,1); n = size(X,2);
```

```
% Ks = 1/(p+q);
Ks = C; Km = C*nu;
A = [ones(1,m) -ones(1,m) zeros(1,2*m);
     ones(1,m) ones(1,m)  zeros(1,2*m) ;
     eye(m) zeros(m,m) eye(m) zeros(m,m);
     zeros(m,m) eye(m)  zeros(m,m) eye(m)];
c = [0; Km; (Ks/m)*ones(2*m,1)];
XX1 = X*X';
P = [XX1 -XX1; -XX1 XX1];
Pa = [P zeros(2*m,2*m); zeros(2*m, 4*m )];
qa = [y; -y; zeros(2*m,1)];
end
```

Appendix C

Zorn's Lemma; Some Applications

C.1 Statement of Zorn's Lemma

Zorn's lemma is a particularly useful form of the axiom of choice, especially for algebraic applications. Readers who want to learn more about Zorn's lemma and its applications to algebra should consult either Lang [109], Appendix 2, §2 (pp. 878-884) and Chapter III, §5 (pp. 139-140), or Artin [7], Appendix §1 (pp. 588-589). For the logical ramifications of Zorn's lemma and its equivalence with the axiom of choice, one should consult Schwartz [150], (Vol. 1), Chapter I, §6, or a text on set theory such as Enderton [56], Suppes [173], or Kuratowski and Mostowski [108].

Given a set, S, a *partial order, $\leq$, on S* is a binary relation on S (i.e., $\leq \subseteq S \times S$) which is

(1) *reflexive*, i.e., $x \leq x$, for all $x \in S$,

(2) *transitive*, i.e, if $x \leq y$ and $y \leq z$, then $x \leq z$, for all $x, y, z \in S$, and

(3) *antisymmetric*, i.e, if $x \leq y$ and $y \leq x$, then $x = y$, for all $x, y \in S$.

A pair $(S, \leq)$, where $\leq$ is a partial order on S, is called a *partially ordered set* or *poset*. Given a poset, $(S, \leq)$, a subset, C, of S is *totally ordered* or a *chain* if for every pair of elements $x, y \in C$, either $x \leq y$ or $y \leq x$. The empty set is trivially a chain. A subset, P, (empty or not) of S is *bounded* if there is some $b \in S$ so that $x \leq b$ for all $x \in P$. Observe that the empty subset of S is bounded if and only if S is nonempty. A *maximal element of* P is an element, $m \in P$, so that $m \leq x$ implies that $m = x$, for all $x \in P$. Zorn's lemma can be stated as follows:

Lemma C.1. *Given a partially ordered set, $(S, \leq)$, if every chain is bounded, then S has a maximal element.*

Proof. See any of Schwartz [150], Enderton [56], Suppes [173], or Kuratowski and Mostowski [108]. $\square$

Remark: As we noted, the hypothesis of Zorn's lemma implies that S is nonempty (since the empty set must be bounded). A partially ordered set such that every chain is bounded is sometimes called *inductive*.

We now give some applications of Zorn's lemma.

C.2 Proof of the Existence of a Basis in a Vector Space

Using Zorn's lemma, we can prove that Theorem 3.7 holds for arbitrary vector spaces, and not just for finitely generated vector spaces, as promised in Chapter 3.

Theorem C.2. *Given any family, $S = (u_i)_{i\in I}$, generating a vector space E and any linearly independent subfamily, $L = (u_j)_{j\in J}$, of S (where $J \subseteq I$), there is a basis, B, of E such that $L \subseteq B \subseteq S$.*

Proof. Consider the set $\mathcal{L}$ of linearly independent families, B, such that $L \subseteq B \subseteq S$. Since $L \in \mathcal{L}$, this set is nonempty. We claim that $\mathcal{L}$ is inductive. Consider any chain, $(B_l)_{l\in\Lambda}$, of linearly independent families B_l in $\mathcal{L}$, and look at $B = \bigcup_{l\in\Lambda} B_l$. The family B is of the form $B = (v_h)_{h\in H}$, for some index set H, and it must be linearly independent. Indeed, if this was not true, there would be some family $(\lambda_h)_{h\in H}$ of scalars, of finite support, so that

$$\sum_{h\in H} \lambda_h v_h = 0,$$

where not all λ_h are zero. Since $B = \bigcup_{l\in\Lambda} B_l$ and only finitely many λ_h are nonzero, there is a finite subset, F, of Λ, so that $v_h \in B_{f_h}$ iff $\lambda_h \neq 0$. But $(B_l)_{l\in\Lambda}$ is a chain, and if we let $f = \max\{f_h \mid f_h \in F\}$, then $v_h \in B_f$, for all v_h for which $\lambda_h \neq 0$. Thus,

$$\sum_{h\in H} \lambda_h v_h = 0$$

would be a nontrivial linear dependency among vectors from B_f, a contradiction. Therefore, $B \in \mathcal{L}$, and since B is obviously an upper bound for the B_l's, we have proved that $\mathcal{L}$ is inductive. By Zorn's lemma (Lemma C.1), the set $\mathcal{L}$ has some maximal element, say $B = (u_h)_{h\in H}$. The rest of the proof is the same as in the proof of Theorem 3.7, but we repeat it for the reader's convenience. We claim that B generates E. Indeed, if B does not generate E, then there is some $u_p \in S$ that is not a linear combination of vectors in B (since S generates E), with $p \notin H$. Then, by Lemma 3.6, the family $B' = (u_h)_{h\in H\cup\{p\}}$ is linearly independent, and since $L \subseteq B \subset B' \subseteq S$, this contradicts the maximality of B. Thus, B is a basis of E such that $L \subseteq B \subseteq S$. $\square$

Another important application of Zorn's lemma is the existence of maximal ideals.

C.3 Existence of Maximal Ideals Containing a Given Proper Ideal

Let A be a commutative ring with identity element. Recall that an ideal $\mathfrak{A}$ in A is a *proper ideal* if $\mathfrak{A} \neq A$. The following theorem holds:

Theorem C.3. *Given any proper ideal, $\mathfrak{A} \subseteq A$, there is a maximal ideal, $\mathfrak{B}$, containing $\mathfrak{A}$.*

Proof. Let $\mathcal{I}$ be the set of all proper ideals, $\mathfrak{B}$, in A that contain $\mathfrak{A}$. The set $\mathcal{I}$ is nonempty, since $\mathfrak{A} \in \mathcal{I}$. We claim that $\mathcal{I}$ is inductive. Consider any chain $(\mathfrak{A}_i)_{i \in I}$ of ideals $\mathfrak{A}_i$ in A. One can easily check that $\mathfrak{B} = \bigcup_{i \in I} \mathfrak{A}_i$ is an ideal. Furthermore, $\mathfrak{B}$ is a proper ideal, since otherwise, the identity element 1 would belong to $\mathfrak{B} = A$, and so, we would have $1 \in \mathfrak{A}_i$ for some i, which would imply $\mathfrak{A}_i = A$, a contradiction. Also, $\mathfrak{B}$ is obviously an upper bound for all the $\mathfrak{A}_i$'s. By Zorn's lemma (Lemma C.1), the set $\mathcal{I}$ has a maximal element, say $\mathfrak{B}$, and $\mathfrak{B}$ is a maximal ideal containing $\mathfrak{A}$. $\square$

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. Addison Wesley, second edition, 1978.

[2] Lars V. Ahlfors and Leo Sario. *Riemann Surfaces*. Princeton Math. Series, No. 2. Princeton University Press, 1960.

[3] George E. Andrews, Richard Askey, and Ranjan Roy. *Special Functions*. Cambridge University Press, first edition, 2000.

[4] Tom Apostol. *Analysis*. Addison Wesley, second edition, 1974.

[5] V.I. Arnold. *Mathematical Methods of Classical Mechanics*. GTM No. 102. Springer Verlag, second edition, 1989.

[6] Emil Artin. *Geometric Algebra*. Wiley Interscience, first edition, 1957.

[7] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[8] Michael F. Atiyah and I. G. Macdonald. *Introduction to Commutative Algebra*. Addison Wesley, third edition, 1969.

[9] A. Avez. *Calcul Différentiel*. Masson, first edition, 1991.

[10] Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer Verlag, second edition, 2004.

[11] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[12] Marcel Berger. *Géométrie 2*. Nathan, 1990. English edition: Geometry 2, Universitext, Springer Verlag.

[13] Marcel Berger and Bernard Gostiaux. *Géométrie différentielle: variétés, courbes et surfaces*. Collection Mathématiques. Puf, second edition, 1992. English edition: Differential geometry, manifolds, curves, and surfaces, GTM No. 115, Springer Verlag.

[14] Rolf Berndt. *An Introduction to Symplectic Geometry*. Graduate Studies in Mathematics, Vol. 26. AMS, first edition, 2001.

[15] J.E. Bertin. *Algèbre linéaire et géométrie classique*. Masson, first edition, 1981.

[16] Dimitri P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, first edition, 2009.

[17] Dimitri P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, first edition, 2015.

[18] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, third edition, 2016.

[19] Dimitri P. Bertsekas, Angelina Nedić, and Asuman E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, first edition, 2003.

[20] Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, first edition, 1997.

[21] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, third edition, 1997.

[22] A. Beutelspacher and U. Rosenbaum. *Projective Geometry*. Cambridge University Press, first edition, 1998.

[23] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, first edition, 2006.

[24] Nicolas Bourbaki. *Algèbre, Chapitre 9*. Eléments de Mathématiques. Hermann, 1968.

[25] Nicolas Bourbaki. *Algèbre, Chapitres 1-3*. Eléments de Mathématiques. Hermann, 1970.

[26] Nicolas Bourbaki. *Algèbre, Chapitres 4-7*. Eléments de Mathématiques. Masson, 1981.

[27] Nicolas Bourbaki. *Espaces Vectoriels Topologiques*. Eléments de Mathématiques. Masson, 1981.

[28] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multiplier. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[29] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, first edition, 2004.

[30] Glen E Bredon. *Topology and Geometry*. GTM No. 139. Springer Verlag, first edition, 1993.

[31] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer-Verlag, first edition, 2011.

[32] G. Cagnac, E. Ramis, and J. Commeau. *Mathématiques Spéciales, Vol. 3, Géométrie*. Masson, 1965.

[33] Élie Cartan. *Theory of Spinors*. Dover, first edition, 1966.

[34] Henri Cartan. *Cours de Calcul Différentiel*. Collection Méthodes. Hermann, 1990.

[35] Henri Cartan. *Differential Forms*. Dover, first edition, 2006.

[36] Chih-Chung Chang and Lin Chih-Jen. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, 13:2119–2147, 2001.

[37] Claude Chevalley. *The Algebraic Theory of Spinors and Clifford Algebras. Collected Works, Vol. 2*. Springer, first edition, 1997.

[38] Yvonne Choquet-Bruhat, Cécile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds, and Physics, Part I: Basics*. North-Holland, first edition, 1982.

[39] Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. AMS, first edition, 1997.

[40] Vasek Chvatal. *Linear Programming*. W.H. Freeman, first edition, 1983.

[41] P.G. Ciarlet. *Introduction to Numerical Matrix Analysis and Optimization*. Cambridge University Press, first edition, 1989. French edition: Masson, 1994.

[42] Timothée Cour and Jianbo Shi. Solving markov random fields with spectral relaxation. In Marita Meila and Xiaotong Shen, editors, *Artifical Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2007.

[43] H.S.M. Coxeter. *Non-Euclidean Geometry*. The University of Toronto Press, first edition, 1942.

[44] H.S.M. Coxeter. *Introduction to Geometry*. Wiley, second edition, 1989.

[45] H.S.M. Coxeter. *The Real Projective Plane*. Springer Verlag, third edition, 1993.

[46] H.S.M. Coxeter. *Projective Geometry*. Springer Verlag, second edition, 1994.

[47] Gaston Darboux. *Principes de Géométrie Analytique*. Gauthier-Villars, first edition, 1917.

[48] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM Publications, first edition, 1997.

[49] Jean Dieudonné. *Algèbre Linéaire et Géométrie Elémentaire*. Hermann, second edition, 1965.

[50] Jean Dieudonné. *Sur les Groupes Classiques.* Hermann, third edition, 1967.

[51] Jacques Dixmier. *General Topology.* UTM. Springer Verlag, first edition, 1984.

[52] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice Hall, 1976.

[53] Manfredo P. do Carmo. *Riemannian Geometry.* Birkhäuser, second edition, 1992.

[54] David S. Dummit and Richard M. Foote. *Abstract Algebra.* Wiley, second edition, 1999.

[55] Gerald A. Edgar. *Measure, Topology, and Fractal Geometry.* Undergraduate Texts in Mathematics. Springer Verlag, first edition, 1992.

[56] Herbert B. Enderton. *Elements of Set Theory.* Academic Press, 1997.

[57] Charles L. Epstein. *Introduction to the Mathematics of Medical Imaging.* SIAM, second edition, 2007.

[58] Gerald Farin. *Curves and Surfaces for CAGD.* Academic Press, fourth edition, 1998.

[59] Olivier Faugeras. *Three-Dimensional Computer Vision, A geometric Viewpoint.* the MIT Press, first edition, 1996.

[60] Gerd Fischer. *Mathematical Models, Commentary.* Vieweg & Sohn, first edition, 1986.

[61] Gerd Fischer. *Mathematische Modelle.* Vieweg & Sohn, first edition, 1986.

[62] Gerd Fischer. *Plane Algebraic Curves.* Student Mathematical Library. AMS, first edition, 2001.

[63] James Foley, Andries van Dam, Steven Feiner, and John Hughes. *Computer Graphics. Principles and Practice.* Addison-Wesley, second edition, 1993.

[64] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach.* Prentice Hall, first edition, 2002.

[65] Jean Fresnel. *Méthodes Modernes En Géométrie.* Hermann, first edition, 1998.

[66] William Fulton. *Algebraic Curves.* Advanced Book Classics. Addison Wesley, first edition, 1989.

[67] William Fulton. *Algebraic Topology, A first course.* GTM No. 153. Springer Verlag, first edition, 1995.

[68] William Fulton and Joe Harris. *Representation Theory, A first course.* GTM No. 129. Springer Verlag, first edition, 1991.

[69] D. Gabay. Applications of the method of multipliers to variational inequalities. *Studies in Mathematics and Applications*, 15(C):299–331, 1983.

[70] Jean H. Gallier. *Curves and Surfaces In Geometric Modeling: Theory And Algorithms*. Morgan Kaufmann, 1999.

[71] Jean H. Gallier. *Discrete Mathematics*. Universitext. Springer Verlag, first edition, 2011.

[72] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[73] Jean H. Gallier. Notes on Convex Sets, Polytopes, Polyhedra, Combinatorial Topology, Voronoi Diagrams, and Delaunay Triangulations. Technical report, University of Pennsylvania, CIS Department, Philadelphia, PA 19104, 2016. Book in Preparation.

[74] Jean H. Gallier. Spectral Graph Theory of Unsigned and Signed Graphs. Applications to Graph Clustering: A survey. Technical report, University of Pennsylvania, 2019. http://www.cis.upenn.edu/ jean/spectral-graph-notes.pdf.

[75] Walter Gander, Gene H. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114/115:815–839, 1989.

[76] Roger Godement. *Cours d'Algèbre*. Hermann, first edition, 1963.

[77] Chris Godsil and Gordon Royle. *Algebraic Graph Theory*. GTM No. 207. Springer Verlag, first edition, 2001.

[78] Gene H. Golub. Some modified eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

[79] Gene H. Golub and Frank Uhlig. The QR algorithm: 50 years later its genesis by john francis and vera kublanovskaya and subsequent developments. *IMA Journal of Numerical Analysis*, 29:467–485, 2009.

[80] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[81] A. Gray. *Modern Differential Geometry of Curves and Surfaces*. CRC Press, second edition, 1997.

[82] Donald T. Greenwood. *Principles of Dynamics*. Prentice Hall, second edition, 1988.

[83] Larry C. Grove. *Classical Groups and Geometric Algebra*. Graduate Studies in Mathematics, Vol. 39. AMS, first edition, 2002.

[84] Jacques Hadamard. *Leçons de Géométrie Elémentaire. I Géométrie Plane*. Armand Colin, thirteenth edition, 1947.

[85] Jacques Hadamard. *Leçons de Géométrie Elémentaire. II Géométrie dans l'Espace.* Armand Colin, eighth edition, 1949.

[86] N. Halko, P.G. Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[87] Joseph Harris. *Algebraic Geometry, A first course.* GTM No. 133. Springer Verlag, first edition, 1992.

[88] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition, 2009.

[89] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity. The Lasso and Generalizations.* CRC Press, first edition, 2015.

[90] Sigurdur Helgason. *Groups and Geometric Analysis. Integral Geometry, Invariant Differential Operators and Spherical Functions.* MSM, Vol. 83. AMS, first edition, 2000.

[91] Nicholas J. Higham. *Functions of Matrices. Theory and Computation.* SIAM, first edition, 2008.

[92] D. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination.* Chelsea Publishing Co., 1952.

[93] Morris W. Hirsh and Stephen Smale. *Differential Equations, Dynamical Systems and Linear Algebra.* Academic Press, first edition, 1974.

[94] Kenneth Hoffman and Ray Kunze. *Linear Algebra.* Prentice Hall, second edition, 1971.

[95] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, first edition, 1990.

[96] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, first edition, 1994.

[97] Claus Hulek. *Elementary Algebraic Geometry.* Student Mathematical Library. AMS, first edition, 2003.

[98] Nathan Jacobson. *Basic Algebra I.* Dover Press, second edition, 1985.

[99] Nathan Jacobson. *Basic Algebra II.* Dover Press, second edition, 1989.

[100] Ramesh Jain, Rangachar Katsuri, and Brian G. Schunck. *Machine Vision.* McGraw-Hill, first edition, 1995.

[101] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer Verlag, fourth edition, 2005.

[102] D. Kincaid and W. Cheney. *Numerical Analysis*. Brooks/Cole Publishing, second edition, 1996.

[103] Felix Klein. *Vorlesungen Über Nicht-Euklidische Geometrie*. AMS Chelsea, first edition, 1927.

[104] Anthony W. Knapp. *Lie Groups Beyond an Introduction*. Progress in Mathematics, Vol. 140. Birkhäuser, second edition, 2002.

[105] A.N. Kolmogorov and S.V. Fomin. *Introductory Real Analysis*. Dover, first edition, 1975.

[106] Erwin Kreyszig. *Differential Geometry*. Dover, first edition, 1991.

[107] P. G. Kumpel and J. A. Thorpe. *Linear Algebra*. W. B. Saunders, first edition, 1983.

[108] K. Kuratowski and A. Mostowski. *Set Theory*. Studies in Logic, Vol. 86. Elsevier, 1976.

[109] Serge Lang. *Algebra*. Addison Wesley, third edition, 1993.

[110] Serge Lang. *Differential and Riemannian Manifolds*. GTM No. 160. Springer Verlag, third edition, 1995.

[111] Serge Lang. *Real and Functional Analysis*. GTM 142. Springer Verlag, third edition, 1996.

[112] Serge Lang. *Undergraduate Analysis*. UTM. Springer Verlag, second edition, 1997.

[113] Peter Lax. *Linear Algebra and Its Applications*. Wiley, second edition, 2007.

[114] N. N. Lebedev. *Special Functions and Their Applications*. Dover, first edition, 1972.

[115] Daniel Lehmann and Rudolphe Bkouche. *Initiation à la Géométrie*. Puf, first edition, 1988.

[116] David G. Luenberger. *Optimization by Vector Space Methods*. Wiley, first edition, 1997.

[117] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*. Verlag, fourth edition, 2016.

[118] Saunders Mac Lane and Garrett Birkhoff. *Algebra*. Macmillan, first edition, 1967.

[119] M.-P. Malliavin. *Algèbre Commutative. Applications en Géométrie et Théorie des Nombres.* Masson, first edition, 1985.

[120] Jerrold E. Marsden and Thomas J.R. Hughes. *Mathematical Foundations of Elasticity.* Dover, first edition, 1994.

[121] William S. Massey. *Algebraic Topology: An Introduction.* GTM No. 56. Springer Verlag, second edition, 1987.

[122] William S. Massey. *A Basic Course in Algebraic Topology.* GTM No. 127. Springer Verlag, first edition, 1991.

[123] Jiri Matousek and Bernd Gartner. *Understanding and Using Linear Programming.* Universitext. Springer Verlag, first edition, 2007.

[124] Dimitris N. Metaxas. *Physics-Based Deformable Models.* Kluwer Academic Publishers, first edition, 1997.

[125] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM, first edition, 2000.

[126] John W. Milnor. *Topology from the Differentiable Viewpoint.* The University Press of Virginia, second edition, 1969.

[127] R. Mneimné and F. Testard. *Introduction à la Théorie des Groupes de Lie Classiques.* Hermann, first edition, 1997.

[128] T. Molla. Class notes, math 588 example 5. Technical report, 2015. http://myweb.usf.edu/molla/2015_spring_math588/example5.pdf.

[129] Shigeyuki Morita. *Geometry of Differential Forms.* Translations of Mathematical Monographs No 201. AMS, first edition, 2001.

[130] James R. Munkres. *Analysis on Manifolds.* Addison Wesley, 1991.

[131] James R. Munkres. *Topology.* Prentice Hall, second edition, 2000.

[132] Ivan Niven, Herbert S. Zuckerman, and Hugh L. Montgomery. *An Introduction to the Theory of Numbers.* Wiley, fifth edition, 1991.

[133] Joseph O'Rourke. *Computational Geometry in C.* Cambridge University Press, second edition, 1998.

[134] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization. Algorithms and Complexity.* Dover, first edition, 1998.

[135] Beresford N. Parlett. *The Symmetric Eigenvalue Problem.* SIAM Publications, first edition, 1997.

[136] Dan Pedoe. *Geometry, A comprehensive Course*. Dover, first edition, 1988.

[137] M. Penna and R. Patterson. *Projective Geometry and its Applications to Computer Graphics*. Prentice Hall, first edition, 1986.

[138] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[139] Eugène Rouché and Charles de Comberousse. *Traité de Géométrie*. Gauthier-Villars, seventh edition, 1900.

[140] Walter Rudin. *Real and Complex Analysis*. McGraw Hill, third edition, 1987.

[141] Walter Rudin. *Functional Analysis*. McGraw Hill, second edition, 1991.

[142] Pierre Samuel. *Projective Geometry*. Undergraduate Texts in Mathematics. Springer Verlag, first edition, 1988.

[143] Pierre Samuel. *Algebraic Theory of Numbers*. Dover, first edition, 2008.

[144] Giovanni Sansone. *Orthogonal Functions*. Dover, first edition, 1991.

[145] Berhnard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, first edition, 2002.

[146] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, and Alex J. Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[147] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[148] Alexander Schrijver. *Theory of Linear and Integer Programming*. Wiley, first edition, 1999.

[149] Laurent Schwartz. *Topologie Générale et Analyse Fonctionnelle*. Collection Enseignement des Sciences. Hermann, 1980.

[150] Laurent Schwartz. *Analyse I. Théorie des Ensembles et Topologie*. Collection Enseignement des Sciences. Hermann, 1991.

[151] Laurent Schwartz. *Analyse II. Calcul Différentiel et Equations Différentielles*. Collection Enseignement des Sciences. Hermann, 1992.

[152] Laurent Schwartz. *Analyse III. Calcul Intégral*. Collection Enseignement des Sciences. Hermann, 1993.

[153] Laurent Schwartz. *Analyse IV. Applications à la Théorie de la Mesure*. Collection Enseignement des Sciences. Hermann, 1993.

[154] Jennifer Seberry, Beata J. Wysocki, and Tadeusz A. Wysocki. On some applications of Hadamard matrices. *Metrika*, 62:221–239, 2005.

[155] H. Seifert and W. Threlfall. *A Textbook of Topology*. Academic Press, first edition, 1980.

[156] Denis Serre. *Matrices, Theory and Applications*. GTM No. 216. Springer Verlag, second edition, 2010.

[157] Jean-Pierre Serre. *A Course in Arithmetic*. Graduate Text in Mathematics, No. 7. Springer, first edition, 1973.

[158] Igor R. Shafarevich. *Basic Algebraic Geometry 1*. Springer Verlag, second edition, 1994.

[159] John Shawe-Taylor and Nello Cristianini. *Kernel methods for Pattern Analysis*. Cambridge University Press, first edition, 2004.

[160] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[161] J.-C. Sidler. *Géométrie Projective*. InterEditions, first edition, 1993.

[162] Ernst Snapper and Robert J. Troyer. *Metric Affine Geometry*. Dover, first edition, 1989.

[163] Daniel Spielman. Spectral graph theory. In Uwe Naumannn and Olaf Schenk, editors, *Combinatorial Scientific Computing*. CRC Press, 2012.

[164] Harold M. Stark. *An Introduction to Number Theory*. MIT Press, first edition, 1994. Eighth Printing.

[165] G.W. Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[166] J.J. Stoker. *Differential Geometry*. Wiley Classics. Wiley-Interscience, first edition, 1989.

[167] J. Stolfi. *Oriented Projective Geometry*. Academic Press, first edition, 1991.

[168] Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. *Wavelets for Computer Graphics Theory and Applications*. Morgan Kaufmann, first edition, 1996.

[169] Gilbert Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, first edition, 1986.

[170] Gilbert Strang. *Linear Algebra and its Applications.* Saunders HBJ, third edition, 1988.

[171] Gilbert Strang. *Linear Algebra and Learning from Data.* Wellesley-Cambridge Press, first edition, 2019.

[172] Gilbert Strang and Nguyen Truong. *Wavelets and Filter Banks.* Wellesley-Cambridge Press, second edition, 1997.

[173] Patrick Suppes. *Axiomatic Set Theory.* Dover, 1972.

[174] Donald E. Taylor. *The Geometry of the Classical Groups.* Sigma Series in Pure Mathematics, Vol. 9. Heldermann Verlag Berlin, 1992.

[175] Claude Tisseron. *Géométries affines, projectives, et euclidiennes.* Hermann, first edition, 1994.

[176] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra.* SIAM Publications, first edition, 1997.

[177] Joel A. Tropp. Improved analysis of the subsampled Hadamard transform. *Advances in Adaptive Data Analysis*, 3:115–126, 2011.

[178] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3D Computer Vision.* Prentice-Hall, first edition, 1998.

[179] B.L. Van Der Waerden. *Algebra, Vol. 1.* Ungar, seventh edition, 1973.

[180] J.H. van Lint and R.M. Wilson. *A Course in Combinatorics.* Cambridge University Press, second edition, 2001.

[181] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions.* Springer, fourth edition, 2014.

[182] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, first edition, 1998.

[183] O. Veblen and J. W. Young. *Projective Geometry, Vol. 1.* Ginn, second edition, 1938.

[184] O. Veblen and J. W. Young. *Projective Geometry, Vol. 2.* Ginn, first edition, 1946.

[185] Lucas Vienne. *Présentation algébrique de la géométrie classique.* Vuibert, first edition, 1996.

[186] Frank Warner. *Foundations of Differentiable Manifolds and Lie Groups.* GTM No. 94. Springer Verlag, first edition, 1983.

[187] David S. Watkins. Understanding the QR algorithm. *SIAM Review*, 24(4):447–440, 1982.

[188] David S. Watkins. The QR algorithm revisited. *SIAM Review*, 50(1):133–145, 2008.

[189] Alan Watt. *3D Computer Graphics*. Addison-Wesley, second edition, 1993.

[190] Ernst Witt. Theorie der quadratischen Formen in beliebigen Körpern. *J. Reine Angew. Math.*, 176:31–44, 1936.

[191] Stella X. Yu. *Computational Models of Perceptual Organization*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA 15213, USA, 2003. Dissertation.

[192] Stella X. Yu and Jianbo Shi. Grouping with bias. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Neural Information Processing Systems, Vancouver, Canada, 3-8 Dec. 2001*. MIT Press, 2001.

[193] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *9th International Conference on Computer Vision, Nice, France, October 13-16*. IEEE, 2003.

[194] Oscar Zariski and Pierre Samuel. *Commutative Algebra, Vol I*. Dover Press, first edition, 2019.

[195] Gunter Ziegler. *Lectures on Polytopes*. GTM No. 152. Springer Verlag, first edition, 1997.

Index