Fundamentals of Linear Algebra and Optimization

Lasso Regression: Learning an Affine Function

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu

May 7, 2020

Lasso Regression for an Affine Function

To learn an affine function $f(x) = x^\top w + b$, we solve the following optimization problem

Lasso Regression for an Affine Function

To learn an affine function $f(x) = x^\top w + b$, we solve the following optimization problem

Program (lasso3):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\ \text{subject to} \quad & \\ & y - Xw - b\mathbf{1}_m = \xi \\ & w \leq \epsilon \\ & -w \leq \epsilon. \end{aligned}$$

Lasso Regression for an Affine Function

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.

Lasso Regression for an Affine Function

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.

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

$$b = \overline{y} - \sum_{j=1}^{n} \overline{X^j} w_j,$$

where $\overline{y}$ is the mean of y and $\overline{X^j}$ is the mean of the jth column of X.

Lasso Regression: Affine Reduction

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 to show that the Program $(\mathbf{lasso3})$ is *equivalent* to applying lasso regression $(\mathbf{lasso2})$ without an intercept term to the centered data, by replacing y by $\widehat{y} = y - \overline{y}\mathbf{1}$ and X by $\widehat{X} = X - \overline{X}$.

Lasso Regression: Affine Reduction

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 to show that the Program $(\mathbf{lasso3})$ is *equivalent* to applying lasso regression $(\mathbf{lasso2})$ without an intercept term to the centered data, by replacing y by $\widehat{y} = y - \overline{y}\mathbf{1}$ and X by $\widehat{X} = X - \overline{X}$.

This is the method described by Hastie, Tibshirani, and Wainwright (Section 2.2).

Lasso Regression: Illustrated Example

Example. 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" $X * ww$.

Lasso Regression: Illustrated Example

The first five rows of X are

$$\begin{pmatrix} -1.1658 & -0.0679 & -1.6118 & 0.3199 & 0.4400 \\ -1.1480 & -0.1952 & -0.0245 & -0.5583 & -0.6169 \\ 0.1049 & -0.2176 & -1.9488 & -0.3114 & 0.2748 \\ 0.7223 & -0.3031 & 1.0205 & -0.5700 & 0.6011 \\ 2.5855 & 0.0230 & 0.8617 & -1.0257 & 0.0923 \end{pmatrix},$$

Lasso Regression: Illustrated Example

and the first five rows of y are

$$y = \begin{pmatrix} -1.0965 \\ 1.2155 \\ 0.4324 \\ 1.1902 \\ 3.1346 \end{pmatrix}.$$

Lasso Regression: Illustrated Example

We ran the program for lasso using ADMM with various values of ρ and τ, including $\rho = 1$ and $\rho = 10$.

Lasso Regression: Illustrated Example

We ran the program for lasso using ADMM 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 $\rho = 1$.

Lasso Regression: Illustrated Example

We ran the program for lasso using ADMM 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 1, 2, and 3.

Lasso Regression: Illustrated Example

We ran the program for lasso using ADMM 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 1, 2, and 3.

This behavior is also observed in Hastie, Tibshirani, and Wainwright.

Lasso Regression: Illustrated Example

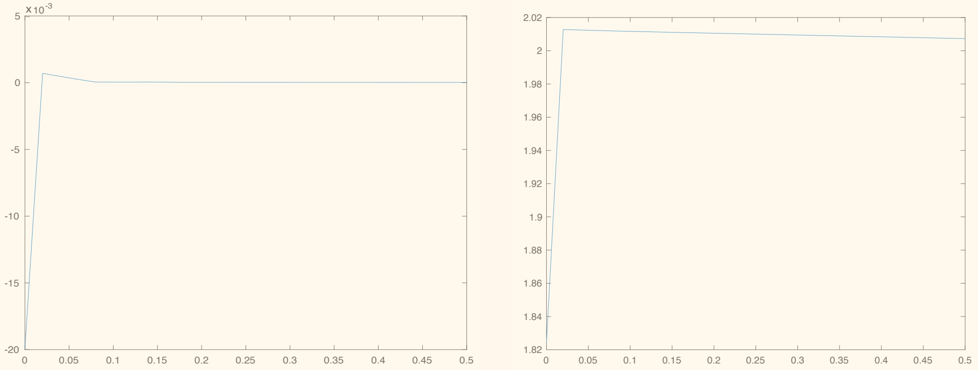

Figure 1: First and second component of w.

Lasso Regression: Illustrated Example

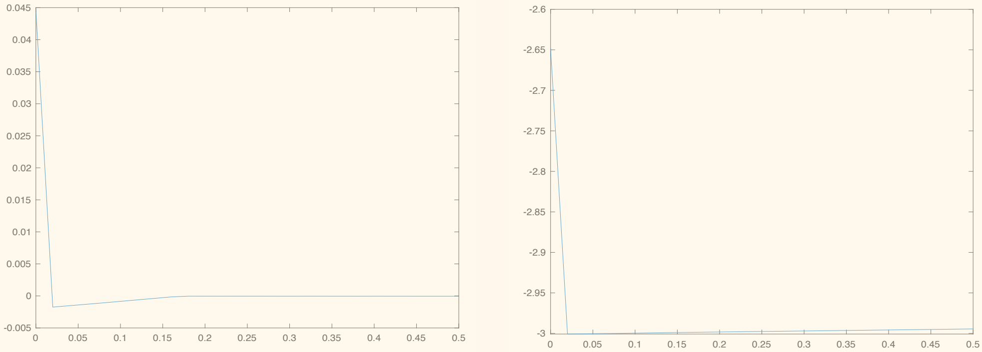

Figure 2: Third and fourth component of w.

Lasso Regression: Illustrated Example

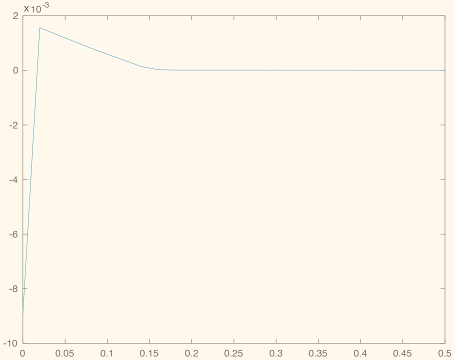

Figure 3: Fifth component of w.

Lasso Regression: Illustrated Example

For $\tau = 0.02$, we have

$$w = \begin{pmatrix} 0.00003 \\ 2.01056 \\ -0.00004 \\ -2.99821 \\ 0.00000 \end{pmatrix}, \quad b = 0.00135.$$

Lasso Regression: Illustrated Example

This weight vector w is very close to the original vector $ww = [0; 2; 0; -3; 0]$ that we used to create y.

Lasso Regression: Illustrated Example

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}.

Comparison of Ridge Regression Methods

It is interesting to compare the behavior of the methods:

Comparison of Ridge Regression Methods

It is interesting to compare the behavior of the methods:

1. Ridge regression $(\mathbf{RR6})$ (which is equivalent to $(\mathbf{RR3})$).

Comparison of Ridge Regression Methods

It is interesting to compare the behavior of the methods:

1. Ridge regression $(\mathbf{RR6})$ (which is equivalent to $(\mathbf{RR3})$).
2. Ridge regression $(\mathbf{RR3}b)$, with b penalized (by adding the term Kb^2 to the objective function).

Comparison of Ridge Regression Methods

It is interesting to compare the behavior of the methods:

1. Ridge regression $(\mathbf{RR6})$ (which is equivalent to $(\mathbf{RR3})$).
2. Ridge regression $(\mathbf{RR3}b)$, with b penalized (by adding the term Kb^2 to the objective function).
3. Least squares applied to $[X\ \mathbf{1}]$.

Comparison of Ridge Regression Methods

It is interesting to compare the behavior of the methods:

1. Ridge regression $(\mathbf{RR6})$ (which is equivalent to $(\mathbf{RR3})$).
2. Ridge regression $(\mathbf{RR3}b)$, with b penalized (by adding the term Kb^2 to the objective function).
3. Least squares applied to $[X\ \mathbf{1}]$.
4. $(\mathbf{lasso3})$.

Comparison of Ridge Regression Methods

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;
```

Comparison of Ridge Regression Methods

The result is shown in Figure 4, 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.

Comparison of Ridge Regression Methods

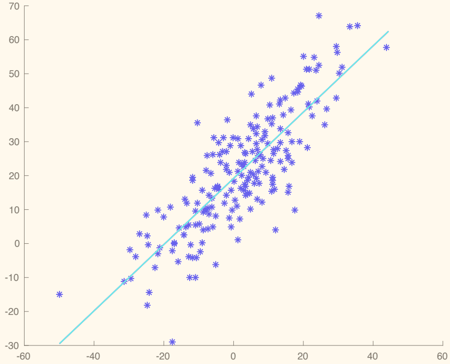

Figure 4: Comparison of the four methods with $K = \tau = 0.1$.

Comparison of Ridge Regression Methods

In order to see a difference we also ran our programs with $K = 1000$ and $\tau = 10000$; see Figure 5.

Comparison of Ridge Regression Methods

Figure 5: Comparison of the four methods with $K = 1000, \tau = 10000$.

Comparison of Ridge Regression Methods

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.

Comparison of Ridge Regression Methods

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 $20, 50$).