# Fundamentals of Linear Algebra and Optimization
# Elastic Net Regression


Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu


December 8, 2020

# *Weakness of Lasso 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.

# *Elastic Net Regression*

This way we obtain a hybrid of lasso and ridge regression called the *elastic net method* and defined as follows:

**Program** (elastic net):

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

# *Elastic Net Regression*

Some of the literature denotes $K$ by $\lambda_2$ and $\tau$ by $\lambda_1$, but we prefer not to adopt this notation since we use $\lambda, \mu$ *etc.* to denote Lagrange multipliers.

Observe that as in the case of ridge regression, minimization is performed over $\xi$, $w$, $\epsilon$ and $b$, but $b$ is *not* penalized in the objective function.

The objective function is *strictly convex* so if an optimal solution exists, then it is *unique*.

# *Elastic Net Regression: Lagrange Multipliers*

Let $\lambda \in \mathbb{R}^m$ be the Lagrange multipliers associated with the equation $y - Xw - b\mathbf{1}_m = \xi$, let $\alpha_+ \in \mathbb{R}^n_+$ be the Lagrange multipliers associated with the inequalities $w \leq \epsilon$, and let $\alpha_- \in \mathbb{R}^n_+$ be the Lagrange multipliers associated with the inequalities $-w \leq \epsilon$.

# *Elastic Net Regression: Lagrangian*

The Lagrangian associated with this optimization problem is

$$L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) = \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda$$
$$+ \epsilon^\top(\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2} K w^\top w,$$

so by setting the gradient $\nabla L_{\xi, w, \epsilon, b}$ to zero we obtain the equations

$$\xi = \lambda$$
$$Kw = -(\alpha_+ - \alpha_- - X^\top \lambda) \qquad (*_w)$$
$$\alpha_+ + \alpha_- - \tau \mathbf{1}_n = 0$$
$$\mathbf{1}_m^\top \lambda = 0.$$

# *Elastic Net Regression: Dual Function*

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 $\lambda = K\mu$.

# *Elastic Net Regression: Dual Program*

After some algebra we find that 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

$$A \begin{pmatrix}\beta_+ \\ \beta_- \\ \mu_+ \\ \mu_-\end{pmatrix} = c, \qquad \beta_+, \beta_- \in \mathbb{R}_+^n, \mu_+, \mu_- \in \mathbb{R}_+^m,$$

# *Elastic Net Regression: Dual Program*

with

$$P = \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top + KI_m & -XX^\top - KI_m \\ X & -X & -XX^\top - KI_m & XX^\top + KI_m \end{pmatrix},$$

$$q = \begin{pmatrix} 0_n \\ 0_n \\ -y \\ y \end{pmatrix}.$$

# *Elastic Net Regression: Dual Program*

and with

$$A = \begin{pmatrix} I_n & I_n & 0_{n,m} & 0_{n,m} \\ 0_n^\top & 0_n^\top & \mathbf{1}_m^\top & -\mathbf{1}_m^\top \end{pmatrix}$$

and

$$c = \begin{pmatrix} \frac{\tau}{K}\mathbf{1}_n \\ 0 \end{pmatrix}.$$

# Solution to Elastic Net Regression

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 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 $j$th column of $X$.

We leave it as an easy exercise to show that $A$ has rank $n+1$. Then we can solve the dual program using ADMM.

# *Elastic Net Regression*

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 $w$.

# *Elastic Net Regression*

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 the previous section 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 $\tau$ are varied (the best way to do this is to use $\theta$ as explained above).

# *Elastic Net Regression*

We have observed that lasso seems to converge much faster than elastic net when $K < 10^{-3}$.

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.

# *Elastic Net Regression*

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 $n$.