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: