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):

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

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

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{2}\left(\|y - \lambda\|_2^2 - \|y\|_2^2\right) \\ &\text{subject to} \\ &\qquad\qquad \left\|X^\top \lambda\right\|_\infty \leq \tau \\ &\qquad\qquad \mathbf{1}_m^\top \lambda = 0, \end{aligned}$$

which is equivalent to