Shashank Singh[1]

2 Safe rules for the LASSO [25 points] (Adona)

(a) The primal problem can be rewritten as

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} f(z) + \lambda\|\beta\|_1 \quad \text{such that} \quad z = X\beta,$$

and hence the dual function is

$$\begin{aligned}
g(u) &= \min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} f(z) + \lambda\|\beta\|_1 + u^T(z - X\beta) \\
&= \left(\min_{z\in\mathbb{R}^n} f(z) + u^T z\right) + \min_{\beta\in\mathbb{R}^p} \lambda(\|\beta\|_1 - u^T X\beta/\lambda) \\
&= -\left(\max_{z\in\mathbb{R}^n} u^T z - f(-z)\right) - \lambda \max_{\beta\in\mathbb{R}^p}(u^T X\beta/\lambda - \|\beta\|_1) \\
&= -f^*(-z) - (\|\cdot\|_1)^* \left(X^T u/\lambda\right) \\
&= -f^*(-z) - I_{\{v:\|v\|_\infty\leq 1\}}(X^T u/\lambda) = -f^*(-z) - I_{\{v:\|v\|_\infty\leq \lambda\}}(X^T u)
\end{aligned}$$

Thus, it follows from the definition of the indicator function that dual problem is

$$\max_{u\in\mathbb{R}^n} g(u) = \max_{\substack{u\in\mathbb{R}^n \\ \|X^T u\|_\infty \leq \lambda}} -f^*(-u). \quad \blacksquare$$

The stationarity KKT condition for β gives

$$\begin{aligned}
0 &\in \partial(f(z) + \lambda\|\beta\|_1) + u^T\partial(z - X\beta) \\
&= \lambda\partial\|\beta\|_1 - u^T\partial X\beta \qquad\qquad = \lambda\partial\|\beta\|_1 - X^T u,
\end{aligned}$$

so that

$$X^t u \in \lambda\partial\|\beta\|_1 = \lambda \begin{cases} [-1, 1] & : \text{ if } \beta_i = 0 \\ \{\text{sign}(\beta_i)\} & : \text{ if } \beta_i \neq 0 \end{cases}.$$

(this last equality was shown in class). $\blacksquare$

(b) By definition, the dual problem is (after replacing u with $-u$ since it is unconstrained)

$$\begin{aligned}
\max_{\mu>0} \min_{u\in\mathbb{R}^n} X_k^T u + \mu(f^*(u) + \gamma) &= \min_{\mu\geq 0} \mu \max_{u\in\mathbb{R}^n} \frac{-X_k^T u}{\mu} - f^*(u) - \gamma \\
&= \min_{\mu\geq 0} \mu f\left(-\frac{X_k}{\mu}\right) - \mu\gamma,
\end{aligned}$$

where we used the fact that, since f is convex and continuous, $f^{**} = f$. $\blacksquare$

[1] sss1@andrew.cmu.edu

(c) Plugging in the LASSO function and noting $\|X_k\| = 1$ gives

$$\begin{aligned} T_{k,+} &= \min_{\mu>0} -\mu\gamma + \frac{\mu}{2}\left\|Y + \frac{X_k}{\mu}\right\|_2^2 \\ &= \frac{1}{2}\min_{\mu>0} -2\mu\gamma + \mu\|Y\|_2^2 + 2Y^T X_k + \frac{1}{\mu}. \end{aligned}$$

Then, setting the appropriate derivative with respect to μ to 0, we have

$$0 = -2\gamma + \|Y\|_2^2 - \mu^{-2} \quad \text{so} \quad \mu = \frac{1}{\sqrt{\|Y\|_2^2 - 2\gamma}}.$$

Plugging this into $T_{k,+}$ gives

$$T_{k,+} = \sqrt{Y^T Y - 2\gamma} + Y^T X_k. \quad \blacksquare$$

(d) Since all the steps in part (c) would work if we replaced X_k with $-X_k$, we have

$$T_{k,-} = \sqrt{Y^T Y - 2\gamma} - Y^T X_k.$$

Thus,

$$\max(T_{k,+}, T_{k,-}) = \sqrt{Y^T Y - 2\gamma} + |Y^T X_k|. \quad \blacksquare$$

(e) To be feasible, we want $sY^T Y \geq 2\gamma$, so choose

$$s = \frac{2\gamma}{\|Y\|_2^2}.$$

Then, $\gamma = \frac{s\|Y\|_2^2}{2}$. $\blacksquare$