Homework 4

Benjamin Noland

1. Let $\rho = \sigma\sqrt{2\log(p/\delta)/n}$, where $0 < \delta < 1/2$. Assume the compatibility condition holds for S and constants $\nu_0 > 0$ and $\xi_0 > 1$, and suppose that the penalty constant $\lambda = A_0\rho$, where $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$. In addition, assume that $p \geq 1/\delta$. Let Ω denote the event

$$\Omega = \left\{ \max_{1 \leq j \leq p} |\langle Y - X\beta^*, X_j \rangle| \leq \rho \right\}.$$

Then Lemma 3 of the lectures implies that under Ω,

$$\|X\hat{\beta} - X\beta^*\|_n^2 + (A_0 - 1)\rho\|\hat{\beta} - \beta^*\|_1 \leq 2A_0\rho \sum_{j \in S} |\hat{\beta}_j - \beta_j^*|.$$

In particular,

$$\begin{aligned} 2A_0\lambda_0 \sum_{j \in S} |\hat{\beta}_j - \beta_j^*| &\geq (A_0 - 1)\rho\|\hat{\beta} - \beta^*\|_1 \\ &= (A_0 - 1)\rho \sum_{j \notin S} |\hat{\beta}_j - \beta_j^*| + (A_0 - 1)\rho \sum_{j \in S} |\hat{\beta}_j - \beta_j^*|, \end{aligned}$$

so that rearrangement yields

$$(A_0 - 1)\rho \sum_{j \notin S} |\hat{\beta}_j - \beta_j^*| \leq (A_0 + 1)\rho \sum_{j \in S} |\hat{\beta}_j - \beta_j^*|,$$

and hence

$$\sum_{j \notin S} |\hat{\beta}_j - \beta_j^*| \leq \frac{A_0 + 1}{A_0 - 1} \sum_{j \in S} |\hat{\beta}_j - \beta_j^*| < \xi_0 \sum_{j \in S} |\hat{\beta}_j - \beta_j^*|,$$

since the fact that $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$ implies that $\xi_0 > (A_0 + 1)/(A_0 - 1)$. So the compatibility condition implies that

$$\nu_0^2 \left(\sum_{j \in S} |\hat{\beta}_j - \beta_j^*| \right)^2 \leq |S| \left((\hat{\beta} - \beta^*)^T \widetilde{\Sigma} (\hat{\beta} - \beta^*) \right) = |S| \|X\hat{\beta} - X\beta^*\|_n^2,$$

where the last equality follows from the fact that $\widetilde{\Sigma} = X^T X/n$. Therefore,

$$\begin{aligned} \|X\hat{\beta} - X\beta^*\|_n^2 + (A_0 - 1)\rho\|\hat{\beta} - \beta^*\|_1 &\\ &\leq 2A_0\rho \sum_{j \in S} |\hat{\beta}_j - \beta_j^*| \\ &= 2A_0\rho\nu_0^{-1}|S|^{1/2}\|X\hat{\beta} - X\beta^*\|_n \\ &= 2A_0\rho\nu_0^{-1}|S|^{1/2}[\|X\hat{\beta} - X\beta^*\|_n^2 + (A_0 - 1)\|\hat{\beta} - \beta^*\|_1]^{1/2}. \end{aligned}$$

Squaring both sides and rearranging then yields,

$$\begin{aligned}\|X\hat{\beta} - X\beta^*\|_n^2 + (A_0 - 1)\rho\|\hat{\beta} - \beta^*\|_1 &\leq 4A_0^2\rho^2\nu^{-2}|S| \\ &= (4A_0^2\nu_0^{-2})(|S|\rho^2) \\ &= O_p(1)|S|\rho^2.\end{aligned}$$

In particular,

$$\|X\hat{\beta} - X\beta^*\|_n^2 \leq O_p(1)|S|\rho^2$$

Since $\mathrm{P}(\Omega) \geq 1 - 2\delta$ by Lemma 2 of the lectures, this inequality holds with probability $\geq 1 - 2\delta$. Moreover, since $p \geq 1/\delta$ by assumption, we have

$$\begin{aligned}\|X\hat{\beta} - X\beta^*\|_n^2 \leq O_p(1)|S|\rho^2 = O_p(1)|S|\sigma^2\frac{2\log(p/\delta)}{n} &= O_p(1)|S|\frac{\log(p/\delta)}{n} \\ = O_p(1)|S|\frac{\log(p) + \log(1/\delta)}{n} &\leq O_p(1)|S|\frac{2\log(p)}{n} \\ = O_p(1)|S|\frac{\log(p)}{n} &= O_p(1)|S|\lambda_0^2,\end{aligned}$$

where the upper bound $\|X\hat{\beta} - X\beta^*\|_n^2 \leq O_p(1)|S|\lambda_0$ is independent of δ, and thus holds with probability 1.

2. Assume the same conditions as in problem 1. Note that we can write

$$S = \{1 \leq j \leq p : |\beta_j^*| > \lambda_0\} = \{1 \leq j \leq p : |\beta_j^*/\lambda_0|^q > 1\}.$$

Thus,

$$|S| \leq \sum_{j=1}^{p}\left|\frac{\beta_j^*}{\lambda_0}\right|^q = \left\|\frac{\beta^*}{\lambda_0}\right\|_q = \|\beta^*\|_q\lambda_0^{-q}.$$

Therefore, by problem 1,

$$\|X\hat{\beta} - X\beta^*\|_n^2 \leq O_p(1)|S|\lambda_0^2 \leq O_p(1)\|\beta^*\|_q\lambda_0^{2-q}.$$

3. Let β be arbitrary, and assume the same conditions as in problem 1, except this time let $S = \{1 \leq j \leq p : \beta_j \neq 0\}$. Let Ω denote the event

$$\Omega = \left\{\max_{1\leq j\leq p}|\langle Y - m^*(X), X_j\rangle| \leq \rho\right\}.$$

Since the error terms are centered and sub-Gaussian, the proof of Lemma 2 from the lectures shows that $P(\Omega) \geq 1 - 2\delta$. Let β be an arbitrary element of the parameter space. The following extension of the basic inequality to non-linear models was proved in the lectures:

$$\frac{1}{2}\|X\hat{\beta} - m^*(X)\|_n^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + \langle Y - m^*(X), X\hat{\beta} - X\beta\rangle_n + \lambda\|\beta\|_1.$$

Using this inequality in the proof of Lemma 3 from the lectures shows that in the event Ω,

$$\frac{1}{2}\|X\hat{\beta} - m^*(X)\|_n^2 + (A_0 - 1)\rho\|\hat{\beta} - \beta\|_1 \leq \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + 2A_0\rho\sum_{j\in S}|\hat{\beta}_j - \beta|.$$

In particular, it follows that

$$(A_0 - 1)\rho\|\hat{\beta} - \beta\|_1 \leq 2A_0\rho \sum_{j \in S} |\hat{\beta}_j - \beta|,$$

and so it follows from the same manipulations carried out in problem 1 that

$$\sum_{j \notin S} |\hat{\beta}_j - \beta_j| \leq \frac{A_0 + 1}{A_0 - 1} \sum_{j \in S} |\hat{\beta}_j - \beta_j| < \xi_0 \sum_{j \in S} |\hat{\beta}_j - \beta_j|,$$

and so by the compatibility condition,

$$\nu_0^2 \left(\sum_{j \in S} |\hat{\beta}_j - \beta_j| \right)^2 \leq |S| \left((\hat{\beta} - \beta)^T \widetilde{\Sigma} (\hat{\beta} - \beta) \right) = |S| \|X\hat{\beta} - X\beta\|_n^2.$$

Thus,

$$\begin{aligned} \frac{1}{2}\|X\hat{\beta} - m^*(X)\|_n^2 + (A_0 - 1)\rho\|\hat{\beta} - \beta\|_1 &\leq \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + 2A_0\rho \sum_{j \in S} |\hat{\beta}_j - \beta_j| \\ &= \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + 2A_0\rho\nu_0^{-1}|S|^{1/2}\|X\hat{\beta} - X\beta\|_n, \end{aligned}$$

from which it follows that

$$\begin{aligned} \frac{1}{2}\|X\hat{\beta} - m^*(X)\|_n^2 &\leq \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + 4A_0^2\rho^2\nu_0^{-2}|S| \\ &= \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + O_p(1)|S|\rho^2. \end{aligned}$$

Since $\mathrm{P}(\Omega) \geq 1 - 2\delta$, this inequality holds with probability $\geq 1 - 2\delta$. Since $p \geq 1/\delta$ by assumption, the same sort of manipulations carried out in problem 1 show that

$$\frac{1}{2}\|X\hat{\beta} - m^*(X)\|_n^2 \leq \frac{1}{2}\|X\beta - m^*(X)\|_n^2 + O_p(1)|S|\lambda_0^2$$

and hence, multiplying through by 2, we see that

$$\|X\hat{\beta} - m^*(X)\|_n^2 \leq \|X\beta - m^*(X)\|_n^2 + O_p(1)|S|\lambda_0^2$$

with probability 1.

4. (i) An application of the chain rule yields the following:

$$\begin{aligned} \nabla l(\beta) = \frac{\partial}{\partial \beta} l(\beta) = \frac{\partial}{\partial \beta} \left[\frac{1}{n} \sum_{i=1}^n \psi(y_i, x_i^T\beta) \right] &= \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} \psi(y_i, x_i^T\beta) \right] \\ = \frac{1}{n} \sum_{i=1}^n \psi_1(y_i, x_i^T\beta) \frac{\partial}{\partial \beta}(x_i^T\beta) &= \frac{1}{n} \sum_{i=1}^n \psi_1(y_i, x_i^T\beta) x_i. \end{aligned}$$

(ii) Consider the tangent hyperplane of l at β, defined by

$$L(\alpha) = l(\beta) + (\alpha - \beta)^T \nabla l(\beta) \quad \text{for any } \alpha.$$

By the definition of the derivative, it follows that any directional derivative of l at β is equal to that of L. In particular,

$$\begin{aligned}
\lim_{t\to 0} \frac{l(\beta) - l((1-t)\beta + t\beta')}{t} &= \lim_{t\to 0} \frac{l(\beta) - l(\beta - t(\beta - \beta'))}{t} \\
&= \lim_{t\to 0} \frac{L(\beta) - L(\beta - t(\beta - \beta'))}{t} \\
&= \lim_{t\to 0} \frac{l(\beta) - l(\beta) - (\beta - t(\beta - \beta') - \beta)^T \nabla l(\beta)}{t} \\
&= \lim_{t\to 0} \frac{t(\beta - \beta')^T \nabla l(\beta)}{t} \\
&= (\beta - \beta')^T \nabla l(\beta) \\
&= \langle \nabla l(\beta), \beta - \beta' \rangle.
\end{aligned}$$

5. The average log-likelihood l is given by

$$l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T \beta - \tau_0(x_i^T \beta)}{\sigma_i^2} \right] + C,$$

where the term C does not depend upon β. The gradient of l is

$$\nabla l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T - \tau_0'(x_i^T \beta) x_i}{\sigma_i^2} \right] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T - \mathrm{E}(y_i|x_i) x_i}{\sigma_i^2} \right].$$

The Bregman divergence of l is therefore given by

$$\begin{aligned}
D_l(\beta, \beta') &= l(\beta) - l(\beta') - \langle \nabla l(\beta'), \beta - \beta' \rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T \beta - \tau_0(x_i^T \beta)}{\sigma_i^2} - \frac{y_i x_i^T \beta' - \tau_0(x_i^T \beta')}{\sigma_i^2} \right] - (\beta - \beta')^T \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T - \mathrm{E}(y_i|x_i) x_i}{\sigma_i^2} \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i x_i^T (\beta - \beta') - \tau_0(x_i^T \beta) + \tau_0(x_i^T \beta') - (\beta - \beta') y_i x_i + (\beta - \beta')^T \mathrm{E}(y_i|x_i) x_i}{\sigma_i^2} \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathrm{E}(y_i|x_i) x_i^T (\beta - \beta') - \tau_0(x_i^T \beta) + \tau_0(x_i^T \beta')}{\sigma_i^2} \right].
\end{aligned}$$

Let p and p' be two distributions from the underlying exponential family. Then

$$\begin{aligned}
p_i(y_i|x_i) &\propto \exp \left[\frac{y_i x_i^T \beta - \tau_0(x_i^T \beta)}{\sigma_i^2} \right] \\
p_i'(y_i|x_i) &\propto \exp \left[\frac{y_i x_i^T \beta' - \tau_0(x_i^T \beta')}{\sigma_i^2} \right]
\end{aligned}$$

for some β and β'. The Kullback-Leibler divergence of p_i and p_i' is given by

$$\begin{aligned}
\mathrm{KL}(p_i \parallel p_i') &= \sum_j p_i(y_j|x_i) \log \left[\frac{p_i(y_j|x_i)}{p_i'(y_j|x_i)} \right] \\
&= \sum_j p_i(y_j|x_i) \left[\frac{y_j x_i^T \beta - \tau_0(x_i^T \beta) - y_j x_i^T \beta' + \tau_0(x_i^T \beta')}{\sigma_i^2} \right] \\
&= \sum_j p_i(y_j|x_i) \left[\frac{y_j x_i^T (\beta - \beta') - \tau_0(x_i^T \beta) + \tau_0(x_i^T \beta')}{\sigma_i^2} \right] \\
&= \frac{1}{\sigma_i^2} \left[\sum_j p_i(y_j|x_i) y_j x_i^T (\beta - \beta') + \sum_j p_i(y_j|x_i)(-\tau_0(x_i^T \beta) + \tau_0(x_i^T \beta')) \right] \\
&= \frac{E(y_i|x_i) x_i^T (\beta - \beta') - \tau_0(x_i^T \beta) + \tau_0(x_i^T \beta')}{\sigma_i^2}.
\end{aligned}$$

So the Bregman divergence can be written

$$D_l(\beta, \beta') = \frac{1}{n} \sum_{i=1}^n \left[\frac{\mathrm{E}(y_i|x_i) x_i^T (\beta - \beta') - \tau_0(x_i^T \beta) + \tau_0(x_i^T \beta')}{\sigma_i^2} \right] = \frac{1}{n} \sum_{i=1}^n \mathrm{KL}(p_i \parallel p_i').$$