Homework 5

Benjamin Noland

1. The Bregman divergence associated with l is given by

$$D(\beta, \beta') = l(\beta) - l(\beta') - \langle \nabla l(\beta'), \beta - \beta' \rangle,$$

where

$$\nabla l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \psi_1(y_i, x_i^T \beta) x_i.$$

Thus,

$$\begin{aligned}
D(\beta, \beta') + D(\beta', \beta) &= [l(\beta) - l(\beta') - \langle \nabla l(\beta'), \beta - \beta' \rangle] - [l(\beta') - l(\beta) - \langle \nabla l(\beta), \beta' - \beta \rangle] \\
&= \langle \nabla l(\beta'), \beta' - \beta \rangle - \langle \nabla l(\beta), \beta' - \beta \rangle \\
&= \langle \nabla l(\beta') - \nabla l(\beta), \beta' - \beta \rangle \\
&= (\nabla l(\beta') - \nabla l(\beta))^T (\beta' - \beta) \\
&= \frac{1}{n} \sum_{i=1}^{n} [\psi_1(y_i, x_i^T \beta') - \psi_1(y_i, x_i^T \beta)] x_i^T (\beta' - \beta).
\end{aligned}$$

Let $u = x_i^T[\beta + t(\beta' - \beta)]$. Then

$$\begin{aligned}
\frac{\partial}{\partial u} \psi_1(y_i, x_i^T[\beta + t(\beta' - \beta)]) &= \psi_2(y_i, x_i^T[\beta + t(\beta' - \beta)]) \frac{\partial}{\partial t}[x_i^T[\beta + t(\beta' - \beta)]] \\
&= \psi_2(y_i, x_i^T[\beta + t(\beta' - \beta)]) x_i^T (\beta' - \beta).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_0^1 \psi_2(y_i, x_i^T[\beta + t(\beta' - \beta)]) x_i^T (\beta' - \beta)\, dt &= \psi_1(y_i, x_i^T[\beta + t(\beta' - \beta)]) \Big|_0^1 \\
&= \psi_1(y_i, x_i^T \beta') - \psi_1(y_i, x_i^T \beta).
\end{aligned}$$

Hence,

$$\begin{aligned}
D(\beta, \beta') + D(\beta', \beta) &= \frac{1}{n} \sum_{i=1}^{n} [\psi_1(y_i, x_i^T \beta') - \psi_1(y_i, x_i^T \beta)] x_i^T (\beta' - \beta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^1 \psi_2(y_i, x_i^T[\beta + t(\beta' - \beta)]) x_i^T (\beta' - \beta)\, dt \right] x_i^T (\beta' - \beta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^1 \psi_2(y_i, x_i^T[\beta + t(\beta' - \beta)])\, dt \right] (x_i^T (\beta' - \beta))^2.
\end{aligned}$$

2. Define

$$f(x, y; \beta) = \mathrm{P}\{Y = y|x\} = \mathrm{P}\{Y = 1|x\}^y \mathrm{P}\{Y = 0|x\}^{1-y}$$
$$= \left(\frac{e^{x^T\beta}}{1 + e^{x^T\beta}} \right)^y \left(\frac{1}{1 + e^{x^T\beta}} \right)^{1-y},$$

where $y \in \{0, 1\}$. The log-likelihood is therefore given by

$$\begin{aligned}
\log L(\beta; x, y) &= \sum_{i=1}^{n} \log f(x, y; \beta) \\
&= \sum_{i=1}^{n} \left[y_i \left(\frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}} \right) + (1 - y_i) \left(\frac{1}{1 + e^{x_i^T\beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i \left[x_i^T\beta - \log(1 + e^{x_i^T\beta}) \right] - (1 - y_i) \log(1 + e^{x_i^T\beta}) \right] \\
&= \sum_{i=1}^{n} \left[y_i x_i^T\beta - \log(1 + e^{x_i^T\beta}) \right].
\end{aligned}$$

Define

$$\psi(y, u) = -yu + \log(1 + e^u)$$

and

$$l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, x_i^T\beta) = \frac{1}{n} \sum_{i=1}^{n} [-y_i x_i^T\beta + \log(1 + e^{x_i^T\beta})].$$

Then minimizing $l(\beta)$ is equivalent to maximizing $\log L(\beta; x, y)$, and hence $L(\beta; x, y)$. In addition, we have the following:

$$\psi_1(y, u) = -y + \frac{e^u}{1 + e^u}$$

and hence

$$\psi_2(y, u) = \frac{e^u}{(1 + e^u)^2} = \psi_2(y, u') \frac{e^{u-u'}}{(1 + e^u)^2} (1 + e^{u'})^2.$$

When $u' \leq u$, we have

$$\frac{e^{u-u'}}{(1 + e^u)^2} (1 + e^{u'})^2 \leq e^{u-u'} \leq e^{|u'-u|}.$$

Now assume $u' > u$. Then

$$\begin{aligned}
\frac{e^{u-u'}}{(1 + e^u)^2} (1 + e^{u'})^2 &= \frac{(e^{(u-u')/2} + e^{(u+u')/2})^2}{(1 + e^u)^2} \leq \frac{(e^{(u-u')/2} + e^{u'})^2}{(1 + e^u)^2} \\
&\leq \left(e^{(u-u')/2} + \frac{e^{u'}}{1 + e^u} \right)^2 \leq \left(e^{(u-u')/2} + e^{u'-u} \right)^2 \\
&\leq (2e^{u'-u})^2 \leq 4e^{2(u'-u)} = 4e^{2|u'-u|}
\end{aligned}$$

I could not get rid of the leading constant here. If in the case $u' > u$ I could find a bound of the form

$$\frac{e^{u-u'}}{(1 + e^u)^2} (1 + e^{u'})^2 \leq e^{C_0|u'-u|},$$

then taking $C = \max\{1, C_0\}$ would complete the proof.

3. We have the following:

$$D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) = \langle \nabla l(\beta^*) - \nabla l(\hat{\beta}), \beta^* - \hat{\beta} \rangle$$
$$= \frac{1}{n} \sum_{i=1} \left[\frac{e^{x_i^T \beta^*}}{1 + e^{x_i^T \beta^*}} - \frac{e^{x_i^T \hat{\beta}}}{1 + e^{x_i^T \hat{\beta}}} \right] x_i^T (\beta^* - \hat{\beta}),$$

where

$$\nabla l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i x_i + \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} x_i \right].$$

Assumptions:

- $|X_{ij}| \leq D$ for every $1 \leq i \leq n$ and $1 \leq j \leq p$ and some constant $D > 0$.
- $\psi_2(y, u) \leq \psi_2(y, u') e^{c_1 |u - u'|}$ for every y, u, and u', and some constant $c > 0$ (this is satisfied by problem 2).
- The following compatibility condition holds: there exist constants $\nu_0 > 0$ and $\xi_0 > 1$ such that if $b \in \mathbb{R}^p$ satisfies
$$\sum_{j \notin S} |b_j| \leq \xi_0 \sum_{j \in S} |b_j|,$$
then
$$\nu_0^2 \left(\sum_{j \in S} |b_j| \right)^2 \leq |S| (b^T \widetilde{\Sigma}_{\beta^*} b),$$
where
$$\widetilde{\Sigma}_{\beta^*} = \frac{\partial^2}{\partial \beta \partial \beta^T} l(\beta) \Big|_{\beta = \beta^*}.$$
- $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$, and the tuning parameter $\lambda = A_0 \rho$, where $\rho = \sigma \sqrt{2 \log(p/\delta)/n}$ for some $0 < \delta < 1/2$.
- $\psi_1(y_i, x_i^T \beta^*)$ is sub-Gaussian with mean 0 and variance σ^2 for every $1 \leq i \leq n$.

Let Ω denote the event

$$\Omega = \left\{ \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_1(y_i, x_i^T \beta^*) \right| \leq \rho \right\}.$$

Then $\mathrm{P}(\Omega) \geq 1 - 2\delta$ as discussed in the lectures. Then by Lemma 3 with the value of S given in the problem,

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

in the event Ω. In particular,

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

so that upon rearrangement,

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

and thus,

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

The compatibility condition therefore implies that

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

By a result from the lectures,

$$D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) \geq C(\hat{\beta}, \beta^*) (\hat{\beta} - \beta^*)^T \widetilde{\Sigma}_{\beta^*} (\hat{\beta} - \beta^*),$$

where

$$C(\hat{\beta}, \beta^*) = \frac{1 - e^{cD \|\hat{\beta} - \beta^*\|_1}}{cD \|\hat{\beta} - \beta^*\|_1}.$$

Therefore, if we assume that $C(\hat{\beta}, \beta^*) \neq 0$ with probability 1,

$$\begin{aligned} \nu_0^2 \left(\sum_{j \in S} |\hat{\beta}_j - \beta_j^*| \right)^2 &\leq |S| (\hat{\beta} - \beta^*)^T \widetilde{\Sigma}_{\beta^*} (\hat{\beta} - \beta^*) \\ &\leq \frac{1}{C(\hat{\beta}, \beta^*)} [D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta})]. \end{aligned}$$

Thus,

$$\begin{aligned} D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) &\leq 2A_0 \rho \sum_{j \in S} |\hat{\beta}_j - \beta_j^*| \\ &\leq 2A_0 \rho |S|^{1/2} \frac{1}{C(\hat{\beta}, \beta^*)^{1/2}} [D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta})]^{1/2}, \end{aligned}$$

so that upon rearrangement,

$$D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) \leq 4A_0^2 \rho^2 |S| \frac{1}{C(\hat{\beta}, \beta^*)}.$$

If we assume in addition that $1/C(\hat{\beta}, \beta^*) = O_p(1)$ and $p \geq 1/\delta$, then

$$\begin{aligned} D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) &\leq 4A_0^2 \rho^2 |S| \frac{1}{C(\hat{\beta}, \beta^*)} \\ &= 4A_0^2 \frac{2\sigma^2 \log(p/\delta)}{n} |S| \frac{1}{C(\hat{\beta}, \beta^*)} \\ &= 4A_0^2 \frac{2\sigma^2 (\log(p) + \log(1/\delta))}{n} |S| \frac{1}{C(\hat{\beta}, \beta^*)} \\ &\leq 4A_0^2 \frac{4\sigma^2 \log(p)}{n} |S| \frac{1}{C(\hat{\beta}, \beta^*)} \\ &= O_p(1) |S| \lambda_0, \end{aligned}$$

so that

$$D(\hat{\beta}, \beta^*) + D(\beta^*, \hat{\beta}) \leq O_p(1)|S|\lambda_0$$

with probability 1.

5. (i) The Bregman divergence associated with K_n is given by

$$\begin{aligned}
D_K(\beta, \beta') &= K_n(\beta) - K_n(\beta') - \langle \nabla K_n(\beta'), \beta - \beta' \rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-\beta^T x_i} + (1 - y_i)\beta^T x_i] - \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-(\beta')^T x_i} + (1 - y_i)(\beta')^T x_i] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}[-y_i e^{-(\beta')^T x_i} x_i + (1 - y_i)x_i]^T(\beta - \beta') \\
&= \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-\beta^T x_i} - y_i e^{-(\beta')^T x_i} + y_i e^{-(\beta')^T x_i} x_i^T(\beta - \beta')],
\end{aligned}$$

since

$$\nabla K_n(\beta') = \frac{1}{n}\sum_{i=1}^{n}[-y_i e^{-(\beta')^T x_i} x_i + (1 - y_i)x_i].$$

On the other hand, the Bregman divergence associated with l is given by

$$\begin{aligned}
D(\beta, \beta') &= l(\beta) - l(\beta') - \langle \nabla l(\beta'), \beta - \beta' \rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}[-y_i x_i^T\beta + \log(1 + e^{x_i^T\beta})] - \frac{1}{n}\sum_{i=1}^{n}[-y_i x_i^T\beta' + \log(1 + e^{x_i^T(\beta')})] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left[-y_i x_i + \frac{e^{x_i^T\beta'}}{1 + e^{x_i^T\beta'}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{1 + e^{x_i^T\beta}}{1 + e^{x_i^T\beta'}}\right) - \frac{e^{x_i^T\beta'}}{1 + e^{x_i^T\beta'}} x_i^T(\beta - \beta')\right],
\end{aligned}$$

since

$$\nabla l(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i x_i + \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}} x_i\right].$$

There is no discernible relationship between D_K and D_l.

(ii) We have the following:

$$\begin{aligned}
D_K(\hat{\beta}, \beta^*) + D_K(\beta^*, \hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-\hat{\beta}^T x_i} - y_i e^{-(\beta^*)^T x_i} + y_i e^{-(\beta^*)^T x_i} x_i^T(\hat{\beta} - \beta^*)] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-(\beta^*)^T x_i} - y_i e^{-\hat{\beta}^T x_i} + y_i e^{-\hat{\beta}^T x_i} x_i^T(\beta^* - \hat{\beta})] \\
&= \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-(\beta^*)^T x_i} - y_i e^{-\hat{\beta}^T x_i}] x_i^T(\beta^* - \hat{\beta}).
\end{aligned}$$

By the same argument as in problem 3, and under the same conditions, we see that

$$D_K(\hat{\beta}, \beta^*) + D_K(\beta^*, \hat{\beta}) \leq O_p(1)|S|\lambda_0.$$

This is because the argument in problem 3 was done using enough generality to encompass this case as well. However, this lead to some potentially very restrictive assumptions.

(iii) The conditions are the same in both cases, although it is likely the conditions can be *substantially* weakened by exploiting the form of the Bregman divergence in each case.