STATISTICAL LEARNING II Rajen D. Shah
Example Sheet 1 (of 2) 2020

In all of the below, assume that any design matrices X are $n \times p$ and have their columns centred and then scaled to have ℓ_2-norm $\sqrt{n}$, and that any responses $Y \in \mathbb{R}^n$ are centred.

1. Show that if $\lambda \geq \lambda_{\max} := \|X^T Y\|_\infty / n$, then $\hat{\beta}^{\mathrm{L}}_\lambda = 0$.
Solution: We need only check that $\hat{\beta}^{\mathrm{L}}_\lambda = 0$ satisfies the KKT conditions and this is clear. The solution is unique as the fitted values of the Lasso are unique.

2. Show that when the columns of X are orthogonal (so necessarily $p \leq n$) and scaled to have ℓ_2-norm $\sqrt{n}$, the kth component of the Lasso estimator is given by

$$\hat{\beta}^L_{\lambda,k} = (|\hat{\beta}^{\mathrm{OLS}}_k| - \lambda)_+ \mathrm{sgn}(\hat{\beta}^{\mathrm{OLS}}_k)$$

where $(\cdot)_+ = \max(0, \cdot)$. What is the corresponding estimator if the ℓ_1 penalty $\|\beta\|_1$ in the Lasso objective is replaced by the ℓ_0 penalty $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?
Solution: Note that

$$\frac{1}{2n}\|Y - X\beta\|_2^2 = \sum_{k=1}^p \frac{1}{2}(\hat{\beta}^{\mathrm{OLS}}_k - \beta_k)^2 + \frac{1}{2n}\|Y - X\hat{\beta}^{\mathrm{OLS}}\|_2^2.$$

Thus for the first part we need to find the minimiser of

$$\frac{1}{2}(\hat{\beta}^{\mathrm{OLS}}_k - \beta_k)^2 + \lambda|\beta_k|.$$

We write $\hat{\beta}$ for $\hat{\beta}^{\mathrm{L}}_\lambda$ for simplicity. Note that $|\hat{\beta}_k|$ is unique. By the KKT conditions,

$$\hat{\beta}^{\mathrm{OLS}}_k - \hat{\beta}_k = \lambda \hat{\nu}_k$$

where $|\hat{\nu}_k| \leq 1$ and $\hat{\nu}_k = \mathrm{sgn}(\hat{\beta}_k)$ if $\hat{\beta}_k \neq 0$. Thus $\hat{\beta}_k = 0$ when $|\hat{\beta}^{\mathrm{OLS}}_k| \leq \lambda$. If $\hat{\beta}^{\mathrm{OLS}}_k > \lambda$, $\hat{\beta}_k = \hat{\beta}^{\mathrm{OLS}}_k - \lambda$; if $\hat{\beta}^{\mathrm{OLS}}_k < -\lambda$, $\hat{\beta}_k = \hat{\beta}^{\mathrm{OLS}}_k + \lambda$.

Now let $\hat{\beta}$ be the optimising β with the ℓ_0 penalty. Clearly when $(\hat{\beta}^{\mathrm{OLS}}_k)^2/2 < \lambda$, $\hat{\beta}_k = 0$ is optimal. When $(\hat{\beta}^{\mathrm{OLS}}_k)^2/2 = \lambda$, two solutions $\hat{\beta}_k = \hat{\beta}^{\mathrm{OLS}}_k$ or $\hat{\beta}_k = 0$ both minimise the objective. When $(\hat{\beta}^{\mathrm{OLS}}_k)^2/2 > \lambda$ then $\hat{\beta}_k = \hat{\beta}^{\mathrm{OLS}}_k$.

3. Let $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$ and let $S = \{k : \beta^0 \neq 0\}$, $N := \{1, \ldots, p\} \setminus S$. Without loss of generality assume $S = \{1, \ldots, |S|\}$. Assume that X_S has full column rank and let $\Omega = \{\|X^T\varepsilon\|_\infty / n \leq \lambda_0\}$. Show that, when $\lambda > \lambda_0$, if the following two conditions hold

$$\sup_{\tau : \|\tau\|_\infty \leq 1} \|X_N^T X_S (X_S^T X_S)^{-1} \tau\|_\infty < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}$$
$$(\lambda + \lambda_0)\|\{(\tfrac{1}{n} X_S^T X_S)^{-1}\}_k\|_1 < |\beta^0_k| \qquad \text{for } k \in S,$$

then on Ω the (unique) Lasso solution satisfies $\mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_\lambda) = \mathrm{sgn}(\beta^0)$.
Solution: Suppressing the dependence of $\hat{\beta}^{\mathrm{L}}_\lambda$ on λ and dropping the superscript L for ease of notation, we can write the KKT conditions as

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix}\begin{pmatrix} \beta_S - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} + \frac{1}{n}\begin{pmatrix} X_S^T \varepsilon \\ X_N^T \varepsilon \end{pmatrix} = \lambda \begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix}, \tag{1}$$

where $\|\hat{\nu}\|_\infty \leq 1$ and writing $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$, we have $\text{sgn}(\hat{\beta}_{\hat{S}}) = \hat{\nu}_{\hat{S}}$. Now if we do have $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$, it must be the case that (considering the first block of (1)),

$$\frac{1}{n} X_S^T X_S(\beta_S - \hat{\beta}_S) + \frac{1}{n} X_S^T \varepsilon = \lambda \text{sgn}(\beta_S^0),$$

and, substituting this into the second block of (1),

$$X_N^T X_S (X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\} + \tfrac{1}{n} X_N^T \varepsilon = \lambda \hat{\nu}_N.$$

Now we work on Ω and claim that

$$\begin{aligned}(\hat{\beta}_S, \hat{\nu}_S) &= (\beta_S^0 - (\tfrac{1}{n} X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\},\ \text{sgn}(\beta_S^0)),\\ (\hat{\beta}_N, \hat{\nu}_N) &= (0,\ [X_N^T X_S (X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\} + \tfrac{1}{n} X_N^T \varepsilon]/\lambda),\end{aligned}$$

satisfy (1). We first check that $\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta_S^0)$. This holds because

$$\begin{aligned}|[(\tfrac{1}{n} X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\}]_k| &\leq \|\{(\tfrac{1}{n} X_S^T X_S)^{-1}\}_k\|_1 \{\lambda \|\text{sgn}(\beta_S^0)\|_\infty + \|\tfrac{1}{n} X_S^T \varepsilon\|_\infty\}\\ &\leq \|\{(\tfrac{1}{n} X_S^T X_S)^{-1}\}_k\|_1 (\lambda + \lambda_0).\end{aligned}$$

Next

$$\begin{aligned}\|X_N^T X_S (X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\} + \tfrac{1}{n} X_N^T \varepsilon\|_\infty &\leq \|X_N^T X_S (X_S^T X_S)^{-1} \{\lambda \text{sgn}(\beta_S^0) - \tfrac{1}{n} X_S^T \varepsilon\}\|_\infty + \lambda_0\\ &\leq (\lambda + \lambda_0) \sup_{\tau : \|\tau\|_\infty \leq 1} \|X_N^T X_S (X_S^T X_S)^{-1} \tau\|_\infty + \lambda_0\\ &< \lambda,\end{aligned}$$

by the first assumption given in the question. Thus the KKT conditions are satisfied. Because we have the strict inequality $\|\hat{\nu}_N\|_\infty < 1$, S is the equicorrelation set. Since X_S has full column rank, we know the Lasso solution is unique.

4. Find the KKT conditions for the group Lasso.
Solution: For $G \subset \{1, \ldots, p\}$, consider the function $\beta \mapsto \|\beta_G\|_2$. The subdifferential of this function at a β with $\beta_G \neq 0$ is singleton a vector v with $v_{G^c} = 0$ and

$$v_G = \frac{\beta_G}{\|\beta_G\|_2}.$$

We claim that the subdifferential when $\beta_G = 0$ is $\{v : v_{G^c} = 0 \text{ and } \|v_G\|_2 \leq 1\}$. Indeed, if $v_{G^c} \neq 0$ then taking y with $y_G = 0$, $y_{G^c} - \beta_{G^c} = v_{G^c}$, we have

$$0 = \|y_G\|_2 < v^T (y - \beta) = \|v_{G^c}\|_2^2.$$

Now if $v_{G^c} = 0$ and $\|v_G\|_2 \leq 1$, then

$$\|y_G\|_2 \geq \|v_G\|_2 \|y_G\|_2 \geq v^T y.$$

Conversely, if $\|v_G\|_2 > 1$, then taking $y_G = v_G$ (and $y_{G^c} = 0$), we have

$$\|y_G\|_2 < \|v_G\|_2 \|y_G\|_2 = v^T y.$$

Since the subdifferential of a sum of convex functions is the set sum of the subdifferentials of the individual functions, we see that the KKT conditions for the group Lasso objective

$$\frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^q m_j \|\beta_{G_j}\|_2,$$

are that $\hat{\beta}$ is a minimiser if and only if, for $j = 1, \ldots, q$,

$$\frac{1}{n} X_{G_j}^T (Y - X\hat{\beta}) = \lambda m_j \hat{\nu}_{G_j},$$

where $\hat{\nu} \in \mathbb{R}^p$ is such that $\|\hat{\nu}_{G_j}\|_2 \leq 1$, and if $\hat{\beta}_{G_j} \neq 0$ then

$$\hat{\nu}_{G_j} = \frac{\hat{\beta}_{G_j}}{\|\hat{\beta}_{G_j}\|_2}.$$

5. (a) Consider the Lasso and let $\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| = \lambda\}$ be the so-called equicorrelation set at λ. Suppose that $\text{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$ for all $\lambda > 0$. Argue that the Lasso solution is unique for all $\lambda > 0$.
Solution: From the KKT conditions, we know that $\hat{E}_\lambda$ contains $\hat{S}_\lambda := \{j : \hat{\beta}_j^{\text{L}} \neq 0\}$. We know from lectures that $X\hat{\beta}_\lambda^{\text{L}} = X_{\hat{S}_\lambda}\hat{\beta}_{\hat{S}_\lambda}^{\text{L}}$ is unique, but then as $X_{\hat{S}_\lambda}$ has full column rank, we know $\hat{\beta}_{\hat{S}_\lambda}^{\text{L}}$ and hence $\hat{\beta}_\lambda^{\text{L}}$ is unique.

(b) Under the assumptions above, let $\hat{\beta}_{\lambda_1}^{\text{L}}$ and $\hat{\beta}_{\lambda_2}^{\text{L}}$ be two Lasso solutions at different values of the regularisation parameter. Suppose that $\text{sgn}(\hat{\beta}_{\lambda_1}^{\text{L}}) = \text{sgn}(\hat{\beta}_{\lambda_2}^{\text{L}})$. Show that then for all $t \in [0, 1]$,

$$t\hat{\beta}_{\lambda_1}^L + (1-t)\hat{\beta}_{\lambda_2}^L = \hat{\beta}_{t\lambda_1+(1-t)\lambda_2}^L.$$

Hint: Check the KKT conditions.
Solution: We need only check that $t\hat{\beta}_{\lambda_1}^{\text{L}} + (1-t)\hat{\beta}_{\lambda_2}^{\text{L}}$ satisfies the KKT conditions for the Lasso at $t\lambda_1 + (1-t)\lambda_2$. To ease notation, let us write $\hat{\beta}^{(j)} = \hat{\beta}_{\lambda_j}^{\text{L}}$, $j = 1, 2$. Now we know that for $j = 1, 2$,

$$\frac{1}{n} X^T (Y - X\hat{\beta}^{(j)}) = \lambda_j \hat{\nu}^{(j)}$$

where, writing $S = \{k : \hat{\beta}^{(1)} \neq 0\}$, $\hat{\nu}_S^{(1)} = \hat{\nu}_S^{(2)} = \text{sgn}(\hat{\beta}_S^{(1)})$, and $\|\hat{\nu}^{(j)}\|_\infty \leq 1$. Thus

$$\begin{aligned}\frac{1}{n} X^T [Y - X\{t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}\}] &= t\frac{1}{n} X^T (Y - X\hat{\beta}^{(1)}) + (1-t)\frac{1}{n} X^T (Y - X\hat{\beta}^{(2)}) \\ &= t\lambda_1 \hat{\nu}^{(1)} + (1-t)\lambda_2 \hat{\nu}^{(2)}.\end{aligned}$$

Now the indices of the nonzero components of $t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)}$ are S and

$$\text{sgn}(t\hat{\beta}_S^{(1)} + (1-t)\hat{\beta}_S^{(2)}) = \hat{\nu}_S^{(1)} = \hat{\nu}_S^{(2)} = \frac{t\lambda_1 \hat{\nu}_S^{(1)} + (1-t)\lambda_2 \hat{\nu}_S^{(2)}}{t\lambda_1 + (1-t)\lambda_2}.$$

Furthermore, by the triangle inequality,

$$\|t\lambda_1 \hat{\nu}^{(1)} + (1-t)\lambda_2 \hat{\nu}^{(2)}\|_\infty \leq t\lambda_1 \|\hat{\nu}^{(1)}\|_\infty + (1-t)\lambda_2 \|\hat{\nu}^{(2)}\|_\infty \leq t\lambda_1 + (1-t)\lambda_2.$$

Thus the pair

$$\left(t\hat{\beta}^{(1)} + (1-t)\hat{\beta}^{(2)},\ \frac{t\lambda_1 \hat{\nu}^{(1)} + (1-t)\lambda_2 \hat{\nu}^{(2)}}{t\lambda_1 + (1-t)\lambda_2}\right)$$

satisfies the KKT conditions at $t\lambda_1 + (1-t)\lambda_2$.

(c) Conclude that the solution path $\lambda \mapsto \hat{\beta}^L_\lambda$ is piecewise linear with a finite number of knots (points λ where the solution path is not linear at λ) and these occur when the sign of the Lasso solution changes.
Solution: Since there are 3^p sign patterns a Lasso solution can take (each component can be either positive, negative or equal to 0), there are a finite number of knots.

6. When proving the theorems on the prediction error of the Lasso, we started with the so-called basic inequality that

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Show that in fact we can improve this to

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Solution: We start with the KKT conditions for the Lasso

$$\frac{1}{n}X^T(Y - X\hat{\beta}) = \lambda\hat{\nu},$$

where, writing $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$, we have $\text{sgn}(\hat{\beta}_{\hat{S}} = \hat{\nu}_{\hat{S}}$, and also $\|\hat{\nu}\|_\infty \leq 1$. Now we multiply (both sides) by ${\beta^0}^T - \hat{\beta}^T$. Note that $\hat{\beta}^T\hat{\nu} = \hat{\beta}^T_{\hat{S}}\text{sgn}(\hat{\beta}_{\hat{S}}) = \|\hat{\beta}\|_1$. Furthermore, by Hölder's inequality, $|{\beta^0}^T\hat{\nu}| \leq \|\beta^0\|_1\|\hat{\nu}\|_\infty \leq \|\beta^0\|_1$. Substituting $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$ yields

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \frac{1}{n}\varepsilon^T X(\beta^0 - \hat{\beta}) \leq \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Rearranging gives the result.

An alternative solution is as follows. Let Q denote the Lasso objective as in lectures. We have $Q(\hat{\beta}) \leq Q((1-t)\beta^0 + t\hat{\beta})$ for all t. Thus

$$\frac{1}{2n}\|X\beta^0 - X\hat{\beta} + \varepsilon\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|t(X\beta^0 - X\hat{\beta}) - \varepsilon\|_2^2 + t\lambda\|\hat{\beta}\|_1 + (1-t)\|\beta^0\|_1.$$

Dividing by $1 - t$ and rearranging we have

$$\frac{1+t}{2n}\|X\beta^0 - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1$$

for all $t < 1$. Letting $t \uparrow 1$ then gives the result.

7. Under the assumptions of Theorem 21 on the prediction and estimation properties of the Lasso under a compatibility condition, show that, with probability $1 - 2p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{9A^2\log(p)}{4\phi^2}\frac{\sigma^2 s}{n}.$$

Solution: We follow the proof of Theorem 21, but starting with the improved "basic inequality" in the previous question. We arrive at

$$\frac{2}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}_N - \beta^0_N\|_1 \leq 3\lambda\|\hat{\beta}_S - \beta^0_S\|_1.$$

Using the compatibility condition, the RHS is at most

$$\|\hat{\beta}_S - \beta^0_S\|_1 \leq \frac{\sqrt{s}\|X(\beta^0 - \hat{\beta})\|_2/\sqrt{n}}{\phi}.$$

Substituting this into the previous inequality, we get

$$\frac{2}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}_N - \beta^0_N\|_1 \leq \frac{3\lambda\sqrt{s}\|X(\beta^0 - \hat{\beta})\|_2/\sqrt{n}}{\phi}$$

whence

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{9\lambda^2 s}{4\phi^2}$$

as required.

8. (a) Show that

$$\max_{\theta:\|X^T\theta\|_\infty \leq \lambda} G(\theta) = \frac{1}{2n}\|Y - X\hat{\beta}^{\mathrm{L}}_\lambda\|_2^2 + \lambda\|\hat{\beta}^{\mathrm{L}}_\lambda\|_1,$$

where

$$G(\theta) = \frac{1}{2n}\|Y\|_2^2 - \frac{1}{2n}\|Y - n\theta\|_2^2.$$

Show that the unique θ maximising G is $\theta^* = (Y - X\hat{\beta}^{\mathrm{L}}_\lambda)/n$. *Hint: Treat the Lasso optimisation problem as minimising* $\|Y - z\|_2^2/(2n) + \lambda\|\beta\|_1$ *subject to* $z - X\beta = 0$ *over* $(\beta, z) \in \mathbb{R}^p \times \mathbb{R}^n$ *and consider the Lagrangian.*

Solution: Taking the hint, we write the Lagrangian for the Lasso problem

$$L(\beta, z, \theta) = \frac{1}{2n}\|Y - z\|_2^2 + \lambda\|\beta\|_1 + \theta^T(z - X\beta).$$

The minimising β and z, β^* and z^* satisfy

$$\frac{1}{n}(Y - z^*) = \theta,$$
$$\lambda\nu^* = X^T\theta,$$

provided θ is such that $\|\nu^*\|_\infty \leq 1$ so $\|X^T\theta\|_\infty \leq \lambda$. Substituting into the Lagrangian and using the fact that $\beta^{*T}X^T\theta = \lambda\beta^{*T}\nu^* = \lambda\|\beta\|_1$ we get

$$L(\beta^*, z^*, \theta) = \frac{n}{2}\|\theta\|_2^2 + \theta^T(Y - n\theta) = G(\theta),$$

provided $\|X^T\theta\|_\infty \leq \lambda$. Thus we have that

$$\max_{\theta:\|X^T\theta\|_\infty \leq \lambda} G(\theta) \leq \frac{1}{2n}\|Y - X\hat{\beta}_\lambda\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_1.$$

To get equality we take $\theta = \theta^* = (Y - X\hat{\beta}_\lambda)/n$ (dropping the superscript L for the Lasso solution for clarity) for then

$$G(\theta^*) = \frac{1}{2n}\|Y - X\hat{\beta}_\lambda\|_2^2 + \frac{1}{n}\hat{\beta}_\lambda^T X^T(Y - X\hat{\beta}^{\mathrm{L}}_\lambda),$$

the final term equalling $\lambda\|\hat{\beta}_\lambda\|_1$ by the KKT conditions. Uniqueness of the maximiser follows from the facts that $-G$ is strictly convex and $\{\theta : \|X^T\theta\|_\infty \leq \lambda\}$ is a convex set.

(b) Let $\tilde{\theta}$ be such that $\|X^T\tilde{\theta}\|_\infty \leq \lambda$. Explain why if

$$\max_{\theta: G(\theta) \geq G(\tilde{\theta})} |X_k^T\theta| < \lambda,$$

then we know that $\hat{\beta}^{\mathrm{L}}_{\lambda,k} = 0$. By considering $\tilde{\theta} = Y\lambda/(n\lambda_{\max})$ with $\lambda_{\max} = \|X^TY\|_\infty/n$, show that $\hat{\beta}^{\mathrm{L}}_{\lambda,k} = 0$ if

$$\frac{1}{n}|X_k^TY| < \lambda - \frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}.$$

Solution: θ^* must be in the set $\{\theta : G(\theta) \geq G(\tilde{\theta})\}$ so if the inequality in the question is true, then we know $|X_k^T(Y - X\hat{\beta}_\lambda)|/n < \lambda$ whence by the KKT conditions for the Lasso, $\hat{\beta}_{\lambda,k}$ must be zero. With the given choice of $\tilde{\theta}$, we know $\|X^T\tilde{\theta}\|_\infty \leq \lambda$.
We now need to show that when

$$\frac{1}{n}|X_k^TY| < \lambda - \frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$

and $G(\theta) \geq G(Y\lambda/(n\lambda_{\max}))$, we have $|X_k^T\theta| < \lambda$. Note the condition $G(\theta) \geq G(Y\lambda/(n\lambda_{\max}))$ is equivalent to

$$\|Y - n\theta\|_2 \leq (1 - \lambda/\lambda_{\max})\|Y\|_2.$$

Now, under the conditions above,

$$\begin{aligned}
|X_k^T\theta| &= |X_k^T(\theta - Y/n + Y/n)| \\
&\leq |X_k^T(\theta - Y/n)| + |X_k^TY|/n \\
&< \|X_k\|_2(1 - \lambda/\lambda_{\max})\|Y\|_2/n + \lambda - \frac{\|Y\|_2}{\sqrt{n}}\frac{\lambda_{\max} - \lambda}{\lambda_{\max}} \\
&= \lambda,
\end{aligned}$$

using the fact that $\|X_k\|_2 = \sqrt{n}$ to get the final equality.

9. The elastic net estimator in the linear model minimises

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2/2)$$

over $\beta \in \mathbb{R}^p$, where $\alpha \in [0, 1]$ is fixed.

(a) Suppose X has two columns X_j and X_k that are identical and $\alpha < 1$. Explain why the minimising β^* above is unique and has $\beta^*_k = \beta^*_j$.
Solution: The minimum is unique as the objective above is strictly convex, and existence can be shown via the same argument used to show the existence of Lasso solutions. Suppose then that β^* is the unique minimiser. Let $\beta' = \beta^*$ in all components except $\beta'_j = \beta^*_k$ and $\beta'_k = \beta^*_j$. The objective is strictly convex in β so $\beta'/2+\beta^*/2$ has an objective value at least as large small as that of β^*, so $\beta' = \beta^*$ by uniqueness.

(b) Let $\hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \ldots$ be the solutions from iterations of a coordinate descent procedure to minimise the elastic net objective. For a fixed variable index k, let $A = \{1, \ldots, k-1\}$ and $B = \{k+1, \ldots, p\}$. Show that for $m \geq 1$,

$$\hat{\beta}^{(m)}_k = \frac{S_{\lambda\alpha}\Big(n^{-1}X_k^T(Y - X_A\hat{\beta}^{(m)}_A - X_B\hat{\beta}^{(m-1)}_B)\Big)}{1 + \lambda(1 - \alpha)},$$

where $S_t(u) = \text{sgn}(u)(|u| - t)_+$ is the soft-thresholding operator.
Solution: We have that

$$\hat{\beta}_k^{(m)} = \underset{\beta \in \mathbb{R}}{\text{argmin}} \{\|Y - X_A\hat{\beta}_A^{(m)} - X_B\hat{\beta}_B^{(m-1)} - \beta X_k\|_2^2/(2n) + \lambda(\alpha|\beta| + (1-\alpha)\beta^2/2)\}$$

The minimiser $\hat{\beta}_k^{(m)}$ must satisfy the subgradient optimality condition:

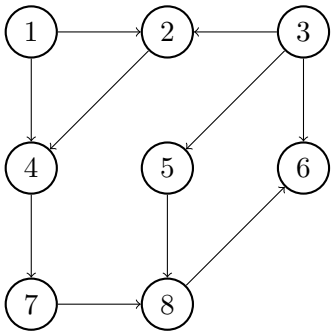
$$-\frac{1}{n}X_k^T(Y - X_A\hat{\beta}_A^{(m)} - X_B\hat{\beta}_B^{(m-1)}) + \hat{\beta}_k^{(m)} + \lambda(1-\alpha)\hat{\beta}_k^{(m)} + \lambda\alpha\hat{\nu} = 0,$$

where $\hat{\nu} \in [-1, 1]$ and if $\hat{\beta}_k^{(m)} \neq 0$, $\hat{\nu} = \text{sgn}(\hat{\beta}_k^{(m)})$. Rearranging, we have

$$\hat{\beta}_k^{(m)} = \frac{\frac{1}{n}X_k^T(Y - X_A\hat{\beta}_A^{(m)} - X_B\hat{\beta}_B^{(m-1)}) - \lambda\alpha\hat{\nu}}{1 + \lambda(1-\alpha)},$$

and we may check that the given expression for $\hat{\beta}_k^{(m)}$ satisfies this. Note that $\hat{\beta}_k^{(m)}$ is the unique minimiser as the objective is strictly convex.

10. For the following DAG $\mathcal{G}$



write down

(a) the descendants of 3;
Solution: $\{2, 4, 5, 6, 7, 8\}$.

(b) all sets of variables that d-separate 1 and 3;
Solution: Note neither 2 nor any of its descendants can be in a d-separating set. Thus we may take $\emptyset, \{5\}$.

(c) all sets of variables that d-separate $\{1, 4\}$ and 6;
Solution: $\{3\} \cup \{\text{at least one of } \{7, 8\}\} \cup \{\text{any subset of } \{2, 5\}\}$.

(d) all the v-structures.
Solution: $1 \to 2 \leftarrow 3$, $8 \to 6 \leftarrow 3$, $7 \to 8 \leftarrow 5$.

11. Let $Z = (Z_1, \ldots, Z_p)^T \in \{0, 1\}^p$ be a binary random vector with probability mass function given by

$$\mathbb{P}(Z_1 = z_1, \ldots, Z_p = z_p) = \exp\left(\Theta_{00} + \sum_{k=1}^{p}\Theta_{0k}z_k + \sum_{k=1}^{p}\sum_{j=1}^{k-1}\Theta_{jk}z_jz_k - \Phi(\Theta)\right)$$

where $\exp(-\Phi(\Theta))$ is a normalising constant. Show that

$$\text{logit}(\mathbb{P}(Z_k = 1 | Z_{-k} = z_{-k})) = \Theta_{0k} + \sum_{j:j<k}\Theta_{jk}z_j + \sum_{j:j>k}\Theta_{kj}z_j,$$

where $\text{logit}(q) = \log\{q/(1-q)\}$ for $q \in (0,1)$. Conclude that, for $j < k$,

$$Z_j \perp\!\!\!\perp Z_k | Z_{-jk} \iff \Theta_{jk} = 0.$$

Note that for discrete random variables we can replace the densities in our definition of conditional independence with probability mass functions (which are in any case densities with respect to counting measure). How might we go about estimating the Θ_{jk}?
Solution: The result follows from noting that

$$\text{logit}(\mathbb{P}(Z_k = 1 | Z_{-k} = z_{-k})) = \log\{\mathbb{P}(Z_k = 1, Z_{-k} = z_{-k})\} - \log\{\mathbb{P}(Z_k = 0, Z_{-k} = z_{-k})\}.$$

Now iff. $\Theta_{jk} = 0$, the distribution of $Z_k | Z_{-k}$ does not depend on Z_j, and so

$$Z_j \perp\!\!\!\perp Z_k | Z_{-jk} \iff \Theta_{jk} = 0.$$

We can try to estimate the Θ_{jk} by logistic regression.