

# **DSO Screening Exam: 2018 In-Class Exam**

Gregory Faletto

**Exercise 1 (Probability).** (a) (i) We have

$$e^{-y}T(y) \xrightarrow{d} \text{Exponential}(\lambda) \iff \lim_{y \rightarrow \infty} \Pr(e^{-y}T(y) \leq t) = 1 - e^{-\lambda t}$$

$$\iff \lim_{y \rightarrow \infty} \Pr\left(\frac{\inf\{x \geq 0 : M(x) \geq y\}}{e^y} \leq t\right) = 1 - \frac{1}{e^{\lambda t}}$$

(ii)

(b)

$$X \sim \text{Bin}(n, p)$$

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \sum_{x=0}^{\infty} \frac{1}{1+x} \cdot \Pr(X=x) = \sum_{x=0}^n \frac{1}{1+x} \cdot \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \sum_{x=0}^n \frac{1}{1+x} \cdot \binom{n}{x} \left(\frac{p}{1-p}\right)^x$$

Recall that

$$\mathbb{E}(X) = np \iff \sum_{x=0}^n x \cdot \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \sum_{x=0}^n x \binom{n}{x} \left(\frac{p}{1-p}\right)^x = np.$$

⋮

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \int_0^{\infty} \Pr([X+1]^{-1} > x) dx = \int_0^1 \Pr\left(\frac{1}{X+1} > x\right) dx = \int_0^1 \Pr\left(X < \frac{1}{x} - 1\right) dx$$

$$= \lim_{a \rightarrow 0^+} \int_a^1 \Pr(X < \lceil x^{-1} - 1 \rceil) dx$$

$$= \lim_{a \rightarrow 0^+} \int_a^{(n+1)^{-1}} \Pr(X < \lceil x^{-1} - 1 \rceil) dx + \int_{(n+1)^{-1}}^1 \Pr(X < \lceil x^{-1} - 1 \rceil) dx$$

$$= \frac{1}{n+1} \cdot 1 + \int_{(n+1)^{-1}}^1 \Pr(X < \lceil x^{-1} - 1 \rceil) dx$$

This is simply the area under  $n$  rectangles with heights  $1, \Pr(X \leq n-1), \Pr(X \leq n-2), \dots, \Pr(X=0)$  and bases  $(n+1)^{-1}, n^{-1} - (n+1)^{-1}, (n-1)^{-1} - n^{-1}, \dots, 1/2$ . That is, we can write this as

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1}{n+1} \cdot 1 + \left(\frac{1}{n} - \frac{1}{n+1}\right) \cdot \Pr(X \leq n-1) + \left(\frac{1}{n-1} - \frac{1}{n}\right) \cdot \Pr(X \leq n-2)$$

$$+ \dots + \left(\frac{1}{2} - \frac{1}{3}\right) \cdot \Pr(X \leq 1) + \frac{1}{2} \cdot \Pr(X=0)$$

⋮

Recall that

$$\mathbb{E}(X) = np \iff \sum_{x=0}^{n-1} \sum_{j=x+1}^n \binom{n}{j} p^j (1-p)^{n-j} = (1-p)^n \sum_{x=0}^{n-1} \sum_{j=x+1}^n \binom{n}{j} \left(\frac{p}{1-p}\right)^j = np.$$

**Exercise 2 (Mathematical Statistics).** (a) Let

$$T_n(X_1, \dots, X_n) := \sum_{i=1}^n X_i.$$

(b) Note that if  $T_n(X_1, \dots, X_n) = y$ , then  $\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n)] = 0$  unless  $\sum_{i=1}^n x_i = y$ . Therefore we have

$$\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n) \mid T_n(X_1, \dots, X_n) = y] = \frac{\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n)]}{\Pr(T_n(X_1, \dots, X_n) = y)}$$

$$\frac{\prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}$$

which does not depend on  $p$ . That is, the distribution of  $X_1, \dots, X_n$  conditional on  $T_n(X_1, \dots, X_n)$  is independent of  $p$ . Therefore  $T_n(X_1, \dots, X_n)$  is sufficient for  $p$ .

(c) Likelihood function:

$$\begin{aligned} \mathcal{L}(p) &= \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} \\ \implies \ell(p) &= \sum_{i=1}^n [X_i \log(p) + (1-X_i) \log(1-p)] = \log(p) \sum_{i=1}^n X_i + \log(1-p) \left( n - \sum_{i=1}^n X_i \right) \\ \implies \frac{d}{dp} \ell(p) &= \frac{1}{p} \sum_{i=1}^n X_i - \frac{1}{1-p} \left( n - \sum_{i=1}^n X_i \right) = 0 \implies \frac{1}{p} \sum_{i=1}^n X_i = \frac{1}{1-p} \left( n - \sum_{i=1}^n X_i \right) \\ &\iff \sum_{i=1}^n X_i - p \sum_{i=1}^n X_i = pn - p \sum_{i=1}^n X_i \iff \boxed{\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i.} \end{aligned}$$

We can find its asymptotic distribution using the Central Limit Theorem:

**Theorem 1. Central Limit Theorem (Grimmett and Stirzaker theorem 5.10.4.)** Let  $X_1, X_2, \dots$  be a sequence of independent identically distributed random variables with finite mean  $\mu$  and finite non-zero variance  $\sigma^2$ , and let  $S_n = \sum_{i=1}^n X_i$ . Then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Since  $\mathbb{E}(X_i) = p$ ,  $\text{Var}(X_i) = p(1-p)$ , we have

$$\begin{aligned} \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} &\xrightarrow{d} \mathcal{N}(0, 1) \iff \sum_{i=1}^n X_i - np \xrightarrow{d} \mathcal{N}(0, np(1-p)) \\ &\iff \frac{1}{n} \sum_{i=1}^n X_i - p \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{n}\right) \iff \boxed{\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)} \end{aligned}$$

**Exercise 3 (Mathematical Statistics).** (a) We have

$$X \mid \mu \sim \mathcal{N}(\mu, \mathbf{I}_n)$$

Let  $X = (X_1, \dots, X_n)^T$  and let  $\mu = (\mu_1, \dots, \mu_n)^T$ . Notice that

$$\begin{aligned} \mathbb{E}(X^T X \mid \mu) &= \mathbb{E}(X_1^2 + X_2^2 + \dots + X_n^2) = \sum_{i=1}^n \mathbb{E}(X_i^2) = \sum_{i=1}^n (\text{Var}(X_i) + \mathbb{E}(X_i)^2) = \sum_{i=1}^n (1 + \mu_i^2) \\ &= n + \|\mu\|_2^2 \implies \mathbb{E}(X^T X - n \mid \mu) = \|\mu\|_2^2 \end{aligned}$$

Therefore given  $\mu$ ,  $\boxed{X^T X - n}$  is unbiased for  $\|\mu\|_2^2$ .

(b) We have

$$\mu \sim \mathcal{N}(0, k\mathbf{I}_n)$$

$$\mathbb{E} \left[ (\|\mu\|_2^2 - t(X))^2 \mid X \right] = \mathbb{E} [\|\mu\|_2^4 - 2\|\mu\|_2^2 t(X) + t(X)^2 \mid X] = \mathbb{E} [\|\mu\|_2^4 \mid X] - 2t(X)\mathbb{E} [\|\mu\|_2^2 \mid X] + t(X)^2$$

The estimator minimizing this is  $t(X) = \mathbb{E} [\|\mu\|_2^2 \mid X] = \mathbb{E} [\mu^T \mu \mid X]$ , which we need to find. We have that  $\mu \sim \mathcal{N}(0, k\mathbf{I}_n)$ , so

$$f_{\mu^T \mu}(t) = f_{\sum_{i=1}^n \mu_i^2}(t) = f_{\sum_{i=1}^n \left[\frac{\mu_i}{k}\right]^2} \left( \frac{t}{k^2} \right) = f_{\chi_n^2} \left( \frac{t}{k^2} \right)$$

where  $f_{\chi_n^2}$  is the density of a  $\chi^2$  random variable with  $n$  degrees of freedom.

$$\implies \Pr(\mu^T \mu = t \mid X = x) = \frac{\Pr(\mu^T \mu = t \cap X = x)}{\Pr(X = x)} = \frac{f_{\mu^T \mu}(t) f_{\mu \mid \mu^T \mu}(m \mid t) \Pr(X = x \mid \mu = m)}{\Pr(X = x)}$$

$$\implies \mathbb{E}(\mu^T \mu \mid X) = \int_0^\infty t \Pr(\mu^T \mu = t \mid X) dt$$

$\vdots$

Given  $\mu$ , we have

$$X \mid \mu \sim \mathcal{N}(\mu, \mathbf{I}_n) \implies (X - \mu)^T (X - \mu) \mid \mu \sim \chi_n^2 \iff X^T X - 2\mu^T X + \mu^T \mu \mid \mu \sim \chi_n^2.$$

$\vdots$

Also, we have that  $\mu \sim \mathcal{N}(0, k\mathbf{I}_n)$ . The joint distribution of  $X$  and  $\mu$  is then

$$f_{X, \mu}(x, m) = f_{X \mid \mu=m}(x \mid m) f_{\mu}(m) =$$

(c)

(d)

**Exercise 4 (High-Dimensional Statistics).** (a)

(b)

(c)

(d)

**Exercise 5 (Optimization).** (a) We can express the original optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (1)$$

as

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 \\ & \text{subject to} \quad z = X\beta. \end{aligned} \quad (2)$$

We will also refer to another expression of the lasso optimization problem,

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - X\beta\|_2^2 \\ & \text{subject to} \quad \|\beta\|_1 \leq t \end{aligned} \quad (3)$$

for some  $t > 0$ . The Lagrangian of (2) is

$$\mathcal{L}(\beta, z, \nu) = \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + \nu^T (z - X\beta),$$

so the Lagrange dual function is

$$\begin{aligned} \inf_{\beta, z} \{\mathcal{L}(x, \nu)\} &= \inf_{\beta, z} \left\{ \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + \nu^T (z - X\beta) \right\} \\ &= \inf_{\beta, z} \left\{ \frac{1}{2} (y - z)^T (y - z) + \nu^T z + \lambda \|\beta\|_1 - \nu^T X\beta \right\} \end{aligned}$$

This minimization is separable:

$$= \inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + \nu^T z \right\} + \inf_{\beta} \{ \lambda \|\beta\|_1 - \nu^T X\beta \} \quad (4)$$

We will handle each part of (4) separately. First, the left side:

$$\inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + \nu^T z \right\} = \inf_z \left\{ \frac{1}{2} z^T z + (\nu - y)^T z + \frac{1}{2} y^T y \right\}$$

Since this is a convex quadratic form, differentiate with respect to  $z$  and set equal to zero:

$$z + (\nu - y) = 0 \implies z = y - \nu \quad (5)$$

$$\begin{aligned}
&\Rightarrow \inf_z \left\{ \frac{1}{2} z^T z + (\nu - y)^T z + \frac{1}{2} y^T y \right\} = \frac{1}{2} (y - \nu)^T (y - \nu) + (\nu - y)^T (y - \nu) + \frac{1}{2} y^T y \\
&= \frac{1}{2} (y^T y - 2\nu^T y + \nu^T \nu) + 2\nu^T y - y^T y - \nu^T \nu + \frac{1}{2} y^T y = -\frac{1}{2} \nu^T \nu + \nu^T y = \frac{1}{2} y^T y - \frac{1}{2} y^T y + \nu^T y - \frac{1}{2} \nu^T \nu \\
&= \frac{1}{2} y^T y - \frac{1}{2} (y^T y - 2\nu^T y + \nu^T \nu) = \frac{1}{2} y^T y - \frac{1}{2} (y - \nu)^T (y - \nu) = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - \nu\|_2^2
\end{aligned}$$

Next we will minimize the right side of (4):

$$\begin{aligned}
\inf_{\beta} \{ \lambda \|\beta\|_1 - \nu^T X \beta \} &= \inf_{\beta} \left\{ \lambda \sum_{i=1}^p |\beta_i| - \sum_{i=1}^p [\nu^T X]_i \beta_i \right\} = \inf_{\beta} \left\{ \sum_{i=1}^p \left( \lambda |\beta_i| - [\nu^T X]_i \beta_i \right) \right\} \\
&= \inf_{\beta} \left\{ \sum_{i=1}^p \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\nu^T X]_i \right) \beta_i \right\} = \sum_{i=1}^p \inf_{\beta_i} \left\{ \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\nu^T X]_i \right) \beta_i \right\}
\end{aligned}$$

where  $\mathbb{I}\{\beta_i < 0\}$  is an indicator function. Notice that when  $\beta_i$  is negative, if  $\left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\nu^T X]_i \right) = -\left( \lambda + [\nu^T X]_i \right)$  is positive there is no lower bound on the quantity we are minimizing; otherwise, when  $\beta_i$  is negative the infimum is 0. When  $\beta_i$  is positive, if  $\left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\nu^T X]_i \right) = \left( \lambda - [\nu^T X]_i \right)$  is negative there is no lower bound on the quantity we are minimizing; otherwise, when  $\beta_i$  is negative the infimum is 0. That is, the only dual feasible points satisfy for all  $i$

$$-\left( \lambda + [\nu^T X]_i \right) \leq 0, \quad \lambda - [\nu^T X]_i \geq 0 \iff [\nu^T X]_i \geq -\lambda, \quad [\nu^T X]_i \leq \lambda$$

which is equivalent to the condition

$$\|\nu^T X\|_{\infty} \leq \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta, z} \{ \mathcal{L}(x, \nu) \} = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - \nu\|_2^2 \quad (6)$$

subject to the constraint

$$\|\nu^T X\|_{\infty} \leq \lambda.$$

This quantity represents a lower bound on the minimum value of the original optimization problem for all  $\nu \in \mathbb{R}^p$ . The dual problem is to find the best lower bound by maximizing over  $\nu$ ; that is, the dual problem is

$$\begin{aligned}
&\underset{\nu \in \mathbb{R}^p}{\text{maximize}} && \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - \nu\|_2^2 \\
&\text{subject to} && \|\nu^T X\|_{\infty} \leq \lambda.
\end{aligned}$$

**still need to finish; actually trivial based on (5):** Lastly, suppose  $\hat{\beta}$  and  $\hat{\nu}$  satisfy

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad ,$$

$$\hat{\nu} = \underset{\nu \in \mathbb{R}^p}{\arg \max} \quad \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - \nu\|_2^2 \quad = \quad \underset{\nu \in \mathbb{R}^p}{\arg \min} \quad -\frac{1}{2} \|y\|_2^2 + \frac{1}{2} \|y - \nu\|_2^2$$

$$\text{subject to} \quad \|\nu^T X\|_\infty \leq \lambda \quad \quad \quad \text{subject to} \quad \|\nu^T X\|_\infty \leq \lambda$$

⋮

- (b) (i) **Not necessarily unique.** Per Tibshirani [2013], if  $\text{rank}(X) < p$ , the lasso solution is not necessarily unique. Intuitively, this is because the columns of  $X$  are linearly dependent, so there may exist more than one linear combination of the columns that minimizes (1). **Jacob suggestion: counterexample.  $X$  is two columns that are equal; then convex combinations of two solutions are equal as long as same sign (can't be opposite sign because then  $\ell_1$  could be smaller by setting one equal to 0.**
- (ii) **Necessarily unique. flesh out more** By a result from part (a),  $\hat{u} = y - X\hat{\beta}$ . By part (iii), even though  $\hat{\beta}$  is not unique,  $X\hat{\beta}$  is (see also Lemma 1 in Tibshirani [2013]). Therefore  $\hat{u}$  is unique. **Jacob's solution: strictly convex optimization problem, so argument maximizing is unique. dual is always convex;**
- (iii) **Necessarily unique** (except in the trivial case  $\lambda = 0$ ). Per part 5(b)(iv),  $\|\hat{\beta}\|_1$  is unique. (1) is convex, so the minimum  $\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1$  is unique. Therefore  $\|y - X\hat{\beta}\|_2^2$  must be unique. **Jacob's solution: if  $\hat{\nu}$  is unique then its  $\ell_2$  norm is unique.**
- (iv) **Necessarily unique change this argument: if you solve 1 and take the norm of that and choose that value for  $t$  then you will get the same solution. because if there existed a better solution then it would have made the objective function 1 better.** (except in the trivial case  $\lambda = 0$ ). Whenever  $\lambda > 0$ , the lasso objective function (1) is the dual of (3) with  $t$  less than the  $\ell_1$  norm of the OLS solution (if it exists). (3) is convex and Slater's condition holds because every point  $\{\beta \in \mathbb{R}^p \mid \|\beta\|_1 < t\}$  is feasible. Therefore for this dual function, strong duality holds; that is, the optimal values of (1) and (3) are equal. Further, they are optimized over the same variable and they are both convex, so the solution set of (1) is identical to the solution set of (3).  
 Since the objective function of (3) is continuous and the feasible region  $\|\beta\|_1 \leq t$  is compact, a minimum is guaranteed to exist. Since (3) is convex and the global minimum lies outside the region for  $t < t_0$ , the minimum will lie on the boundary; that is,  $\|\hat{\beta}\|_1 = t$  for (3) and therefore for (1). See Osborne et al. [2000] for details.
- (c) (i) Since  $\beta^*$  is clearly feasible for (1) and  $\hat{\beta}$  achieves the minimum, we have

$$\frac{1}{2} \|y - X\beta^*\|_2^2 + \lambda \|\beta^*\|_1 \geq \frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \iff \frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2} \|\varepsilon\|_2^2 + \lambda \|\beta^*\|_1$$

- (ii) From part (a), since the optimal value of the dual (6) is a lower bound for the optimal value of the primal (1), we have

$$\frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - \hat{\nu}\|_2^2 \leq \frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1. \quad (7)$$

**Jacob's solution: plug in  $\varepsilon$  for  $\nu$  in the lower bound equation, then you get  $X\beta^*$**

so we are done if  $\|X\beta^*\|_2^2 \geq \|y - \hat{\nu}\|_2^2$ . Also by part (a),  $\hat{\nu} = y - X\hat{\beta}$ , so  $\|y - \hat{\nu}\|_2^2 = \|X\hat{\beta}\|_2^2$ , so we are done if  $\|X\beta^*\|_2^2 \geq \|X\hat{\beta}\|_2^2$ . From part (c)(i), we have

$$\frac{1}{2} \left[ (y - X\hat{\beta})^T (y - X\hat{\beta}) - (y - X\beta^*)^T (y - X\beta^*) \right] \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff 2y^T X(\beta^* - \hat{\beta}) + (\hat{\beta})^T X^T X \hat{\beta} - (\beta^*)^T X^T X \beta^* \leq 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \quad (8)$$

So (8) is a sufficient condition for the result. Examining the right side of the right side of (8), we have

$$2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) = 2y^T(y - \varepsilon - X\hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \quad (9)$$

Borrowing notation from Osborne et al. [2000], note that the subdifferential of (1) is given by

$$\begin{aligned} \partial_\beta \mathcal{L}(\beta, \lambda) &= \partial_\beta \left( \frac{1}{2}(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_1 \right) = \partial_\beta \left( \frac{1}{2}(y^T y - 2y^T X\beta + \beta^T X^T X\beta) + \lambda\|\beta\|_1 \right) \\ &= -y^T X + X^T X\beta + \lambda v = -X^T(y - X\beta) + \lambda v \end{aligned}$$

where  $v = (v_1, \dots, v_p)^T$  with  $v_i = \text{sgn}(\beta_i)$  if  $\beta_i \neq 0$  and  $v_i \in [-1, 1]$  if  $\beta_i = 0$ . By the convexity of (1), for  $\hat{\beta}$  minimizing (1) it must hold that

$$\begin{aligned} 0 = -X^T(y - X\hat{\beta}) + \lambda v &\iff \lambda v^T \hat{\beta} = (y - X\hat{\beta})^T X \hat{\beta} \iff \lambda = \frac{(y - X\hat{\beta})^T X \hat{\beta}}{\|\hat{\beta}\|_1} \\ &\iff \|\hat{\beta}\|_1 = \frac{(y - X\hat{\beta})^T X \hat{\beta}}{\lambda} \end{aligned} \quad (10)$$

where we used  $v^T \hat{\beta} = \|\hat{\beta}\|_1$ . Substituting (10) into (9), we have

$$\begin{aligned} &2y^T(y - \varepsilon - X\hat{\beta}) - 2\lambda \left( \|\beta^*\|_1 - \frac{(y - X\hat{\beta})^T X \hat{\beta}}{\lambda} \right) \\ &= 2y^T y - 2y^T \varepsilon - 2y^T X\hat{\beta} - 2\lambda\|\beta^*\|_1 + 2y^T X\hat{\beta} - 2\hat{\beta}^T X^T X \hat{\beta} = 2y^T(y - \varepsilon) - 2\lambda\|\beta^*\|_1 - 2\hat{\beta}^T X^T X \hat{\beta} \\ &= 2y^T X\beta^* - 2\lambda\|\beta^*\|_1 - 2\|X\hat{\beta}\|_2^2 = 2(X\beta^* + \varepsilon)^T X\beta^* - 2\lambda\|\beta^*\|_1 - 2\|X\hat{\beta}\|_2^2 \\ &= 2\|X\beta^*\|_2^2 + 2\varepsilon^T X\beta^* - 2\lambda\|\beta^*\|_1 - 2\|X\hat{\beta}\|_2^2 = 2 \left( \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \right) - 2(\lambda\|\beta^*\|_1 - \varepsilon^T X\beta^*) \\ &= 2 \left( \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \right) - 2 \sum_{i=1}^p \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i \right) \beta_i^* \end{aligned}$$

Because  $\lambda \geq \|X^T \varepsilon\|_\infty$  (that is, for all  $i \in \{1, \dots, p\}$ ,  $\lambda - [\varepsilon^T X]_i \geq 0$ ) we have that for all  $i \in \{1, \dots, p\}$ ,

$$\left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i \right) \beta_i^* \geq 0$$

by the following argument. If  $\beta_i^* = 0$ , the result is trivial. If  $\beta_i^* > 0$ , we have



$$(-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i = \lambda - [\varepsilon^T X]_i \geq \lambda - |X^T \varepsilon|_i \geq 0$$

$$\implies \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i \right) \beta_i^* \geq 0.$$

Lastly, if  $\beta_i^* < 0$ , we have

$$(-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i = - \left( \lambda + [\varepsilon^T X]_i \right) \leq - \left( \lambda - |X^T \varepsilon|_i \right) \leq 0$$

$$\implies \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i \right) \beta_i^* \geq 0.$$

Therefore we have

$$2 \left( \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \right) - 2 \sum_{i=1}^p \left( (-1)^{\mathbb{I}\{\beta_i < 0\}} \lambda - [\varepsilon^T X]_i \right) \beta_i^* \leq 2 \left( \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \right)$$

$\vdots$

$$\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2(X\beta^* + \varepsilon)^T X\beta^* - 2y^T X\hat{\beta} - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2\|X\beta^*\|_2^2 + 2\varepsilon^T X\beta^* - 2y^T X\hat{\beta} - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

(iii) **haven't finished, but this is just a bonus question** We already have from part (c)(i)

$$\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2} \|\varepsilon\|_2^2 + \lambda \|\beta^*\|_1$$

Note that

$$\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\beta^*\|_1$$

(iv) **haven't finished, but this is just a bonus question**

(d) *Proof.*

**Definition 1 (Convex function in  $\mathbb{R}^n$ ).** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . We say that  $f$  is **convex** if, for any  $x, y \in \mathbb{R}^n$  and for any  $t \in [0, 1]$ , we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \quad (11)$$

Note that

$$\|tx + (1-t)y\|_1 \leq \|tx\|_1 + \|(1-t)y\|_1 = t\|x\|_1 + (1-t)\|y\|_1 \quad (12)$$

where the first step follows by the Triangle Inequality (which all norms satisfy, including the  $\ell_1$  norm) and the second step follows by the homogeneity property of norms. Therefore  $\|\theta\|_1$  is convex. Next, by (12) and the monotonicity of  $g(\theta) = \theta^2$  when  $\theta \geq 0$ ,

$$\begin{aligned} f(tx + (1-t)y) &= (\|tx + (1-t)y\|_1)^2 \leq (t\|x\|_1 + (1-t)\|y\|_1)^2 \\ &= t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1 \end{aligned}$$

and

$$tf(x) + (1-t)f(y) = t\|x\|_1^2 + (1-t)\|y\|_1^2$$

Taking the difference of these yields

$$\begin{aligned} tf(x) + (1-t)f(y) - f(tx + (1-t)y) &\geq t\|x\|_1^2 + (1-t)\|y\|_1^2 - (t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1) \\ &= (t-t^2)\|x\|_1^2 + [(1-t) - (1-t)^2]\|y\|_1^2 - 2t(1-t)\|x\|_1\|y\|_1 \\ &= (t-t^2)(\|x\|_1^2 + \|y\|_1^2 - 2\|x\|_1\|y\|_1) = t(1-t)(\|x\|_1 - \|y\|_1)^2 \geq 0 \\ &\iff tf(x) + (1-t)f(y) \geq f(tx + (1-t)y) \end{aligned}$$

which proves convexity. □

## References

- M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. ISSN 1537-2715. doi: 10.1080/10618600.2000.10474883. URL <https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20>.
- R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.