Modern Statistical Methods
Revision questions

1. $Y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p} \quad \text{rank}(X) = p$

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbb{1} \qquad \mathbb{E}(\varepsilon) = 0 \quad \text{Var}(\varepsilon) = \sigma^2 I$$
$$\bar{\varepsilon} = \mathbb{1}^T\varepsilon/n$$

$$\hat{\beta}^{OLS} := \arg\min_\beta \|Y - X\beta\|_2^2$$

$$\frac{\partial}{\partial \beta^T}(Y - X\beta)^T(Y - X\beta) = -2X^TY + 2X^TX\beta = 0$$

$$\Rightarrow \hat{\beta}^{OLS} = (X^TX)^{-1}X^TY \qquad (X^TX)^{-1} \text{ exists as } X \text{ full column rank}$$

$$\hat{\beta}^R_\lambda := \arg\min_\beta \left(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right)$$

$$\frac{\partial}{\partial \beta^T}\left((Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta\right) = -2X^TY + 2X^TX\beta + 2\lambda\beta = 0$$

$$\Rightarrow \hat{\beta}^R_\lambda = (X^TX + \lambda I)^{-1}X^TY \qquad \text{inverse always exists}$$

$$\mathbb{E}\{(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T\} - \mathbb{E}\{(\hat{\beta}^R_\lambda - \beta^0)(\hat{\beta}^R_\lambda - \beta^0)^T\} \quad \text{is pos. def.?}$$

$$\hat{\beta}^{OLS} - \beta^0 = (X^TX)^{-1}X^T(X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbb{1}) - \beta^0 = (X^TX)^{-1}X^T(\varepsilon - \bar{\varepsilon}\mathbb{1})$$

$$\mathbb{E}\{(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T\} = \mathbb{E}\{(X^TX)^{-1}X^T(\varepsilon - \bar{\varepsilon}\mathbb{1})(\varepsilon - \bar{\varepsilon}\mathbb{1})^TX(X^TX)^{-1}\}$$

$$= \sigma^2(X^TX)^{-1} \qquad \text{as } X^T\mathbb{1} = 0 \ (X \text{ cols centred})$$

$$\hat{\beta}_\lambda^R - \beta^0 = (X^TX + \lambda I)^{-1} X^T(X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbb{1}) - \beta^0$$

$$= (X^TX + \lambda I)^{-1}(X^TX\beta^0 + \lambda I\beta^0 - \lambda I\beta^0 + X^T\varepsilon - \bar{\varepsilon}X^T\mathbb{1}) - \beta^0$$

$$= (X^TX + \lambda I)^{-1}(-\lambda\beta^0 + X^T\varepsilon)$$

$$\mathbb{E}\{(\hat{\beta}_\lambda^R - \beta^0)(\hat{\beta}_\lambda^R - \beta^0)^T\} = \mathbb{E}\{(X^TX + \lambda I)^{-1}(-\lambda\beta^0 + X^T\varepsilon)(\varepsilon^TX - \lambda\beta^{0T})(X^TX + \lambda I)^{-T}\}$$

$$= (X^TX + \lambda I)^{-1}\lambda^2\beta^0\beta^{0T}(X^TX + \lambda I)^{-1} + \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$$

$$= (X^TX + \lambda I)^{-1}(\lambda^2\beta^0\beta^{0T} - \sigma^2\lambda I)(X^TX + \lambda I)^{-1} + \sigma^2(X^TX + \lambda I)^{-1}$$

$$\mathbb{E}\{(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T\} - \mathbb{E}\{(\hat{\beta}_\lambda^R - \beta^0)(\hat{\beta}_\lambda^R - \beta^0)^T\}$$

$$= \sigma^2(X^TX)^{-1} - \sigma^2(X^TX + \lambda I)^{-1} + \lambda(X^TX + \lambda I)^{-1}(\sigma^2 I - \lambda\beta^0\beta^{0T})(X^TX + \lambda I)^{-1}$$

$$= (X^TX + \lambda I)^{-1}\left((X^TX + \lambda I)\sigma^2(X^TX)^{-1} - \sigma^2\right) + \dots$$

$$= (X^TX + \lambda I)^{-1}\left(\sigma^2\lambda(X^TX)^{-1}(X^TX + \lambda I)\right)(X^TX + \lambda I)^{-1} + \dots$$

$$= \lambda(X^TX + \lambda I)^{-1}\left(\sigma^2\left[I + \lambda(X^TX)^{-1} + I\right] - \lambda\beta^0\beta^{0T}\right)(X^TX + \lambda I)^{-1}$$

$$= \lambda(X^TX + \lambda I)^{-1}\left[\sigma^2\{2I + \lambda(X^TX)^{-1}\} - \lambda\beta^0\beta^{0T}\right](X^TX + \lambda I)^{-1}$$

Since $\lambda > 0$, $(X^TX + \lambda I)^T = (X^TX + \lambda I) \neq 0$, then this is positive definite iff

$\sigma^2\{2I + \lambda(X^TX)^{-1}\} - \lambda\beta^0\beta^{0T} \succ 0$ which is always true for sufficiently small $\lambda > 0$.

$X^TX \succeq 0$ by pos def of Euclidean norm

$X^TX + \lambda I \succ 0$ for $\lambda > 0$

$A \succ 0 \Leftrightarrow A^{-1} \succ 0$ from eigenvalues

Therefore,

$$\mathbb{E}\{(x^{*T}\hat{\beta}_\lambda^R - x^{*T}\beta^0)^2\} < \mathbb{E}\{(x^{*T}\hat{\beta}^{OLS} - x^{*T}\beta^0)^2\} \quad \text{for all } x^* \in \mathbb{R}^p$$

(and indeed all $\|x^*\|_2 = 1$)

$$\lambda(X^TX+\lambda I)^{-1}\{\sigma^2(2I+\lambda(X^TX)^{-1})-\lambda\beta^0\beta^{0T}\}(X^TX+\lambda I)^{-1} = M$$

Want to show that $\exists x^*, \beta^0$ with

$$-x^{*T} M x^* > \delta$$

Take $x^* = \dfrac{(X^TX+\lambda I)\beta^0}{\|(X^TX+\lambda I)\beta^0\|_2}$

Then (let c_1 be the max e-val)

$$-x^{*T} M x^* = \lambda \frac{\lambda\|\beta^0\|_2^4 - \beta^{0T}(\sigma^2 2I + \sigma^2\lambda(X^TX)^{-1})\beta^0}{\|(X^TX+\lambda I)\beta^0\|_2^2}$$

(let c_2 be the max e-val)

$$\geq \lambda \frac{\lambda\|\beta^0\|_2^2(\|\beta^0\|_2^2 - c_1)}{\|\beta^0\|_2^2 c_2} > \delta \quad \text{for some } \beta^0$$

2. $FWER = \mathbb{P}(N_{01} \geq 1)$ $\quad N_{01}$ - rejected true hypotheses

Bonferroni correction: reject H_i if $p_i \leq \alpha/m$.

Then

$FWER = \mathbb{P}(N_{01} \geq 1) \leq \mathbb{E}(N_{01})$ Markov inequality

$= \mathbb{E}\left(\sum_{i \in I_0} \mathbb{1}_{\{p_i \leq \alpha/m\}}\right)$ where I_0 are true null hypotheses, $|I_0| = m_0$

$= \sum_{i \in I_0} \mathbb{E}\left(\mathbb{1}_{\{p_i \leq \alpha/m\}}\right)$

$= \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha/m) \leq \sum_{i \in I_0} \frac{\alpha}{m} = \frac{\alpha m_0}{m} \leq \alpha$

Intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ (all $H_i : i \in I$ are true)

Closure of $\{H_i\}_{i=1}^m$ is $\{H_I : I \subseteq \{1, \dots, m\}, I \neq \emptyset\}$

Closed testing procedure:

Reject H_I iff $\forall J \subseteq I$

H_J is rejected by the local test ϕ_J

Local test ϕ_I: α-level test for each I taking values in $\{0, 1\}$, under H_I

$\mathbb{P}_{H_I}(\phi_I = 1) \leq \alpha$

If I_0 are true null hypotheses, in order for any false rejection H_{I_0} must have been rejected

$FWER = \mathbb{P}(N_{01} \geq 1) = \mathbb{P}(\text{at least one false rejection}) \leq \mathbb{P}(\phi_{I_0} = 1) \leq \alpha$.

$H_I : I \in \mathcal{I}$, $\forall I, J \in \mathcal{I}$ either $I \cap J = \emptyset$, $I \subseteq J$ or $J \subseteq I$

p_I

$p_I^{adj} := \max_{J : J \in \mathcal{I}, J \supseteq I} \frac{m}{|J|} p_J$: Reject H_I : $p_I^{adj} \leq \alpha$. Prob at least $1-\alpha$, no false rejections

$FWER = \mathbb{P}(N_{01} \geq 1) \leq \sum_{I \in \mathcal{I}} \mathbb{P}(p_I^{adj} \leq \alpha)$ or otherwise all $J \subseteq I$ not rejected

$= \sum_{I \in \mathcal{I} : \nexists J \supseteq I} \mathbb{P}\left(\frac{m}{|I|} p_I \leq \alpha\right) \leq \sum_{I \in \mathcal{I} : \nexists J \supseteq I} \frac{\alpha |I|}{m} = \alpha \frac{\sum |I|}{m} \leq \alpha$.

3. $n, p > 1$, $k \in \{1, \dots, p\}$, $j \in \{1, \dots, n\}$

$Z \in \mathbb{R}^p$, $Z_{-k} \in \mathbb{R}^{p-1}$ $\quad$ $X \in \mathbb{R}^{n \times p}$, X_k kth col, $X_{-j,k}$ kth col, jth row removed

$Z \sim N_p(\mu, \Sigma)$, $\Sigma > 0$

CIG is the undirected graph satisfying the pairwise Markov property w.r.t. Z

Z satisfies the pairwise Markov property w.r.t. $\mathcal{G}$ if for any pair $j, k \in V$ with $j \neq k$ and $(j,k), (k,j) \notin E$

$$Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk} .$$

$Z \in \mathbb{R}^p$

$$Z_k \mid Z_{-k} = z_{-k}$$

Consider general $Z_A \mid Z_B = z_B$. Write

$$Z_A = M Z_B + (Z_A - M Z_B) \text{ s.t.}$$

$$\mathrm{Cov}(Z_B, Z_A - M Z_B) = \Sigma_{B,A} - \Sigma_{B,B} M^T = 0$$

$$\Rightarrow M^T = \Sigma_{B,B}^{-1} \Sigma_{B,A} \quad \text{i.e.} \quad M = \Sigma_{A,B} \Sigma_{B,B}^{-1}$$

Then $Z_B \perp\!\!\!\perp Z_A - M Z_B$ and

$$E(Z_A - M Z_B) = \mu_A - \Sigma_{A,B} \Sigma_{B,B}^{-1} \mu_B$$

$$\mathrm{Var}(Z_A - M Z_B) = \Sigma_{A,A} + \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,B} \Sigma_{B,B}^{-1} \Sigma_{B,A} - 2 \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,A}$$

$$= \Sigma_{A,A} - \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,A}$$

$$\therefore Z_A \mid Z_B = z_B \sim N_{|A|}\left(\mu_A + \Sigma_{A,B} \Sigma_{B,B}^{-1} (z_B - \mu_B), \Sigma_{A,A} - \Sigma_{A,B} \Sigma_{B,B}^{-1} \Sigma_{B,A}\right)$$

$$\Rightarrow Z_k \mid Z_{-k} = z_{-k} \sim N\left(\mu_k + \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} (z_{-k} - \mu_{-k}), \Sigma_{k,k} - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \Sigma_{-k,k}\right)$$

Nodewise regression given data X,

conditioning on $Z_{-k} = z_{-k}$ gives

$$Z_k = m_k + z_{-k}^T \Sigma_{-k,-k}^{-1} \Sigma_{-k,k} + \varepsilon_k \qquad (*)$$

where $m_k = \mu_k - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \mu_{-k}$

$$\varepsilon_k \mid Z_{-k} = z_{-k} \sim N(0, \Sigma_{k,k} - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \Sigma_{-k,k})$$

In nodewise regression we estimate $\Sigma_{-k,-k}^{-1} \Sigma_{-k,k}$ for each k by regressing X_k on X_{-k}.

$$\frac{1}{2n} \sum_{k=1}^{p} \| X_k - \mu_k \mathbb{1} - X_{-k} \Theta_{-k,k} \|_2^2 + \lambda \sum_{j<k} \sqrt{\Theta_{jk}^2 + \Theta_{kj}^2}$$

estimates the intercept and coefficient vector. The penalty is similar to Lasso but the ℓ_2-norm for symmetric terms prefers Θ to be symmetric.

$$\Theta_{-k,k} = \Sigma_{-k,-k}^{-1} \Sigma_{-k,k}$$

Hence

$$(\Theta_{-k,k})_j = 0 \iff \begin{cases} \Omega_{j,k} = 0 & : j < k \\ \Omega_{j+1,k} = 0 & : j \geq k \end{cases}$$

and

$$\Omega = \Sigma^{-1}, \qquad Z_k \perp\!\!\!\perp Z_j \mid Z_{-j,k} \iff \Omega_{jk} = 0$$

Allowing to estimate CIG.

$$\Sigma = \begin{pmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{pmatrix} \succ 0 \qquad S = \Sigma_{k,k} - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \Sigma_{-k,k} \quad \text{Schur comp.}$$

$$S \succ 0$$

$$\Omega = \Sigma^{-1} = \begin{pmatrix} S^{-1} & -S^{-1} \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \\ -\Sigma_{-k,-k}^{-1} \Sigma_{-k,k} S^{-1} & \Sigma_{-k,-k}^{-1} + \Sigma_{-k,-k}^{-1} \Sigma_{-k,k} S^{-1} \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \end{pmatrix}$$

$\therefore \Omega_{k,k}^{-1} = S = \Sigma_{k,k} - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \Sigma_{-k,k}$, $\Omega_{k,-k} - \Omega_{k,k} \Omega_{k,k}^{-1} = \Sigma_{-k,-k}^{-1} \Sigma_{-k,k} = \Theta_{-k,k}$

and $\Omega_{k,k} \in \mathbb{R} \setminus \{0\}$ as $S \succ 0$.

4. $Y \in \mathbb{R}^n$ $X \in \mathbb{R}^{n \times p}$ $G_1, \dots, G_q$ partition of $\{1, \dots, p\}$

Group Lasso, $\lambda > 0$, $m_1, \dots, m_q > 0$ weights

$$Q_\lambda(\beta) = \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{q} \|\beta_{G_i}\|_2 m_i$$

Multipliers balance different group sizes.

Penalty encourages either an entire group G_j to have $\hat{\beta}_{G_j} = 0$ or $\hat{\beta}_k \neq 0 \ \forall k \in G_j$.

Useful for categorical predictors or when expanding predictors using basis functions.

Block-wise coordinate descent: Starting with a guess $\beta^{(0)}$, solve for $m \geq 1$:

$$\beta_{G_1}^{(m)} = \underset{\beta_{G_1} \in \mathbb{R}^{|G_1|}}{\operatorname{argmin}} Q_\lambda(\beta_{G_1}, \beta_{G_2}^{(m-1)}, \dots, \beta_{G_q}^{(m-1)})$$

$$\vdots$$

$$\beta_{G_j}^{(m)} = \underset{\beta_{G_j} \in \mathbb{R}^{|G_j|}}{\operatorname{argmin}} Q_\lambda(\beta_{G_1}^{(m)}, \dots, \beta_{G_j}, \dots, \beta_{G_q}^{(m-1)})$$

As other penalties do not change, for rth block the problem is

$$\min_b \left\{ \frac{1}{2n} \|Y - X_{G_r^c} \beta_{G_r^c}^* - X_{G_r} b\|_2^2 + \lambda m_r \|b\|_2 \right\}$$

$$= \min_b \left\{ \frac{1}{2} \|w - Ab\|_2^2 + \eta \|b\|_2 \right\}$$

where $w = \frac{1}{\sqrt{n}}(Y - X_{G_r^c} \beta_{G_r^c}^*)$ and $\beta_{G_r^c}^* = \begin{cases} \beta_{G_j}^{(m)} & : m_j < r \\ \beta_{G_j}^{(m-1)} & : j > r \end{cases}$

$$A = \frac{1}{\sqrt{n}} X_{G_r}, \quad \eta = \lambda m_r$$

$$\frac{\partial}{\partial b^T} \left\{ \frac{1}{2}(w - Ab)^T(w - Ab) + \eta \sqrt{b^T b} \right\} = -A^T w + A^T A b + \frac{\eta b}{\sqrt{b^T b}} = 0$$

$$\Rightarrow b^* = (\theta^* I + A^T A)^{-1} A^T w \quad \text{where} \quad \theta^* = \frac{\eta}{\|b^*\|_2} > 0.$$

$A = UDV^T$, $\gamma = DU^T w$

Rewriting from above

$$\eta^2 = w^T A (\theta^* I + A^T A)^{-3} A^T w \, \theta^{*3}$$

$$\eta^2 = w^T UDV^T (\theta^* I + VD^T U^T UDV^T)^{-3} VD^T U^T w \, \theta^{*3}$$

$$\eta^2 = \gamma^T V^T (\theta^* I + V D^2 V^T)^{-3} V \gamma \, \theta^{*3}$$

$$\eta^2 = \gamma^T V^T (I + \theta^{*-1} V D^2 V^T)^{-3} V \gamma$$

$$\eta^2 = \gamma^T (I + \theta^{*-1} D^2)^{-3} \gamma$$

$$\Rightarrow \sum_k \frac{\gamma_k^2}{(D_{kk}^2 + \theta^*)^3} = \frac{\eta^2}{\theta^{*3}}$$

$$b^* = (\theta^* I + VD^2 V^T)^{-1} V \gamma$$

$$= (V(\theta^* I + D^2) V^T)^{-1} V \gamma$$

$$= V (\theta^* I + D^2)^{-1} \gamma \qquad \theta^* = \frac{\eta}{\|b^*\|_2}$$

$$\|b^*\|_2^2 = \frac{\eta^2}{\theta^*} = \gamma^T (\theta^* I + D^2)^{-2} \gamma$$

$$= \sum_k \frac{\gamma_k^2}{(\theta^* + D_{kk}^2)^2}$$

5. $FDP = \frac{N_{01}}{\max(R,1)}$ $\quad \mathbb{E}(FDP) = FDR$ $\quad R = \max\{i : p_{(i)} \le \frac{i\alpha}{m}\}$

Reject $H_{(1)}, \dots, H_{(R)}$

All p-vals independent

$p_1, \dots, p_{m_0} \sim U[0,1]$

$$\{p_1 \le \alpha r/m \,,\, R = r\} = \{p_1 \le \tfrac{\alpha r}{m} \,,\, p_{(r)} \le \tfrac{\alpha r}{m} \,,\, p_{(s)} > \tfrac{\alpha s}{m} \ \forall s > r\}$$

Let $p'_{(1)}, \dots, p'_{(m-1)}$ be the order stats of $p_2, \dots, p_m$

Let $R' = \max\{i : p'_{(i)} \le \frac{(i+1)\alpha}{m}\}$

$$\dots = \{p_1 \le \tfrac{\alpha r}{m} \,,\, R' = r-1\}$$

$p''_{(1)}, \dots, p''_{(m-2)}$ be the order stats $p_3, \dots, p_m$

$R'' = \max\{i : p''_{(i)} \le \frac{(i+2)\alpha}{m}\}$

$$\{p_1 \le \tfrac{\alpha r}{m} \,,\, p_2 \le \tfrac{\alpha r}{m} \,,\, R = r\} = \{p_1 \le \tfrac{\alpha r}{m} \,,\, p_2 \le \tfrac{\alpha r}{m} \,,\, R'' = r-2\}$$

$$\mathbb{E}\left((FDP)^2\right) = \sum_{r=1}^{m} \mathbb{E}\left(\frac{N_{01}^2}{r^2} \mathbb{1}_{\{R=r\}}\right)$$

$$= \sum_{r=1}^{m} \frac{1}{r^2} \mathbb{E}\left(\left(\sum_{j=1}^{m_0} \mathbb{1}_{\{p_j \le \frac{\alpha r}{m} \,,\, R=r\}}\right)^2\right)$$

$$= \sum_{r=1}^{m} \frac{1}{r^2} \left(\sum_{j=1}^{m_0} \mathbb{P}\left(p_j \le \tfrac{\alpha r}{m} \,,\, R' = r-1\right) + 2 \sum_{j=1}^{m_0} \sum_{k=1}^{j-1} \mathbb{P}\left(p_j \le \tfrac{\alpha r}{m} \,,\, p_k \le \tfrac{\alpha r}{m} \,,\, R'' = r-2\right) \right)$$

independent units

$$= \sum_{r=1}^{m} \frac{1}{r^2} \left(\frac{\alpha r}{m} m_0 \mathbb{P}(R' = r-1) + \frac{m_0(m_0-1)\alpha^2}{m^2} \mathbb{P}(R'' = r-2) \right)$$

$$= \frac{\alpha m_0}{m} \mathbb{E}\left(\frac{1}{R'+1}\right) + \frac{\alpha^2 m(m_0-1)}{m^2}$$

6. $Y \in \mathbb{R}^n \qquad X \in \mathbb{R}^{n \times p}$

$$Y = X\beta^0 + z - \bar{z}\mathbb{1} \qquad \varepsilon \sim N_n(0, \sigma^2 I)$$

Lasso estimator

$$\hat{\beta}^L_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

$S = \{k \in \{1, \dots, p\} : \beta^0_k \neq 0\}$, $N = \{1, \dots, p\} \setminus S$, $s = |S|$

Arbitrary $A \subseteq \{1, \dots, p\}$, $b \in \mathbb{R}^p$, $b_A \in \mathbb{R}^{|A|}$

Assume $\exists\, \phi > 0$ s.t.

$\forall b \in \mathbb{R}^p$ with $\|b_N\|_1 \leq 3\|b_S\|_1$ have

$$\|b_S\|_1^2 \leq \frac{s\|Xb\|_2^2}{n\phi^2}$$

Show:

If $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$ for some $A > 0$, then w. prob at least $1 - 2p^{-(A^2/8 - 1)}$

have $\frac{1}{n}\|X(\hat{\beta}^L_\lambda - \beta^0)\|_2^2 + \lambda\|\hat{\beta}^L_\lambda - \beta^0\|_1 \leq \frac{16A^2}{\phi^2} \frac{\sigma^2 s \log p}{n}$.

Basic inequality

$$\frac{1}{2n}\|Y - X\hat{\beta}^L_\lambda\|_2^2 + \lambda\|\hat{\beta}^L_\lambda\|_1 \leq \frac{1}{2n}\|Y - X\beta^0\|_2^2 + \lambda\|\beta^0\|_1$$

$$-\frac{1}{n} Y^T X\hat{\beta}^L_\lambda + \frac{1}{2n}\|X\hat{\beta}^L_\lambda\|_2^2 + \cancel{\frac{\|\varepsilon\|_2^2}{2n}} + \frac{1}{n} Y^T X\beta^0 - \frac{1}{2n}\|X\beta^0\|_2^2 \leq \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}^L_\lambda\|_1$$

$$\frac{1}{2n}(X\hat{\beta}^L_\lambda - 2Y)^T X\hat{\beta}^L_\lambda - \frac{1}{2n}(X\beta^0 - 2Y)^T X\beta^0 \leq \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}^L_\lambda\|_1$$

$$\frac{1}{2n}(X\hat{\beta}^L_\lambda - 2X\beta^0 - 2\varepsilon + \cancel{2\varepsilon\mathbb{1}}^{\,0})^T X\hat{\beta}^L_\lambda - \frac{1}{2n}(-X\beta^0 - 2\varepsilon)^T X\beta^0 \leq \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}^L_\lambda\|_1$$

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

Work on the event $\Omega = \{2\|X^T\varepsilon\|_\infty/n \le \lambda\}$

Hölder inequality $\Rightarrow \varepsilon^T X(\hat\beta^L_\lambda - \beta^0) \le \|X^T\varepsilon\|_\infty \|\hat\beta^L_\lambda - \beta^0\|_1$

$$\Rightarrow \frac{1}{n}\|X(\hat\beta^L_\lambda - \beta^0)\|_2^2 + 2\lambda\|\hat\beta^L_\lambda\|_1 \le \lambda\|\hat\beta^L_\lambda - \beta^0\|_1 + 2\lambda\|\beta^0\|_1$$

Now $\frac{1}{n}\|X(\hat\beta^L_\lambda - \beta^0)\|_2^2 \le 3\lambda\|\hat\beta^L_\lambda - \beta^0\|_1$ and

$$a + 2\|\hat\beta\|_1 \le \|\hat\beta - \beta^0\|_1 + 2\|\beta^0\|_1 \qquad \text{where } a = \frac{1}{n\lambda}\|X(\hat\beta^L_\lambda - \beta^0\|\|_2^2$$

$$a + 2(\|\hat\beta_N\|_1 + \|\hat\beta_S\|_1) \le \|\hat\beta_S - \beta^0_S\|_1 + \|\hat\beta_N - \beta^0_N\|_1 + 2\|\beta^0_N\| + 2\|\beta^0_S\|_1$$

$$\le \|\hat\beta_S - \beta^0_S\|_1 + \|\hat\beta_N\|_1 + 2\|\beta^0_S\|_1$$

$$a + \|\hat\beta_N\|_1 \le \|\hat\beta_S - \beta^0_S\|_1 + 2\|\beta^0_S\| - 2\|\hat\beta_S\|_1$$

$$a + \|\hat\beta_N - \beta^0_N\|_1 \le 3\|\hat\beta_S - \beta^0_S\|_1$$

$$a + \|\hat\beta - \beta^0\|_1 \le 4\|\hat\beta_S - \beta^0_S\|_1$$

$$\Rightarrow \frac{1}{n}\|X(\hat\beta^L_\lambda - \beta^0)\|_2^2 + \lambda\|\hat\beta^L_\lambda - \beta^0\|_1 \le 4\lambda\|\hat\beta^L_{\lambda,S} - \beta^0_S\|_1$$

Applying the assumption to $\hat\beta_S - \beta^0$, $\le \frac{4\lambda}{\phi}\sqrt{\frac{s}{n}}\|X(\hat\beta^L_\lambda - \beta^0)\|_2$

and $\frac{1}{\sqrt{n}}\|X(\hat\beta^L_\lambda - \beta^0)\|_2 \le \frac{4\lambda}{\phi}\sqrt{s}$

$$\therefore \frac{1}{n}\|X(\hat\beta^L_\lambda - \beta^0)\|_2^2 + \lambda\|\hat\beta^L_\lambda - \beta^0\|_1 \le \frac{16\lambda^2 s}{\phi^2}$$ the required result.

Finally $\mathbb{P}(\Omega_0) \ge 1 - 2p^{-(A^2/8-1)}$ as $\mathbb{P}(2\|X^T\varepsilon\|_\infty/n > \lambda) \le \sum_{j=1}^p \mathbb{P}(2|X_j^T\varepsilon|/n > \lambda)$

$X_j^T\varepsilon/n$ sub-Gaussian v.p. $(\sigma^2\|X_j\|_2^2/n^2)^{1/2} = \sigma/n$ $\le 2p^{-(A^2/8-1)}$