36-755, Fall 2017
Homework 3 Solution

Due Wed Oct 18 by 5:00pm in Jisu's mailbox

Points: 100+5 pts total for the assignment.

1. **Median and sample quantiles.**

(a) Suppose that $(X_1, \ldots, X_n)$ is an i.i.d. sample from a distribution P (if you like, you may assume P to be absolutely continuous). Let $X_{(1)} \leq X_{(2)} < \ldots < X_{(n)}$ be the order statistics and set $\alpha \in (0, 1)$. Determine a $1 - \alpha$ confidence interval for the median of P of the form

$$\left(X_{(k_1)}, X_{(k_2)}\right)$$

for some choice of $k_1 < k_2$. Determine k_1 and k_2 by relating this problem to a $\mathrm{Bin}(n, 1/2)$ distribution and use concentration.

(b) **Bonus problem (it means this is optional).** Letting m be the median of P, assumed unique for convenience. Assume also that there exists an $\eta > 0$ such that the c.d.f. F of P is differentiable at all $x \in I = (m - \eta, m + \eta)$, with $\inf_{x \in I} F'(x) \geq C > 0$. Compute a high probability bound on the length of the confidence interval found in the previous point. You may use the following result, known as the DKW inequality. If $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution over the real line with c.d.f. F and F_n denotes the corresponding empirical c.d.f., then

$$\mathbb{P}\left(\|F - F_n\|_\infty \geq t\right) \leq 2e^{-2nt^2}.$$

What happens when η or C gets smaller?

(c) Consider the same setting as the previous exercise and let F be the c.d.f. of P and $p \in (0, 1)$. The pth quantile and p-th sample quantile are, respectively,

$$\xi_p = \inf\{x \colon F(x) \geq p\}$$

and

$$\hat{\xi}_p = \inf\{x \colon F_n(x) \geq p\},$$

res[ectively, where F_n is the sample c.d.f. (i.e. $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x)$). Show that, for any $\epsilon > 0$,

$$\mathbb{P}\left(|\hat{\xi}_p - \xi_p| > \epsilon\right) \leq 2\exp\left\{-2n\delta_\epsilon^2\right\},$$

where $\delta_\epsilon = \min\left\{F(x_p + \epsilon) - p, p - F(\xi_p - \epsilon)\right\}$.

Write, for instance, $\mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right) = \mathbb{P}\left(p > F_n(\xi_p + \epsilon)\right)$. *Then, notice that* $F_n(x)$ *is a sum of i.i.d. Bernoulli and use Hoeffding yet again...*

Points: 20+5 pts = 10 + 5 + 10.

Solution.

(a)

Let $\xi_{\frac{1}{2}} = \inf\left\{x \colon F(x) \geq \frac{1}{2}\right\}$ be a median. Then

$$k_1 < \sum_{i=1}^n I\left(X_i \leq \xi_{\frac{1}{2}}\right) < k_2 \iff k_1 < \sum_{i=1}^n I\left(X_{(i)} \leq \xi_{\frac{1}{2}}\right) < k_2 \iff X_{(k_1)} < \xi_{\frac{1}{2}} < X_{(k_2)}.$$

Now since the cdf F is right continuous, $\mathbb{P}\left(X_i \leq \xi_{\frac{1}{2}}\right) = F\left(\inf\left\{x\colon F(x) \geq \frac{1}{2}\right\}\right) = \frac{1}{2}$, hence $I\left(X_i \leq \xi_{\frac{1}{2}}\right) \sim Bernouli\left(\frac{1}{2}\right)$. Hence by applying Hoeffding's inequality, we have a bound for tail probability as

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} I\left(X_i \leq \xi_{\frac{1}{2}}\right) - \frac{n}{2}\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n}\right).$$

Hence by plugging in $t = \sqrt{\frac{n}{2}\log\left(\frac{2}{\alpha}\right)}$, we have a bound as

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} I\left(X_i \leq \xi_{\frac{1}{2}}\right) - \frac{n}{2}\right| \geq \sqrt{\frac{n}{2}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence we lower bound the probability $\mathbb{P}\left(X_{(k_1)} < \xi_{\frac{1}{2}} < X_{(k_2)}\right)$ as

$$\mathbb{P}\left(X_{(k_1)} < \xi_{\frac{1}{2}} < X_{(k_2)}\right) \geq 1 - \alpha$$

with $k_1 = \left\lfloor \frac{n}{2} - \sqrt{\frac{n\log(2/\alpha)}{2}} \right\rfloor$ and $k_2 = \left\lceil \frac{n}{2} + \sqrt{\frac{n\log(2/\alpha)}{2}} \right\rceil$. i.e. $\left(X_{(k_1)}, X_{(k_2)}\right)$ is $1-\alpha$ confidence interval for $\xi_{\frac{1}{2}}$.

(b)

For any $\epsilon > 0$, note that from DKW inequality,

$$\mathbb{P}\left(\|F - F_n\|_\infty \leq \sqrt{\frac{\log\left(2/\epsilon\right)}{2n}}\right) \geq 1 - \epsilon.$$

Let $A := \left\{\omega : \|F - F_n\|_\infty \leq \sqrt{\frac{\log(2/\epsilon)}{2n}}\right\}$ be the event where $\|F - F_n\|_\infty$concentrates, and suppose we are conditioned on A. Then from $F_n(X_{(k_1)}) = \frac{k_1}{n}$,

$$F(X_{(k_1)}) \geq F_n(X_{(k_1)}) + \|F - F_n\|_\infty \geq \frac{k_1}{n} + \|F - F_n\|_\infty,$$

so that

$$X_{(k_1)} \geq F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right).$$

And similarly,

$$X_{(k_2)} \leq F^{-1}\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right),$$

and hence under the event A,

$$X_{(k_2)} - X_{(k_1)} \leq F^{-1}\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) - F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right).$$

Suppose further that

$$n \geq \frac{\left(\sqrt{\log(2/\alpha)} + \sqrt{\log(2/\epsilon)}\right)^2 + 8C\delta}{2C^2\delta^2}.$$

Note that this implies $\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}} \geq \frac{1}{2} - C\delta$. Then $F(m-\delta) = F(m) - \int_{m-\delta}^{m} F'(x)dx \leq F(m) - C\delta = \frac{1}{2} - C\delta$, hence

$$F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) \geq m - \delta.$$

Then from

$$\frac{1}{2} - \left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) = \int_{F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right)}^{m} F'(x)dx \geq C\left(m - F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right)\right),$$

we can lower bound as

$$F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) \geq m - \frac{1}{C}\left(\frac{1}{2} - \left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right)\right).$$

And similarly,

$$F^{-1}\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) \leq m + \frac{1}{C}\left(\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) - \frac{1}{2}\right).$$

Hence $F^{-1}\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) - F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right)$ can be further upper bounded as

$$F^{-1}\left(\frac{k_2}{n} + \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) - F^{-1}\left(\frac{k_1}{n} - \sqrt{\frac{\log(2/\epsilon)}{2n}}\right) \leq \frac{1}{C}\left(\frac{k_2 - k_1}{n} + \sqrt{\frac{2\log(2/\epsilon)}{n}}\right).$$

Hence we have the high probability bound as

$$\mathbb{P}\left(X_{(k_2)} - X_{(k_1)} \leq \frac{1}{C}\left(\frac{k_2 - k_1}{n} + \sqrt{\frac{2\log(2/\epsilon)}{n}}\right)\right) \geq 1 - \epsilon.$$

In particular, when $k_1 = \left\lfloor \frac{n}{2} - \sqrt{\frac{n\log(2/\alpha)}{2}} \right\rfloor$ and $k_2 = \left\lceil \frac{n}{2} + \sqrt{\frac{n\log(2/\alpha)}{2}} \right\rceil$, $\frac{k_2 - k_1}{n} \leq \sqrt{\frac{2\log(2/\alpha)}{n}} + \frac{2}{n}$, and hence when $n \geq \frac{\left(\sqrt{\log(2/\alpha)} + \sqrt{\log(2/\epsilon)}\right)^2 + 8C\delta}{2C^2\delta^2}$,

$$\mathbb{P}\left(X_{(k_2)} - X_{(k_1)} \leq \frac{1}{C}\left(\sqrt{\frac{2\log(2/\alpha)}{n}} + \sqrt{\frac{2\log(2/\epsilon)}{n}} + \frac{2}{n}\right)\right) \geq 1 - \epsilon.$$

Note that when η or C gets smaller, the lower bound on n grows, and hence we need more sample to achieve the high probability bound. Also, the high probability bound grows as C gets smaller, hence we are more uncertain with the $1 - \alpha$ confidence interval in (a).

(c)

Note that

$$\hat{\xi_p} > \xi_p + \epsilon \iff \inf\{x \colon F_n(x) \geq p\} > \xi_p + \epsilon \iff F_n(\xi_p + \epsilon) < p,$$

hence

$$\mathbb{P}\left(\hat{\xi_p} > \xi_p + \epsilon\right) = \mathbb{P}\left(p > F_n(\xi_p + \epsilon)\right).$$

Then since $1(X_i \leq x) \sim \mathit{Bernoulli}(F(x))$, so by Hoeffding's inequality, the tail probability $\mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right)$ can be bounded as

$$\begin{aligned}
\mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right) &= \mathbb{P}\left(F_n(\xi_p + \epsilon) < p\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq \xi_p + \epsilon) - F(\xi_p + \epsilon) < p - F(\xi_p + \epsilon)\right) \\
&\leq \exp\left(-2n(F(\xi_p + \epsilon) - p)^2\right).
\end{aligned}$$

Also, note that

$$\hat{\xi}_p \leq \xi_p - \epsilon \iff \inf\{x \colon F_n(x) \geq p\} \leq \xi_p - \epsilon \iff F_n(\xi_p - \epsilon) \geq p.$$

Hence similarly, the tail probability $\mathbb{P}\left(\hat{\xi}_p \leq \xi_p - \epsilon\right)$ can be bounded as

$$\begin{aligned}
\mathbb{P}\left(\hat{\xi}_p \leq \xi_p - \epsilon\right) &= \mathbb{P}\left(F_n(\xi_p - \epsilon) \geq p\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq \xi_p - \epsilon) - F(\xi_p - \epsilon) \geq p - F(\xi_p - \epsilon)\right) \\
&\leq \exp\left(-2n(p - F(\xi_p - \epsilon))^2\right).
\end{aligned}$$

Hence the tail probability $\mathbb{P}\left(\left|\hat{\xi}_p - \xi_p\right| > \epsilon\right)$ can be bounded as

$$\begin{aligned}
\mathbb{P}\left(\left|\hat{\xi}_p - \xi_p\right| > \epsilon\right) &\leq \mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right) + \mathbb{P}\left(\hat{\xi}_p \leq \xi_p - \epsilon\right) \\
&\leq \exp\left(-2n(F(\xi_p + \epsilon) - p)^2\right) + \exp\left(-2n(p - F(\xi_p - \epsilon))^2\right) \\
&\leq 2\exp\left(-2n\delta_\epsilon^2\right),
\end{aligned}$$

where $\delta_\epsilon = \min\left\{F(x_p + \epsilon) - p, p - F(\xi_p - \epsilon)\right\}$.

2. Consider the linear regression model
$$Y = X\theta^* + \epsilon$$
where $\theta \in \mathbb{R}^d$, X is fixed and $\epsilon \in \mathbb{R}^n$ consists of independent zero-mean variables with finite variance. The ridge estimator is defined as
$$\hat{\theta}_{\text{ridge}} = \hat{\theta}_{\text{ridge}}(\lambda) = \text{argmin}_{\theta \in \mathbb{R}^d}\left\{\frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\theta\|^2\right\},$$
where $\lambda > 0$.

   (a) Show that $\hat{\theta}_{\text{ridge}}$ is uniquely defined for any $\lambda > 0$ and find a closed-form expression. Will the solution exist and be unique if $d > n$?

   (b) Compute the bias of $\hat{\theta}_{\text{ridge}}$.

Points: 15 pts = 8 + 7.

Solution.

(a)

Let $\Omega := \left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}$. Note that for any $\eta \in \mathbb{R}^d$, $\eta^\top(\frac{1}{n}X^\top X + \lambda I)\eta = \frac{1}{n}\|X\eta\|^2 + \lambda\|\eta\|^2 \geq 0$ and equality holds if and only if $\eta = 0$, so $\frac{1}{n}X^\top X + \lambda I$ is positive definite and hence the inverse Ω exists. Note that the objective function for ridge estimator can be written as

$$\begin{aligned}
&\frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\theta\|^2 \\
&= \frac{1}{n}\theta^\top X^\top X\theta - \frac{2}{n}Y^\top X\theta + Y^\top Y + \lambda\theta^\top\theta \\
&= \left(\theta - \Omega\frac{1}{n}X^\top Y\right)^\top \left(\frac{1}{n}X^\top X + \lambda I\right)\left(\theta - \Omega\frac{1}{n}X^\top Y\right) + Y^\top\left(I - \frac{1}{n^2}X\Omega X^\top\right)Y.
\end{aligned}$$

Then since $\frac{1}{n}X^\top X + \lambda I$ is positive definite as mentioned earlier,

$$\frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\theta\|^2 \geq Y^\top\left(I - \frac{1}{n^2}X\Omega X^\top\right)Y,$$

and equality holds if and only if $\theta - \Omega\frac{1}{n}X^\top Y = 0$, i.e. $\theta = \Omega\frac{1}{n}X^\top Y = \left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}\frac{1}{n}X^\top Y$. Hence the objective function has the unique minimizer

$$\hat{\theta}_{\text{ridge}} = \left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}\frac{1}{n}X^\top Y,$$

which also exists and is unique when $d > n$.

(b)

The bias $\mathbb{E}\left[\hat{\theta}_{ridge}\right] - \theta^*$ can be computed as

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}_{ridge}\right] - \theta^* &= \left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}\frac{1}{n}X^\top\mathbb{E}[Y] - \theta^* \\
&= \left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}\left(\frac{1}{n}X^\top X\theta^* - \left(\frac{1}{n}X^\top X + \lambda I\right)\theta^*\right) \\
&= -\lambda\left(\frac{1}{n}X^\top X + \lambda I\right)^{-1}\theta^*.
\end{aligned}$$

3. Considert the distribution-free framework for regression: the pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ has a distribution P on $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$ in the support of X, let $\mu(x) = \mathbb{E}[Y|X = x]$ be the regression function. As we discussed in class, linear regression postulates that $\mu(x) = \beta^\top x$, for some $\beta \in \mathbb{R}^d$. This is a very strong assumption, which is unlikely to hold in most scenarios. What if one still fits a linear regression function?

   (a) Let $\Sigma = \mathbb{V}[X]$, assumed to be invertible. Define

$$\beta^* = \text{argmin}_{\beta\in\mathbb{R}^d}\mathbb{E}\left[(Y - X^\top\beta)^2\right].$$

   The vector β^* contains the coefficients of the best (in an L_2 sense) approximation of Y by linear functions of X (In fact, $X^\top\beta^*$ is the L_2 projection of Y into the linear space of linear functions on X). Show that

$$\beta^* = \Sigma^{-1}\alpha,$$

   where $\alpha = \mathbb{E}[YX] \in \mathbb{R}^d$.

(b) Now observed data in the form of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$. Assume for simplicity that $\mathbb{E}[X] = 0$. The plug-in estimator of β^* is the ordinary least squares estimator

$$\hat{\beta} = \hat{\Sigma}^{-1}\hat{\alpha},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ and $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^m Y_i X_i$. We assume that P belongs to a large non-parametric class of probability distributions satisfying the folowing assumptions:

i. each P in the class has a Lebesgue density (which implies that $\hat{\Sigma}$ is invertible almost surely if $n \geq d$; why?).

ii. Y and all the coordinates of X are bounded in absolute value by some constant K, almost surely (this could be relaxed to a sub-gaussian assumption, but let's keep things simple).

iii. the covariancer matrix of X, Σ, has a positve minimal eigenvalue bounded from below by $\lambda_{\min} > 0$.

Compute a bound for

$$\|\hat{\beta} - \beta^*\|.$$

The bound should depend on d, K and $\lambda_{\min}$, all of which are allowed to change with n. Based on your bound, comment on the dependence on d.

Hint: Recall that $\|Ax\| \leq \|A\|_{\text{op}}\|x\|$, $\|AB\|_{\text{op}} \leq \|A\|_{\text{op}}\|B\|_{\text{op}}$ and that the maximal eigenvaue of Σ^{-1} (which is also its operator norm) is the reciprocal of the minimal eogenvale of Σ. Also, you may find the following result useful (see equation 5.8.2 in the book Matrix Analysis, by Horn and Johnson, 2012): letting $E = \hat{\Sigma} - \Sigma$, if $\|\Sigma^{-1}E\|_{\text{op}} < 1$, we have that

$$\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_{\text{op}} \leq \|\Sigma^{-1}\|_{\text{op}} \frac{\|\Sigma^{-1}E\|_{\text{op}}}{1 - \|\Sigma^{-1}E\|_{\text{op}}}.$$

You may want to use the matrix Bernstein inequality to get sharer rates.

Note: One should be able to infer this result from the main Theorem in the highly recommended paper "Random design analysis of ridge regression", by Daniel Hsu, Sham M. Kakade and Tong Zhang, available here. However, presumably, if you follow the hint you should end up with a simpler proof. I am curious to see what rates you get...

Points: 25 pts = 20 + 5.

Solution.

(a)

Note that

$$\begin{aligned}
\mathbb{E}\left[(Y - X^\top\beta)^2\right] &= \beta^\top\mathbb{E}\left[XX^\top\right]\beta - 2\beta^\top\mathbb{E}\left[XY\right] + \mathbb{E}\left[Y^2\right] \\
&= \beta^\top\Sigma\beta - 2\beta^\top\alpha + \mathbb{E}\left[Y^2\right] \\
&= \beta^\top\Sigma\beta - 2\beta^\top\Sigma\Sigma^{-1}\alpha + \alpha^\top\Sigma^{-1}\alpha + \mathbb{E}\left[Y^2\right] - \alpha^\top\Sigma^{-1}\alpha \\
&= (\beta - \Sigma^{-1}\alpha)^\top\Sigma(\beta - \Sigma^{-1}\alpha) + \mathbb{E}\left[Y^2\right] - \alpha^\top\Sigma^{-1}\alpha \\
&\geq \mathbb{E}\left[Y^2\right] - \alpha^\top\Sigma^{-1}\alpha,
\end{aligned}$$

and equality holds if and only if $(\beta - \Sigma^{-1}\alpha)^\top\Sigma(\beta - \Sigma^{-1}\alpha) = 0$. Since Σ is positive definite, this happens if and only if $\beta - \Sigma^{-1}\alpha = 0$, and hence $\beta^* = \Sigma^{-1}\alpha$.

(b)

$$\begin{aligned}\left\|\hat{\beta}-\beta^*\right\| &= \left\|\hat{\Sigma}^{-1}\hat{\alpha}-\Sigma^{-1}\alpha\right\| \\ &= \left\|\hat{\Sigma}^{-1}\hat{\alpha}-\Sigma^{-1}\hat{\alpha}+\Sigma^{-1}\hat{\alpha}-\Sigma^{-1}\alpha\right\| \\ &\le \left\|(\hat{\Sigma}^{-1}-\Sigma^{-1})\hat{\alpha}\right\| + \left\|\Sigma^{-1}(\hat{\alpha}-\alpha)\right\|.\end{aligned}$$

For bounding RHS, consider bounding $E := \hat{\Sigma}-\Sigma$ and $\hat{\alpha}-\alpha$ first. For bounding E, note that for all $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$,

$$0 \le v^\top \Sigma v = \mathbb{E}\left[v^\top X_i X_i^\top v\right] \le K^2 d$$

holds, and hence $\|\Sigma\|_{op} \le K^2 d$ holds. And

$$0 \le v^\top X_i X_i^\top v = (v^\top X_i)^2 \le \|v\|_2^2 \|X_i\|_2^2 \le \|v\|_2^2\, d\, \|X_i\|_\infty^2 \le K^2 d$$

holds, and then

$$-K^2 d \le v^\top (X_i X_i^\top - \Sigma) v \le K^2 d,$$

and hence $\left\|X_i X_i^\top - \Sigma\right\|_{op} \le K^2 d$. Also, for all $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$,

$$\begin{aligned}v^\top \mathbb{E}\left[(X_i X_i^\top - \Sigma)^2\right] v = v^\top \left(\mathbb{E}\left[(X_i X_i^\top)^2\right] - \mathbb{E}\left[\Sigma^2\right]\right) v &\le v^\top \mathbb{E}\left[\|X_i\|_2^2 X_i X_i^\top\right] v \\ &\le K^2 d v^\top \mathbb{E}\left[X_i X_i^\top\right] v \le K^4 d^2,\end{aligned}$$

And hence $\left\|(X_i X_i^\top - \Sigma)^2\right\|_{op} \le K^4 d^2$. Hence from matrix Bernstein inequality, E can be bounded as

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{\Sigma}-\Sigma\right\|_{op} \ge t\right) &= \mathbb{P}\left(\left\|\sum_{i=1}^n \left(X_i X_i^\top - \Sigma\right)\right\|_{op} \ge nt\right) \\ &\le 2d \exp\left(-\frac{n^2 t^2}{2\left(nK^4 d^2 + \frac{ntK^2 d}{3}\right)}\right) \\ &= 2d \exp\left(-\frac{nt^2}{2K^2 d\left(K^2 d + \frac{t}{3}\right)}\right).\end{aligned}$$

Then from HW1 Problem 4(b),

$$\mathbb{P}\left(\left\|\hat{\Sigma}-\Sigma\right\|_{op} \le \sqrt{\frac{2K^4 d^2}{n}\log\left(\frac{4d}{\delta}\right)} + \frac{2K^2 d}{3n}\log\left(\frac{4d}{\delta}\right)\right) \ge 1 - \frac{\delta}{2}.$$

Then for $n \ge \left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{8K^4 d^2}{\lambda_{\min}^2}\log\left(\frac{4d}{\delta}\right) \ge \frac{8K^4 d^2}{\lambda_{\min}^2}\log\left(\frac{4d}{\delta}\right)$, $\frac{2K^2 d}{3n}\log\left(\frac{4d}{\delta}\right) \le \sqrt{\frac{\lambda_{\min}^2}{18n}\log\left(\frac{4d}{\delta}\right)}$ and hence

$$\mathbb{P}\left(\left\|\hat{\Sigma}-\Sigma\right\|_{op} \le \sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2K^4 d^2}{n}\log\left(\frac{4d}{\delta}\right)}\right) \ge 1 - \frac{\delta}{2}.$$

Now, consider $\hat{\alpha} - \alpha$. Note that for $1 \leq i \leq n$ and $1 \leq j \leq d$, $Y_i X_{ij} - \alpha_j \in SG(K^4)$ and $\{Y_i X_{ij} - \alpha_j\}_{i=1,\dots,n}$ are independent, so $\frac{1}{n}\sum_{i=1}^n Y_i X_{ij} - \alpha_j \in SG\left(\frac{K^4}{n}\right)$. Then $\left(\frac{1}{n}\sum_{i=1}^n Y_i X_{ij} - \alpha_j\right)^2$ is sub-exponential with $\nu^2 = \frac{256K^8}{n^2}$ and $\alpha = \frac{16K^4}{n}$, and hence $\|\hat{\alpha} - \alpha\|^2$ is sub-exponential with $\nu^2 = \frac{256K^8 d^2}{n^2}$ and $\alpha = \frac{16K^4 d}{n}$. Hence applying sub-exponential bound gives

$$\mathbb{P}\left(\|\hat{\alpha} - \alpha\|^2 \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{n^2 t^2}{512K^8 d^2}\right), & \text{if } 0 \leq t \leq \frac{16K^4 d}{n}, \\ 2\exp\left(-\frac{nt}{32K^4 d}\right), & \text{if } t > \frac{16K^4 d}{n}. \end{cases}$$

and hence applying $t = \frac{32K^4 d \log(4/\delta)}{n} \geq \frac{16K^4 d}{n}$ gives

$$\mathbb{P}\left(\|\hat{\alpha} - \alpha\| \leq 4\sqrt{2}K^2 \sqrt{\frac{d}{n}\log\left(\frac{4}{\delta}\right)}\right) \geq 1 - \frac{\delta}{2}.$$

Hence with probability at least $1-\delta$, $\left\|\hat{\Sigma} - \Sigma\right\|_{op} \leq \sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2K^4 d^2}{n}\log\left(\frac{4d}{\delta}\right)}$ and $\|\hat{\alpha} - \alpha\| \leq 4\sqrt{2}K^2\sqrt{\frac{d}{n}\log\left(\frac{4}{\delta}\right)}$. Suppose we are conditioned on this event, and assume

$$n \geq \left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{8K^4 d^2}{\lambda_{\min}^2}\log\left(\frac{4d}{\delta}\right).$$

Then from $\left\|\Sigma^{-1}\right\|_{op} \leq \frac{1}{\lambda_{\min}}$ and $\left\|\Sigma^{-1}E\right\|_{op} \leq \left\|\Sigma^{-1}\right\|_{op}\|E\|_{op} \leq \frac{1}{\lambda_{\min}}\sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2K^4 d^2}{n}\log\left(\frac{4d}{\delta}\right)} \leq \frac{1}{2}$, we get

$$\begin{aligned}\left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\|_{op} &\leq \left\|\Sigma^{-1}\right\|_{op}\frac{\left\|\Sigma^{-1}E\right\|_{op}}{1 - \left\|\Sigma^{-1}E\right\|_{op}} \leq 2\left\|\Sigma^{-1}\right\|_{op}^2\|E\|_{op} \\ &\leq \frac{2}{\lambda_{\min}^2}\sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2K^4 d^2}{n}\log\left(\frac{4d}{\delta}\right)}.\end{aligned}$$

Then $\|\hat{\alpha}\| = \left\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\right\| \leq K^2\sqrt{d}$, and hence

$$\begin{aligned}\left\|(\hat{\Sigma}^{-1} - \Sigma^{-1})\hat{\alpha}\right\| &\leq \left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\|_{op}\|\hat{\alpha}\| \\ &\leq \frac{2K^4}{\lambda_{\min}^2}\sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2d^3}{n}\log\left(\frac{4d}{\delta}\right)}.\end{aligned}$$

Also,

$$\begin{aligned}\left\|\Sigma^{-1}(\hat{\alpha} - \alpha)\right\| &\leq \left\|\Sigma^{-1}\right\|_{op}\|\hat{\alpha} - \alpha\| \\ &\leq \frac{4\sqrt{2}K^2}{\lambda_{\min}}\sqrt{\frac{d}{n}\log\left(\frac{4}{\delta}\right)}.\end{aligned}$$

Hence with probability $1 - \delta$,

$$\begin{aligned}\left\|\hat{\beta} - \beta^*\right\| &\leq \left\|(\hat{\Sigma}^{-1} - \Sigma^{-1})\hat{\alpha}\right\| + \left\|\Sigma^{-1}(\hat{\alpha} - \alpha)\right\| \\ &\leq \frac{2K^4}{\lambda_{\min}^2}\sqrt{\left(1 + \frac{\lambda_{\min}^2}{36K^4 d^2}\right)\frac{2d^3}{n}\log\left(\frac{4d}{\delta}\right)} + \frac{4\sqrt{2}K^2}{\lambda_{\min}}\sqrt{\frac{d}{n}\log\left(\frac{4}{\delta}\right)}.\end{aligned}$$

Hence when $\lambda_{\min}$ is small enough and K is large enough, with probability $1-\delta$,

$$\left\|\hat{\beta}-\beta\right\| = O\left(\frac{K^4}{\lambda_{\min}^2}\sqrt{\frac{d^3}{n}\log\left(\frac{d}{\delta}\right)}\right).$$

Therefore, when other parameters are fixed, $\left\|\hat{\beta}-\beta\right\|$ are bounded in the order of $\sqrt{\frac{d^3 \log d}{n}}$. This was expectable, since roughly speaking, $\hat{\Sigma}^{-1}$ consists of d^2 random entries and $\hat{\alpha}$ consists of d random entries, and we are using naive bound, hence $\hat{\beta} = \hat{\Sigma}^{-1}\hat{\alpha}$ should deal with randomness scaling with d^3 factors, although we have strong assumption that $\lambda_{\min}$ is fixed. When $\lambda_{\min}$ decreases as n and d grow, the rate of convergence will be worse.

4. **Hard thresholding in the sub-gaussian many means problem.** Suppose we observe the vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, where

$$X = \theta^* + \epsilon,$$

with $\theta^* \in \mathbb{R}^d$ unknown and $\epsilon \in SG_d(\sigma^2)$. We would like to estimate θ^* using the hard thresholding estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ with parameter $\tau > 0$, given by:

$$\hat{\theta}_i = \begin{cases} X_i & \text{if } |X_i| > \tau \\ 0 & \text{if } |X_i| \le \tau. \end{cases}$$

This estimator either keeps or kills each coordinate of X.

For $\delta \in (0,1)$, set

$$\tau = 2\sigma\sqrt{2\log(2d/\delta)}.$$

Notice that $\mathbb{P}(\max_i |\epsilon_i| > \tau/2) \le \delta$ (If this surprises you, refresh your memory on maximal inequalities).

(a) Prove that the hard-thresholding estimator is the solution the optimization problem

$$\min_{\theta\in\mathbb{R}^d} \|X-\theta\|^2 + \tau^2\|\theta\|_0.$$

(b) Prove that if $\|\theta^*\|_0 = k$, with probability at least $1-\delta$,

$$\|\hat{\theta}-\theta^*\|^2 \le C\sigma^2 k\log(2d/\delta),$$

for some universal constant $C > 0$. *Hint: show that, for each* $i = 1, \ldots, d$

$$|\hat{\theta}_i - \theta_i^*| \le C' \min\{|\theta_i^*|, \tau\}$$

for some $C' > 0$, *with probability at least* $1-\delta$.

(c) Compare with the oracle estimator $\hat{\theta}^{\mathrm{or}}$, with coordinates given by

$$\hat{\theta}_i^{\mathrm{or}} = \begin{cases} X_i & \text{if } i \in \mathrm{supp}(\theta^*) \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, d$. This estimator is of course not computable, as it requires knwoledge of $\mathrm{supp}(\theta^*)$. It is an estimator that an oracle, who has access to this additional knowledge, would be able to compute. Oracle estimators are idealized estimators, which perform at least as well as any computable estimators. Thus, in order to show that a given estimator performs well, it is enoygh to show that it mimicks closely the performance of an oracle estimator.

(d) Show that if $\min_{i\in\mathrm{supp}(\theta^*)} |\theta_i| > \frac{3}{2}\tau$, then, with probability at least $1-\delta$,

$$\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*).$$

How does $\hat{\theta}$ compare now to the oracle estimator?

Points: 30 pts = 6 + 8 + 8 + 8.

Solution.

(a)

Note that

$$\|X-\theta\|_2^2 + \tau^2\|\theta\|_0 = \sum_{i=1}^d \left((X_i-\theta_i)^2 + \tau^2 I(\theta_i \neq 0)\right).$$

Then

$$\begin{aligned}(X_i-\theta_i)^2 + \tau^2 I(\theta_i\neq 0) &\geq X_i^2 I(\theta_i = 0) + \tau^2 I(\theta_i \neq 0)\\ &\geq \min\left\{X_i^2, \tau^2\right\},\end{aligned}$$

and equality holds if and only if

$$\theta_i = \begin{cases} 0 & \text{if } |X_i| < \tau \\ 0 \text{ or } X_i & \text{if } |X_i| = \tau \\ X_i & \text{if } |X_i| > \tau. \end{cases}$$

Hence

$$\begin{aligned}\|X-\theta\|_2^2 + \tau^2\|\theta\|_0 &= \sum_{i=1}^d \left((X_i-\theta_i)^2 + \tau^2 I(\theta_i\neq 0)\right)\\ &\geq \sum_{i=1}^d \min\left\{X_i^2,\tau^2\right\},\end{aligned}$$

and the hard-thresholding estimator $\hat{\theta}_i = \begin{cases} X_i & \text{if } |X_i| > \tau \\ 0 & \text{if } |X_i| \leq \tau. \end{cases}$ satisfies the equality condition. Hence it is one solution to the optimization problem

$$\min_{\theta\in\mathbb{R}^d} \|X-\theta\|^2 + \tau^2\|\theta\|_0.$$

(b)

Note that $\hat{\theta}_i = X_i I(|X_i| > \tau)$, so when $\max_i |\epsilon_i| \leq \frac{\tau}{2}$ holds, then

$$\begin{aligned}&\left|\hat{\theta}_i - \theta_i^*\right|\\ &= \left|X_i - \theta_i^*\right| I(|X_i| > \tau) + |\theta_i^*| I(|X_i| \leq \tau)\\ &\leq |\epsilon_i| I(|\theta_i^*| + |\epsilon_i| > \tau) + |\theta_i^*| I(|\theta_i^*| - |\epsilon_i| \leq \tau)\\ &\leq \frac{\tau}{2} I\left(|\theta_i^*| > \frac{1}{2}\tau\right) + |\theta_i^*| I\left(|\theta_i^*| < \frac{3}{2}\tau\right)\\ &= |\theta_i^*| I\left(|\theta_i^*| \leq \frac{1}{2}\tau\right) + \left(|\theta_i^*| + \frac{\tau}{2}\right)\left(I\left(\frac{1}{2}\tau < |\theta_i^*| \leq \tau\right) + I\left(\tau < |\theta_i^*| < \frac{3}{2}\tau\right)\right) + \frac{\tau}{2} I\left(|\theta_i^*| \geq \frac{3}{2}\tau\right)\\ &\leq 2|\theta_i^*| I\left(|\theta_i^*| \leq \tau\right) + 2\tau I\left(|\theta_i^*| > \tau\right)\\ &\leq 2\min(|\theta_i^*|, \tau).\end{aligned}$$

And $\|\hat{\theta}-\theta^*\|_2^2$ can be correspondingly upper bounded as

$$\begin{aligned}\|\hat{\theta}-\theta^*\|_2^2 = \sum_{i=1}^d \left|\hat{\theta}_i - \theta_i^*\right|^2 &\leq 4\sum_{i=1}^d \min(|\theta_i^*|,\tau)^2 \\ &\leq 4\sum_{i=1}^d \tau^2 I(|\theta_i^*|>0) = 4\|\theta^*\|_0\tau^2.\end{aligned}$$

Hence $\max_i |\epsilon_i| \leq \frac{\tau}{2}$ implies $\|\hat{\theta}-\theta^*\|_2^2 \leq 4\|\theta^*\|_0\tau^2 = 32\sigma^2 k\log\left(\frac{2d}{\delta}\right)$, i.e.

$$\mathbb{P}\left(\|\hat{\theta}-\theta^*\|_2^2 \leq 32\sigma^2 k\log\left(\frac{2d}{\delta}\right)\right) \geq \mathbb{P}\left(\max_i |\epsilon_i| \leq \frac{\tau}{2}\right) \geq 1-\delta.$$

(c)
Note that $\|\hat{\theta}^{or}-\theta^*\|_2^2 = \sum \epsilon_i^2 I(\theta_i^*>0)$. Then $\epsilon_i^2 \in SE((16\sigma^2)^2, 16\sigma^2)$ from Homework 2 Problem 4, so for θ^* with $\|\theta^*\|_0 = k$,

$$\sum \epsilon_i^2 I(\theta_i^*>0) \in SE\left(256\sigma^4k^2, 16\sigma^2 k\right),$$

and from this, we apply sub-exponential tail bound to get

$$\mathbb{P}\left(\|\hat{\theta}^{or}-\theta^*\|_2^2 \geq t\right) \leq \begin{cases}\exp\left(-\frac{t^2}{512\sigma^4k^2}\right) & t \leq 16\sigma^2k, \\ \exp\left(-\frac{t}{32\sigma^2k}\right) & t \geq 16\sigma^2k.\end{cases}$$

Hence by applying $t = 32\sigma^2k\log\left(\frac{1}{\delta}\right) > 16\sigma^2k$,

$$\mathbb{P}\left(\|\hat{\theta}^{or}-\theta^*\|_2^2 \leq 32\sigma^2k\log\left(\frac{1}{\delta}\right)\right) \geq 1-\delta.$$

(d)
Suppose $\min_{i\in\text{supp}(\theta^*)} |\theta_i^*| > \frac{3}{2}\tau$ holds. Then if $\max_i |\epsilon_i| \leq \frac{\tau}{2}$ further holds, then for $i \in supp(\theta^*)$, $|X_i| \geq |\theta_i^*| - |\epsilon_i| > \tau$, so $\left|\hat{\theta}_i\right| = |X_i| \neq 0$, i.e. $i \in supp(\hat{\theta})$. Conversely, we have seen in (b) that $i \notin supp(\theta^*)$ implies $\left|\hat{\theta}_i - \theta_i^*\right| \leq 2\min(|\theta_i^*|,\tau) = 0$, so $\left|\hat{\theta}_i\right| = 0$ holds. Hence $\max_i |\epsilon_i| \leq \frac{\tau}{2}$ implies $\text{supp}(\hat{\theta}) = \text{supp}(\theta^*)$, and hence

$$\mathbb{P}\left(\text{supp}(\hat{\theta}) = \text{supp}(\theta^*)\right) \geq \mathbb{P}\left(\max_i |\epsilon_i| \leq \frac{\tau}{2}\right) \geq 1-\delta.$$

Now, note that when $supp(\hat{\theta}) = supp(\theta^*)$ holds,

$$\hat{\theta}_i = X_i I(|X_i|>\tau) = X_i I(\hat{\theta}_i \neq 0) = X_i I(\theta_i^* \neq 0) = \hat{\theta}_i^{or},$$

hence$\hat{\theta} = \hat{\theta}^{or}$ under $supp(\hat{\theta}) = supp(\theta^*)$. Therefore, $\|\hat{\theta}-\theta^*\|_2^2 > 32\sigma^2k\log\left(\frac{2d}{\delta}\right)$ implies either $\|\hat{\theta}^{or}-\theta^*\|_2^2 > 32\sigma^2k\log\left(\frac{2d}{\delta}\right)$ or $\text{supp}(\hat{\theta}) \neq \text{supp}(\theta^*)$, so

$$\begin{aligned}&\mathbb{P}\left(\|\hat{\theta}-\theta^*\|_2^2 \leq 32\sigma^2k\log\left(\frac{1}{\delta}\right)\right) \\ &= 1-\mathbb{P}\left(\|\hat{\theta}-\theta^*\|_2^2 > 32\sigma^2k\log\left(\frac{1}{\delta}\right)\right) \\ &\geq 1-\mathbb{P}\left(\|\hat{\theta}^{or}-\theta^*\|_2^2 > 32\sigma^2k\log\left(\frac{1}{\delta}\right)\right) - \mathbb{P}\left(\text{supp}(\hat{\theta}) \neq \text{supp}(\theta^*)\right) \\ &\geq 1-2\delta.\end{aligned}$$

5. **Reading Exercise, graded for effort, not correctness.**

The following paper outlines a general strategy, called primal dual witness construction, for showing model selection consistency for the lasso, aka sparsistency. It means that the LASSO selects the right set of non-zero covariates. It an be extended to other penalized likelihood procedures.

- Wainwright, M. (2009). Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso), IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 55, NO. 5, 2183–2202.

Reproduce the proof of Theorem 1. Notice that the incoherence condition, which is necessary for the result, is a very strong assumption, unlikely to be satisifed in practice. Since this assumption is nearly necessary, conclude that the LASSO in practice should not be expected to be sparistent (even if you believe a linear model!).

Points: 10 pts.

Solution.

Let $S = S(\beta^*)$ be the support of the true vector β^*. We first write the statement of Theorem 1: Assume that $n \times p$ design matrix X satisfies that there exists some incoherence parameter $\gamma \in (0, 1]$ such that

$$|||X_{S^c}^\top X_S(X_S^\top X_S)^{-1}|||_\infty \leq (1-\gamma), \tag{1}$$

where $|||\cdot|||$ denotes the l_∞/l_∞ operator norm[1], and there exists some $C_{\min} > 0$ such that

$$\Lambda_{\min}(\frac{1}{n}X_S^\top X_S) \geq C_{\min}, \tag{2}$$

where $\Lambda_{\min}$ denotes the minimal eigenvalue, and its n-dimensional columns are normalized such that

$$n^{-1/2}\max_{j\in S^c}\|X_j\|_2 \leq 1. \tag{3}$$

Suppose that the sequence of regularization parameters $\{\lambda_n\}$ satisfies

$$\lambda_n > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}. \tag{4}$$

Then for some constant $c_1 > 0$, the following properties hold with probability greater than $1 - 4\exp(-c_1 n\lambda_n^2) \to 1$:

(a) The Lasso has a unique solution $\hat{\beta} \in \mathbb{R}^p$ with its support contained within the true support (i.e. $S(\hat{\beta}) \subset S(\beta^*)$), and satisfies the ℓ_∞ bound:

$$\|\hat{\beta}_S - \beta_S^*\|_\infty \leq \underbrace{\lambda_n\left[\|(X_S^\top X_S/n)^{-1}\|_\infty + \frac{4\sigma}{\sqrt{C_{\min}}}\right]}_{g(\lambda_n)}.$$

(b) If in addition the minimum value of the regression vector $\beta_S$on its support is bounded below as $\beta_{\min} > g(\lambda_n)$, then $\hat{\beta}$ has the correct signed support (i.e., $\mathbb{S}_\pm(\hat{\beta}) = \mathbb{S}_\pm(\beta^*)$).

[1] For an $m \times n$ matrix M, $|||M|||_\infty = \max_{i=1,\cdots,m} \sum_{j=1}^n |M_{ij}|$.

Let $\breve{z}_S$ be an element of the subdifferential of the ℓ_1 norm evaluated at

$$\breve{\beta}_S = \arg\min_{\beta_S \in \mathbb{R}^k} \left\{ \frac{1}{2n} \|y - X_S \beta_S\|_2^2 + \lambda_n \|\beta_S\|_1 \right\}.$$

For each $j \in S^c$, let $Z_j := X_j^\top \left\{ X_S (X_S^\top X_S)^{-1} \breve{z}_S + \Pi_{X_S^\perp}(\frac{\omega}{\lambda_n n}) \right\}$ where $\Pi_{X_S^\perp} := I_{n \times n} - X_S (X_S^\top X_S)^{-1} X_S$ is an orthogonal projection matrix. And for each $i \in S$, let $\Delta_i := e_i^\top \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \left[\frac{1}{n} X_S^\top \omega - \lambda_n \breve{z}_S \right]$[2]. Then as formalized in Lemma 3 in Wainwright [2009], Z_j is the candidate variable solved for Step 3 of the primal-dual construction, and strict dual feasibility holds if and only if $|Z_j| < 1$ for all $j \in S^c$. Also, in proof of Lemma 3 in Wainwright [2009], $\Delta_i = (\breve{\beta}_S - \beta_S^*)_i$ holds. Hence if strict dual feasibility holds, then Lemma 2(a) in Wainwright [2009] implies that Lasso solution $\hat{\beta}$ uniquely exists with $\hat{\beta} = \breve{\beta}_S$ and $Supp(\hat{\beta}) \subset Supp(\beta^*)$, and hence $\|\hat{\beta} - \beta^*\|_\infty = \max_{i \in S} |\Delta_i|$. Hence we show strict dual feasibility and bound $\max_{i \in S} |\Delta_i|$ for showing (a). (b) follows correspondingly since if $|\hat{\beta}_i - \beta_i^*| \le g(\lambda_n) < \beta_{\min}$, then $sign(\hat{\beta}_i) = sign(\beta_i^*)$ follows correspondingly.
(Establishing strict dual feasibility)
We show that $\max_{j \in S^c} |Z_j| < 1 - \frac{\gamma}{2}$ with high probability. Note that we have the decomposition $Z_j = \mu_j + \tilde{Z}_j$, where $\mu_j = X_j^\top X_S (X_S^\top X_S)^{-1} \breve{z}_S$ and $\tilde{Z}_j := X_j^\top \Pi_{X_S^\perp}(\frac{\omega}{\lambda_n n})$ a zero-mean sub-Gaussian noise variable. Since $\breve{z}_S$ is a subgradient vector for ℓ_1 norm, $\|\breve{z}_S\|_\infty \le 1$. Applying (1) yields that for all indices $j \in S^c$

$$|\mu_j| = |X_j^\top X_S (X_S^\top X_S)^{-1} \breve{z}_S| \le |||X_j^\top X_S (X_S^\top X_S)^{-1}||| \|\breve{z}_S\|_\infty \le 1 - \gamma.$$

Hence we obtain

$$\max_{j \in S^c} |Z_j| \le (1 - \gamma) + \max_{j \in S^c} |\tilde{Z}_j|. \tag{5}$$

Since each ω_i are independent with $\omega_i \in SG(\sigma^2)$, from Homework 1 Problem 7 Details, $\tilde{Z}_j = \left(\frac{1}{\lambda_n n} \Pi_{X_S^\perp}(X_j) \right) \omega \in SG\left(\frac{\sigma^2}{\lambda_n^2 n^2} \|\Pi_{X_S^\perp}(X_j)\|_2^2 \right)$ holds. Then from $\|\Pi_{X_S^\perp}(X_j)\|_2 \le \|X_j\|_2$ and (3), we have

$$\begin{aligned} \frac{\sigma^2}{\lambda_n^2 n^2} \|\Pi_{X_S^\perp}(X_j)\|_2^2 &\le \frac{\sigma^2}{\lambda_n^2 n^2} \|X_j\|_2^2 \\ &\le \frac{\sigma^2}{\lambda_n^2 n}, \qquad \text{(using (3))} \end{aligned}$$

so we have $\tilde{Z}_j \in SG\left(\frac{\sigma^2}{\lambda_n^2 n} \right)$. Hence from Homework 2 Problem 1,

$$\mathbb{P}\left(\max_{j \in S^c} |\tilde{Z}_j| \ge \frac{\gamma}{2} \right) \le 2 \exp\left(-\frac{\lambda_n^2 n \gamma^2}{8\sigma^2} + \log(p - k) \right).$$

[2] Note that this is different from its definition in Wainwright [2009].

From this, (5), and (4),

$$\begin{aligned}\mathbb{P}\left(\max_{j\in S^c}|Z_j|\geq 1-\frac{\gamma}{2}\right) &\leq \mathbb{P}\left(\max_{j\in S^c}|\tilde{Z}_j|\geq \frac{\gamma}{2}\right) \qquad \text{(using (5))}\\ &\leq 2\exp\left(-\frac{\lambda_n^2 n\gamma^2}{8\sigma^2}+\log(p-k)\right)\\ &< 2\exp\left(-\frac{\lambda_n^2 n\gamma^2}{8\sigma^2}+\frac{\lambda_n^2 n\gamma^2\log(p-k)}{8\sigma^2\log p}\right) \qquad \text{(using (4))}\\ &= 2\exp\left(-\frac{\lambda_n^2 n\gamma^2}{8\sigma^2}\left(1-\frac{\log(p-k)}{\log p}\right)\right)\\ &\leq 2\exp\left(-c_1 n\lambda_n^2\right)\to 0,\end{aligned}$$

where c_1 is some constant satisfying $c_1 \leq \frac{\gamma^2}{8\sigma^2}\left(1-\frac{\log(p-k)}{\log p}\right)$.

(Establishing ℓ_∞ bounds)

By using triangle inequality, $\max_{i\in S}|\Delta_i|$ is upper bounded as

$$\max_{i\in S}|\Delta_i| \leq ||\left(\frac{1}{n}X_S^\top X_S\right)^{-1} X_S^\top\frac{\omega}{n}||_\infty + |||\left(\frac{1}{n}X_S^\top X_S\right)^{-1}|||_\infty\lambda_n. \tag{6}$$

We focus on the first term. For each $i=1,\cdots,k$, consider the random variable

$$V_k := e_i^\top\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top\omega.$$

Since each ω_i are independent with $\omega_i\in SG(\sigma^2)$, from Homework 1 Problem 7 Details,

$$V_k = \left(e_i^\top\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top\right)\omega \in SG\left(\left\|e_i^\top\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top\right\|_2^2\right).$$

Now, let $\frac{1}{\sqrt{n}}X_S = UDV^\top$ be the singular decomposition of $\frac{1}{\sqrt{n}}X_S$. Then $\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top = \frac{1}{\sqrt{n}}(VD^\top UU^\top DV^\top)^{-1}VD^\top U = \frac{1}{\sqrt{n}}V(D^\top D)^{-1}V^\top VD^\top U = \frac{1}{\sqrt{n}}V(D^\top D)^{-1}D^\top U$, so

$$|||\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top|||_2 = \frac{1}{\sqrt{n}}\Lambda_{\max}(V(D^\top D)^{-1}D^\top U) = \frac{1}{\sqrt{n\Lambda_{\min}(D^\top D)}} = \frac{1}{\sqrt{n\Lambda_{\min}(\frac{1}{n}X_S^\top X_S)}}.$$

Then from (2),

$$\begin{aligned}\|e_i^\top\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top\|_2^2 &\leq \|e_i\|_2^2|||\left(\frac{1}{n}X_S^\top X_S\right)^{-1}\frac{1}{n}X_S^\top|||_2^2\\ &= \frac{1}{n\Lambda_{\min}(\frac{1}{n}X_S^\top X_S)}\\ &\leq \frac{1}{nC_{\min}}, \qquad \text{(using (2))}\end{aligned}$$

And hence $V_k \in SG\left(\frac{\sigma^2}{nC_{\min}}\right)$. Hence from Homework 2 Problem 1 and using (4),

$$\begin{aligned}\mathbb{P}\left(\max_{i=1,\cdots,k} |V_k| > \frac{4\sigma\lambda_n}{\sqrt{C_{\min}}}\right) &\leq 2\exp\left(-8\lambda_n^2 n + \log k\right) \\ &\leq 2\exp\left(-c_2\lambda_n^2 n\right). \end{aligned} \tag{7}$$

[3]

Hence we get (a) and (b) from above two claims.

References

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ell_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2016018.

[3] This doesn't make sense for me. I feel like λ_n should satisfy additional condition such as $\lambda_n > \max\left\{\frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}, \sqrt{\frac{\log p}{8n}}\right\}$.