# 36-755, Fall 2017
# Homework 4

Due Wed Nov 1 by 5:00pm in Jisu's mailbox

**Points:** 100 + 3 pts total for the assignment.

1. In earlier works on the lasso, people have used even stronger assumptions than the restricted eigenvalue property. Here is one. Suppose that the design matrix $X$ is such that, for some integer $k > 0$,

$$\max_{i,j} \left| \frac{X_i^\top X_j}{n} - 1(i = j) \right| \leq \frac{1}{23k} \tag{1}$$

where $X_i$ is the $i$th column of $X$, $i = 1, \ldots, d$. Think about what that means.

(a) Show that this condition implies that, for any subset $S$ of $\{1, \ldots, d\}$ of cardinality no larger than $k < d$ and any $\Delta \in \mathbb{R}^d$ with $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$,

$$\|\Delta\|^2 \leq \frac{2}{n}\|X\Delta\|^2.$$

That is, show that this condition implies the $RE(3, 1/2)$ condition given in class for all non-empty subsets $S$ of $\{1, \ldots, d\}$ of size no larger than $k$. *Instead of the constant* 23 *you may take a larger one if it simplifies your calculations.*

(b) Suppose that the entries of $X$ are now populated by independent Rademacher variables (a Ra detacher variable is one that that takes the values $+1$ and $-1$ with equal probability). Show that, for any $\delta \in (0, 1)$, if

$$n \geq Ck^2(\log(d) + \log(1/\delta)),$$

for some constant $C > 0$, then $X$ satisfies the condition (1), with probability at least $1 - \delta$. *Again, instead of* 23 *feel free to show the result for a different constant if it helps with the calculations.*

**Points:** 35 + 3 pts = 20 + 3 + 15.

**Solution.**

(a)

Note that for a matrix $A \in \mathbb{R}^{d \times d}$ and for all $x, y \in \mathbb{R}^d$,

$$\left| x^\top Ay \right| = \left| \sum_{i,j} x_i A_{ij} y_j \right| \leq \max_{i,j} |A_{ij}| \sum_{i,j} |x_i| \, |y_j| = \max_{i,j} |A_{ij}| \, \|x\|_1 \|y\|_1. \tag{2}$$

Then, $\frac{1}{n}\|X\Delta\|_2^2$ can be expanded as

$$\begin{aligned} \frac{1}{n}\|X\Delta\|_2^2 &= \Delta^\top \left( \frac{X^\top X}{n} \right) \Delta \\ &= \Delta^\top I \Delta - (\Delta_S + \Delta_{S^c})^\top \left( I - \frac{X^\top X}{n} \right) (\Delta_S + \Delta_{S^c}). \end{aligned}$$

Then applying (2) and the condition $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$ gives further lower bound as

$$
\begin{aligned}
\frac{1}{n}\|X\Delta\|_2^2 &\geq \|\Delta\|_2^2 - \max_{i,j}\left|\left(I - \frac{X^\top X}{n}\right)_{ij}\right| (\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1)^2 \qquad \text{(using (2))} \\
&\geq \|\Delta\|_2^2 - 16 \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_1^2 \qquad \text{(using } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\text{)} \\
&\geq \|\Delta\|_2^2 - 16\,|S| \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_2^2 \\
&\geq \|\Delta\|_2^2 - 16k \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_2^2. 
\end{aligned} \tag{3}
$$

Also, by using $\Delta_S^\top \Delta_{S^c} = 0$, $\frac{1}{n}\|X\Delta\|_2^2$ can be alternatively expanded as

$$
\begin{aligned}
\frac{1}{n}\|X\Delta\|_2^2 &= \frac{1}{n}\|X\Delta_S + X\Delta_{S^c}\|_2^2 \\
&= \Delta_S^\top \left(\frac{X^\top X}{n}\right)(\Delta_S + 2\Delta_{S^c}) + \frac{1}{n}\|X\Delta_{S^c}\|_2^2 \\
&= \Delta_S^\top I(\Delta_S + 2\Delta_{S^c}) - \Delta_S^\top \left(I - \frac{X^\top X}{n}\right)(\Delta_S + 2\Delta_{S^c}) + \frac{1}{n}\|X\Delta_{S^c}\|_2^2 \\
&= \|\Delta_S\|_2^2 - \Delta_S^\top \left(I - \frac{X^\top X}{n}\right)(\Delta_S + 2\Delta_{S^c}) + \frac{1}{n}\|X\Delta_{S^c}\|_2^2 \\
&\geq \|\Delta_S\|_2^2 - \Delta_S^\top \left(I - \frac{X^\top X}{n}\right)(\Delta_S + 2\Delta_{S^c}).
\end{aligned}
$$

Then applying (2) and the condition $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$ gives further lower bound as

$$
\begin{aligned}
\frac{1}{n}\|X\Delta\|_2^2 &\geq \|\Delta_S\|_2^2 - \max_{i,j}\left|\left(I - \frac{X^\top X}{n}\right)_{ij}\right| \|\Delta_S\|_1(\|\Delta_S\|_1 + 2\|\Delta_{S^c}\|_1) \qquad \text{(using (2))} \\
&\geq \|\Delta_S\|_2^2 - 7 \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_1^2 \qquad \text{(using } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\text{)} \\
&\geq \|\Delta_S\|_2^2 - 7\,|S| \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_2^2 \\
&\geq \|\Delta_S\|_2^2 - 7k \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_2^2.
\end{aligned} \tag{4}
$$

Hence combining (3) and (4) gives

$$
\begin{aligned}
\frac{2}{n}\|X\Delta\|_2^2 &\geq \frac{1}{n}\|X\Delta\|_2^2 + \frac{1}{n}\|X\Delta\|_2^2 \\
&\geq \|\Delta\|_2^2 + \|\Delta_S\|_2^2 - 23k \max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \|\Delta_S\|_2^2.
\end{aligned}
$$

Then applying $\max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{23k}$ on this gives

$$
\frac{2}{n}\|X\Delta\|_2^2 \geq \|\Delta\|_2^2.
$$

**Details.**

Note that under slightly weaker condition $\max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{14k}$, we can still have a weaker result $\|\Delta_S\|_2^2 \leq \frac{2}{n}\|X\Delta\|_2^2$, which directly follows solely from (4).

Also, note that under slightly stronger condition $\max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{32k}$, solely arguing (3) without (4) is still sufficient for $\|\Delta\|_2^2 \leq \frac{2}{n}\|X\Delta\|_2^2$. There are 3 bonus points for those who showed the claim $\|\Delta\|_2^2 \leq \frac{2}{n}\|X\Delta\|_2^2$ under the condition $\max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{ck}$ with $c < 32$.

Also, note that for a matrix $A \in \mathbb{R}^{d\times d}$ and for all $x, y \in \mathbb{R}^d$, $\left|x^\top Ay\right| \leq \max_{i,j}|A_{ij}|\,\|x\|_2\|y\|_2$ does not generally hold. For example, when $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $x = y = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, then

$$\left|x^\top Ay\right| = 4 > 2 = \max_{i,j}|A_{ij}|\,\|x\|_2\|y\|_2.$$

(b)

For $1 \leq i \leq d$ and $1 \leq k \leq n$, let $X_{ik}$ be $i^{th}$ column and $k^{th}$ row of $X$. Then if $1 \leq i \leq d$, then $X_{ik}^2 = 1$ holds, and hence

$$\frac{X_i^\top X_i}{n} - 1(i=i) = \frac{1}{n}\sum_{i=1}^{n} X_{ik}^2 - 1 = 0.$$

Now, let $1 \leq i < j \leq d$. Then $\{X_{ik}X_{jk}\}_{k=1}^n$ is i.i.d. Rademacher, hence Rademacher being $SG(1)$ and HW1 Problem 6 Details imply that

$$\frac{X_i^\top X_j}{n} - 1(i=j) = \frac{1}{n}\sum_{i=1}^{n} X_{ik}X_{jk} \in SG(\frac{1}{n}).$$

Hence HW2 Problem 1 imply that

$$\begin{aligned} P\left(\max_{i,j}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{23k}\right) &= P\left(\max_{1\leq i<j\leq d}\left|\frac{X_i^\top X_j}{n} - 1(i=j)\right| \leq \frac{1}{23k}\right) \\ &\geq 1 - d(d-1)\exp\left(-\frac{n}{1058k^2}\right). \end{aligned}$$

Hence when $n \geq 2116k^2(\log d + \log(1/\delta))$, then

$$1 - d(d-1)\exp\left(-\frac{n}{1058k^2}\right) \geq 1 - d(d-1)\frac{\delta^2}{d^2} \geq 1 - \delta,$$

and hence $X$ satisfies the condition (1), with probability at least $1-\delta$.

2. Read the paper ”p-Values for High-Dimensional Regression” by Nicolai Meinshausen, Lukas Meier and Peter Bühlmann, JASA 2009, volume 104, issue 448, pages 1671-1681. Reproduce the proof of Theorem 3.2.

**Points:** 10 pts.

**Solution.**

For all $1 \leq b \leq B$ and for all $1 \leq j \leq p$, Let $\tilde{K}_j^{(b)}$ and $K_j^{(b)}$ be as follows:

$$\tilde{K}_j^{(b)} = \tilde{P}_j^{(b)} 1\{S \subseteq \tilde{S}^{(b)}\} + 1\{S \nsubseteq \tilde{S}^{(b)}\}, \tag{5}$$

$$K_j^{(b)} = P_j^{(b)} 1\{S \subseteq \tilde{S}^{(b)}\} + 1\{S \nsubseteq \tilde{S}^{(b)}\}. \tag{6}$$

Hence $K_j^{(b)}$ is the adjusted p-values if the estimated active set contains the true active set, and otherwise, all p-values are set to 1. Then for $f \in N$ and for any $\alpha, \gamma \in (0,1)$, from (6),

$$\mathbb{E}\left[1\left\{K_j^{(b)} \leq \alpha\gamma\right\}\right] \leq \mathbb{E}\left[\mathbb{P}\left[P_j^{(b)} \leq \alpha\gamma |\, S \subset \tilde{S}^{(b)}\right]\right] = \mathbb{E}\left[\min\left\{\frac{\alpha\gamma}{\left|\tilde{S}^{(b)}\right|}, 1\right\}\right], \tag{7}$$

since $\tilde{P}_j^{(b)} | S \subset \tilde{S}^{(b)} \sim Unif(0,1)$ for $j \in N$ and $P_j^{(b)} = \min\left\{\tilde{P}_j^{(b)} |\tilde{S}^{(b)}|, 1\right\}$ (there are some typos in the last equation of p.20). Also, since $\tilde{P}_j^{(b)} = 1$ when $j \notin \tilde{S}^{(b)}$, $P_j^{(b)} = K_j^{(b)} = 1$ as well, and hence

$$\mathbb{E}\left[\max_{j \in N} \frac{1\left\{K_j^{(b)} \leq \alpha\gamma\right\}}{\gamma}\right] \leq \mathbb{E}\left[\sum_{j \in N} \frac{1\left\{K_j^{(b)} \leq \alpha\gamma\right\}}{\gamma}\right] = \mathbb{E}\left[\sum_{j \in N \cap \tilde{S}^{(b)}} \frac{1\left\{K_j^{(b)} \leq \alpha\gamma\right\}}{\gamma}\right].$$

Then applying (7) to this gives

$$\mathbb{E}\left[\max_{j \in N} \frac{1\left\{K_j^{(b)} \leq \alpha\gamma\right\}}{\gamma}\right] \leq \mathbb{E}\left[\sum_{j \in N \cap \tilde{S}^{(b)}} \frac{\alpha}{\left|\tilde{S}^{(b)}\right|}\right] \leq \alpha.$$

For any $\alpha > 0$ and a random variable $U$ taking values in $[0,1]$,

$$\sup_{\gamma \in (\gamma_{min}, 1)} \frac{1\{U \leq \alpha\gamma\}}{\gamma} = \begin{cases} 0, & U \geq \alpha, \\ \alpha/U, & \alpha\gamma_{\min} \leq U < \alpha, \\ 1/\gamma_{\min}, & U < \alpha\gamma_{\min}. \end{cases}$$

Moreover, if $U$ has a uniform distribution on $[0,1]$,

$$\mathbb{E}\left[\sup_{\gamma \in (\gamma_{min}, 1)} \frac{1\{U \leq \alpha\gamma\}}{\gamma}\right] = \int_0^{\alpha\gamma_{\min}} \gamma_{\min}^{-1} dx + \int_{\alpha\gamma_{\min}}^{\alpha} x^{-1} dx = \alpha(1 - \log\gamma_{\min}).$$

Then from (5), $\tilde{K}_j^{(b)} | S \subset \tilde{S}^{(b)} \sim Unif(0,1)$ for $j \in N$, and hence

$$\mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\left\{\tilde{K}_j^{(b)} \leq \alpha\gamma\right\}}{\gamma}\right] \leq \mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\left\{\tilde{K}_j^{(b)} \leq \alpha\gamma\right\}}{\gamma} |\, S \subseteq \tilde{S}^{(b)}\right] = \alpha(1 - \log\gamma_{\min}). \tag{8}$$

Then, note that $K_j^{(b)} = \min\left\{\tilde{K}_j^{(b)} |\tilde{S}^{(b)}|, 1\right\}$ and $K_j = 1$ when $j \notin \tilde{S}^{(b)}$ as it has been for $P_j^{(b)}$ as well, so

$$\begin{aligned}
\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\right] &\leq \sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\mid S\subseteq\tilde{S}^{(b)}\right]\\
&= \sum_{j\in N\cap\tilde{S}^{(b)}}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\mid S\subseteq\tilde{S}^{(b)}\right]\\
&= \sum_{j\in N\cap\tilde{S}^{(b)}}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\left\{\tilde{K}_j^{(b)}|\tilde{S}^{(b)}|\leq\alpha\gamma\right\}}{\gamma}\mid S\subseteq\tilde{S}^{(b)}\right]\\
&\leq \sum_{j\in N\cap\tilde{S}^{(b)}}\frac{\alpha}{\left|\tilde{S}^{(b)}\right|}(1-\log\gamma_{\min})\quad\text{(from (8))}\\
&\leq \alpha(1-\log\gamma_{\min}).
\end{aligned}$$

Averaging over all bootstrap samples yields

$$\begin{aligned}
\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{\frac{1}{B}\sum_{b=1}^{B}1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\right] &\leq \sum_{j\in N}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\sup_{\gamma\in(\gamma_{\min},1)}\frac{1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\right]\\
&\leq \alpha(1-\log\gamma_{\min}).
\end{aligned}$$

Now, define for $u\in(0,1)$ the quantity $\pi_j(u)$ as the fraction of bootstrap samples that yield $K_j^{(b)}$ less than or equal to $u$, i.e.

$$\pi_j(u)=\frac{1}{B}\sum_{b=1}^{B}1\{K_j^{(b)}\leq u\}. \tag{9}$$

Then using a Markov inequality yields

$$\begin{aligned}
\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}1\{\pi_j(\alpha\gamma)\geq\gamma\}\right] &= \sum_{j\in N}\mathbb{P}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{\pi_j(\alpha\gamma)}{\gamma}\geq 1\right]\\
&\leq \sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{\pi_j(\alpha\gamma)}{\gamma}\right],
\end{aligned}$$

and hence from definition of $\pi_j(\cdot)$ in (9),

$$\begin{aligned}
\sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}1\{\pi_j(\alpha\gamma)\geq\gamma\}\right] &\leq \sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{\pi_j(\alpha\gamma)}{\gamma}\right]\\
&= \sum_{j\in N}\mathbb{E}\left[\sup_{\gamma\in(\gamma_{\min},1)}\frac{\frac{1}{B}\sum_{b=1}^{B}1\left\{K_j^{(b)}\leq\alpha\gamma\right\}}{\gamma}\right]\\
&\leq \alpha(1-\log\gamma_{\min}).
\end{aligned}$$

Since the events $\{Q_j(\gamma)\leq\alpha\}$ and $\{\pi_j(\alpha\gamma)\geq\gamma\}$ are equivalent, it follows that

$$\sum_{j\in N}\mathbb{P}\left[\inf_{\gamma\in(\gamma_{\min},1)}Q_j(\gamma)\leq\alpha\right]\leq\alpha(1-\log\gamma_{\min}),$$

implying that

$$\sum_{j\in N} \mathbb{P}\left[\inf_{\gamma\in(\gamma_{\min},1)} Q_j(\gamma)(1-\log\gamma_{\min}) \le \alpha\right] \le \alpha.$$

Then using the definition of $P_j = \min\left\{1,\, (1-\log\gamma_{\min})\inf_{\gamma\in(\gamma_{\min},1)} Q_j(\gamma)\right\}$,

$$\sum_{j\in N} \mathbb{P}\left[P_j \le \alpha\right] \le \alpha,$$

and thus by using the union bound,

$$\mathbb{P}\left[\min_{j\in N} P_j \le \alpha\right] \le \alpha,$$

which completes the proof.

3. **Performance of the best selection procedure.** Consider the regression framework $Y = X\beta^* + \epsilon$, where $X$ is a $n \times d$ deterministic design matrix, and $\epsilon$ a vector in $\mathbb{R}^n$ of independent $SG(\sigma^2)$ error variables. Assume that the true regression coefficient $\beta^*$ belongs to the set $S_0(k) = \{x \in \mathbb{R}^d : , \|x\|_0 = k\}$ of $k$-sparse vectors, where $k \le d$. Consider the estimator

$$\hat\beta = \text{argmin}_{\beta\in B_0(k)} \|Y - X\beta\|^2.$$

This the best least squares solution computed over all subsets of the coordinates of size $k$. Computationally, it requires evaluating $\binom{d}{k}$ least squares. Analyze the performance of $\hat\beta$ by showing that, with probability at lest $1-\delta$

$$\|X(\hat\beta - \beta^*)\|^2 \le C(\delta)\frac{\sigma^2 k}{n}\log\left(\frac{ed}{2k}\right),$$

where $C(\delta)$ is a constant that depends on $\delta$. Notice that, up to a logarithmic term, this is the (optimal) performance of the least squares estimator *if the support of $\beta^*$ were known.* This is something that is quite typical: the statistical price for not knowing the support of $\beta^*$ is only logarithmic (and therefore rather minimal). However, at least for the estimator $\hat\beta$, the computational price is huge. The trade-off between computational and statistical guarantees in a very important topic in the theoretical literature on high-dimensional statistics.
*Hint: follow the proof of the performance of the least squares estimator. You may want to use the fact that $\binom{d}{k} \le \left(\frac{ed}{k}\right)^k$.*

**Points:** 20 pts.

**Solution.**
Note the basic inequality

$$\|X(\hat\beta - \beta^*)\|_2^2 \le 2\epsilon^\top X(\hat\beta - \beta^*).$$

Let $r := \min\{2k, d\}$, and choose $\hat I \subset \{1, \ldots, d\}$ be satisfying $\hat I \supset \text{supp}(\hat\beta - \beta^*)$ and $|\hat I| = r$. Then there exists $\Phi_{\hat I}$ :$n \times r$ matrix where $\left\langle (\Phi_{\hat I})_i \right\rangle \supset \left\langle X_i,\, i \in \hat I \right\rangle$ with $\Phi_{\hat I}^\top \Phi_{\hat I} = I_r$. Then,

$$X(\hat\beta - \beta^*) = \Phi_{\hat I} v$$

for some unique $v \in \mathbb{R}^r$. Simplifying the basic inequality, we have

$$\begin{aligned}\|X(\hat\beta - \beta^*)\| &\le 2\epsilon^\top \frac{X(\hat\beta - \beta^*)}{\|X(\hat\beta - \beta^*)\|} \\ &= 2\epsilon^\top \frac{\Phi_{\hat I} v}{\|\Phi_{\hat I} v\|} \\ &= 2\epsilon_{\hat I}^\top \frac{v}{\|v\|},\end{aligned}$$

where $\epsilon_{\hat I} = \Phi_{\hat I}^\top \epsilon$ and we used the fact that $\|\Phi_{\hat I} v\|^2 = v^\top \Phi_{\hat I}^\top \Phi_{\hat I} v = v^\top v = \|v\|^2$. Hence,

$$\|X(\hat\beta - \beta^*)\| \le 2 \sup_{v \in S^{r-1}} \epsilon_{\hat I}^\top v = 2 \sup_{v \in B_r} \epsilon_{\hat I}^\top v,$$

where $S^{r-1} = \{x \in \mathbb{R}^r : \|x\| = 1\}$ and $B_r = \{x \in \mathbb{R}^r : \|x\| \le 1\}$. Then, as we have seen in the discretization argument in Lecture 6 (Sep 20), let $\mathcal{N}_{1/2}$ be the $\frac{1}{2}$ covering of $B_r$ in Euclidean norm. Then for all $v \in B_r$, there exists $x \in \mathcal{N}_{1/2}$ and $w \in \frac{1}{2} B_d$ such that $v = x + w$. And hence

$$\sup_{v \in B_d} \epsilon_{\hat I}^\top v \le \max_{x \in B_d} \epsilon_{\hat I}^\top x + \sup_{w \in \frac{1}{2} B_d} \epsilon_{\hat I}^\top w = \max_{x \in B_d} \epsilon_{\hat I}^\top x + \frac{1}{2} \sup_{w \in B_d} \epsilon_{\hat I}^\top w,$$

hence paraphrasing this gives

$$\sup_{v \in S^{r-1}} \epsilon_{\hat I}^\top v \le 2 \max_{x \in \mathcal{N}_{1/2}} \epsilon_{\hat I}^\top x,$$

as in Lecture 6 (Sep 20). And hence

$$\|X(\hat\beta - \beta^*)\| \le 4 \max_{x \in \mathcal{N}_{1/2}} \epsilon_{\hat I}^\top x.$$

Hence taking max over all $I \subset [d] := \{1, \dots, d\}$ with $|I| = r$ gives

$$\|X(\hat\beta - \beta^*)\| \le 4 \max_{I \subset [d] \colon |I| = r} \max_{x \in \mathcal{N}_{1/2}} \epsilon_I^\top x.$$

Now, note that $\|x\| = 1$ and $\Phi^\top \Phi = I_r$ implies $\|\Phi x\| = 1$. And from HW 2 Problem 2, $\epsilon_I^\top x = \epsilon^\top \Phi x$ with $\epsilon_i$ being i.i.d. $SG(\sigma^2)$ implies $\epsilon_I^\top x$ is $SG(\|x\|\sigma^2) = SG(\sigma^2)$. Hence from maximal inequality,

$$\begin{aligned}P\left(\frac{1}{n}\|X(\hat\beta - \beta^*)\|^2 \ge t\right) &\le P\left(\max_{I \subset [d] \colon |I| = r} \max_{x \in \mathcal{N}_{1/2}} \epsilon_I^\top x \ge \sqrt{\frac{nt}{16}}\right) \\ &\le \sum_{I \subset [d] \colon |I| = r} \sum_{x \in \mathcal{N}_{1/2}} P\left(\epsilon_I^\top x \ge \sqrt{\frac{nt}{16}}\right) \\ &\le \left(\begin{array}{c} d \\ r \end{array}\right) |\mathcal{N}_{1/2}| \exp\left(-\frac{nt}{32\sigma^2}\right).\end{aligned}$$

Then Theorem 5.10 in Lecture 5 (Sep 18) bounds the covering number as $\left|\mathcal{N}_{1/2}\right| \le \left(1 + \frac{2}{1/2}\right)^r \le 5^{2k}$, and $\binom{d}{r} \le \max\left\{1, \left(\frac{ed}{2k}\right)^{2k}\right\} = \left(\frac{ed}{2k}\right)^{2k}$ holds, and hence

$$P\left(\frac{1}{n}\|X(\hat\beta - \beta^*)\|^2 \ge t\right) \le 25^k \left(\frac{ed}{2k}\right)^{2k} \exp\left(-\frac{t}{32\sigma^2}\right).$$

Hence applying $t = \frac{32\sigma^2}{n}\left(\log\left(\frac{1}{\delta}\right) + k\log 25 + 2k\log\left(\frac{ed}{2k}\right)\right)$ gives the RHS as $\delta$. Hence with probability $1-\delta$,

$$\begin{aligned}\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 &\leq \frac{32\sigma^2}{n}\left(\log\left(\frac{1}{\delta}\right) + k\log 25 + 2k\log\left(\frac{ed}{2k}\right)\right)\\ &\leq \frac{32\sigma^2 k}{n}\left(\log\left(\frac{25}{\delta}\right) + 2\log\left(\frac{ed}{2k}\right)\right)\\ &\leq 32\left(\frac{\log(\frac{25}{\delta})}{2\log(\frac{e}{2})} + 1\right)\frac{\sigma^2 k}{n}\log\left(\frac{ed}{2k}\right).\end{aligned}$$

4. **Matrix Algebra Problems.**

(a) Problem 8.3 (You may assume the result of Problem 8.1 as given):
Prove Weyl's inequality

$$\max_{j=1,\ldots,d}|\gamma_j(Q)-\gamma_j(R)| \leq |||Q-R|||_{op}.$$

(b) Recall the spiked covariance model: $\Sigma = \theta vv^\top + I_d$, where $\theta > 0$ and $v \in \mathbb{S}^{d-1}$. Let $\hat{v}$ be another unit vector in $\mathbb{S}^{d-1}$. Show that

$$v^\top\Sigma v - \hat{v}^\top\Sigma\hat{v} = \theta\sin^2(\angle(v,\hat{v}))$$

where $\angle(v,\hat{v}) = \cos^{-1}(|v^\top\hat{v}|)$

(c) Show that

$$\left\|\hat{v}\hat{v}^\top - vv^\top\right\|_F^2 = 2\sin^2(\angle(v,\hat{v})),$$

where, for a matrix $A = (A_{i,j})$, $\|A\|_F = \sqrt{\sum_{i,j}A_{i,j}^2}$.

**Points:** 35 pts = 15 + 10 + 10.

**Solution.**
(a)
Exercise 8.1 in Wainwright implies that for $j \geq 2$, $\gamma_j(Q)$ can be expressed as

$$\gamma_j(Q) = \min_{\mathbb{V}\in\mathcal{V}_{j-1}}\max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u, Qu\rangle.$$

Note that when we understand $\mathcal{V}_0 = \{\{0\}\}$ and $\{0\}^\perp = \mathbb{R}^d$, then the above statement holds for $j=1$ as well. Suppose two symmetric matrices $Q$, $R$ and $1 \leq j \leq d$ is given. Then for every $\mathbb{V}\in\mathcal{V}_{j-1}$,

$$\begin{aligned}\max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u, Qu\rangle &= \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\left(\langle u, Ru\rangle + \langle u,(Q-R)u\rangle\right)\\ &\leq \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u, Ru\rangle + \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u,(Q-R)u\rangle\\ &\leq \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u, Ru\rangle + \max_{u\in\mathbb{S}^{d-1}}\langle u,(Q-R)u\rangle\\ &= \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}}\langle u, Ru\rangle + |||Q-P|||_{op}.\end{aligned}$$

Hence taking min over $\mathbb{V} \in \mathcal{V}_{j-1}$, $\gamma_j(Q)$ can be upper bounded as

$$\begin{aligned}\gamma_j(Q) &= \min_{\mathbb{V}\in\mathcal{V}_{j-1}} \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}} \langle u, Qu\rangle \\ &\leq \min_{\mathbb{V}\in\mathcal{V}_{j-1}} \left( \max_{u\in\mathbb{V}^\perp\cap\mathbb{S}^{d-1}} \langle u, Pu\rangle + |||Q-P|||_{op} \right) \\ &= \gamma_j(P) + |||Q-P|||_{op}.\end{aligned}$$

And hence $\gamma_j(Q) \leq \gamma_j(P) + |||Q-P|||_{op}$ for all $j$. Similarly, $\gamma_j(P) \leq \gamma_j(Q) + |||Q-P|||_{op}$ holds for all $j$, and hence $\max_{j=1,\ldots,d} |\gamma_j(Q) - \gamma_j(R)| \leq |||Q-R|||_{op}$ holds.

**Details.**

Note that the equation looks similar, but $\gamma_j(Q) - \gamma_j(R) \leq \gamma_j(Q-R)$ may not hold even when $Q$, $R$, $Q-R$ are all positive definite: when $Q = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $R = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, then $Q - R = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, and hence

$$\gamma_2(Q) - \gamma_2(R) = 1 > 0 = \gamma_2(Q-R).$$

(b)

Note that for any $u \in \mathbb{S}^{d-1}$,

$$u^\top \Sigma u = \theta u^\top v v^\top u + u^\top I_d u = \theta (u^\top v)^2 + 1 = \theta \cos^2(\angle(v,u)) + 1$$

Hence

$$\begin{aligned}v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} &= \theta(\cos^2(\angle(v,v)) - \cos^2(\angle(v,\hat{v}))) \\ &= \theta \sin^2(\angle(v,\hat{v})).\end{aligned}$$

(c)

Note that $\|A\|_F^2 = tr(AA^\top)$, hence

$$\begin{aligned}\left\|\hat{v}\hat{v}^\top - vv^\top\right\|_F^2 &= tr\left((\hat{v}\hat{v}^\top - vv^\top)(\hat{v}\hat{v}^\top - vv^\top)^\top\right) \\ &= tr\left(\hat{v}\hat{v}^\top\hat{v}\hat{v}^\top + vv^\top vv^\top - \hat{v}\hat{v}^\top vv^\top - vv^\top\hat{v}\hat{v}^\top\right) \\ &= \hat{v}^\top\hat{v}\hat{v}^\top\hat{v} + v^\top vv^\top v - \hat{v}^\top vv^\top\hat{v} - v^\top\hat{v}\hat{v}^\top v \\ &= 2(1 - (\hat{v}^\top v)^2) = 2\sin^2(\angle(v,\hat{v})).\end{aligned}$$