36-755, Fall 2017
Homework 5 Solution

Due Wed Nov 15 by 5:00pm in Jisu's mailbox

Points: 100+1 pts total for the assignment.

We first review some basic relations with norms and the singular value decomposition on matrices.

Lemma 0.1 *For any matrix $A \in \mathbb{R}^{m\times n}$, let $A = U\Sigma V^\top$ be its singular value decomposition with $U^\top U = UU^\top = I_m$ and $V^\top V = VV^\top = I_n$, and Σ being $m \times n$ diagonal matrix with nonnegative diagonal values $\sigma_1, \cdots, \sigma_{\min(m,n)}$.*

(i) Frobenius norm can be calculated as $\|A\|_F = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2}$.

(ii) Operator 2-norm $\|A\|_{op}$ can be calculated as $\|A\|_{op} = \max_i \sigma_i$.

Proof (i)
Note that Frobenius norm equals $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2} = \sqrt{tr(A^\top A)}$. And hence from the singular value decomposition of A,

$$\begin{aligned}
\|A\|_F = \sqrt{tr(A^\top A)} &= \sqrt{tr(U\Sigma V^\top V\Sigma^\top U^\top)} \\
&= \sqrt{tr(U\Sigma\Sigma^\top U^\top)} = \sqrt{tr(\Sigma\Sigma^\top U^\top U)} \\
&= \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2}.
\end{aligned}$$

(ii)
Let $r = \min(m, n)$, and let columns of U and V be $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$, respectively. Expand any $w \in \mathbb{R}^n$ using $v_1, \ldots, v_n$ as basis, so that $w = \sum_{i=1}^n a_i v_i$ with $a_1, \ldots, a_n \in \mathbb{R}^n$. Then

$$Aw = \sum_{i=1}^n a_i A v_i = \sum_{i=1}^r a_i \sigma_i u_i,$$

and hence

$$\|Aw\|_2 = \sqrt{\sum_{i=1}^r \sigma_i^2 a_i^2} \le \max_{1\le i\le r} \sigma_i \sqrt{\sum_{i=1}^n a_i^2} = \left(\max_{1\le i\le r} \sigma_i\right) \|w\|_2,$$

and sufficient condition for the equality is when $a_1 = 1$ and $a_2 = \cdots = a_n = 0$, i.e. $w = v_1$. Hence

$$\|A\|_{op} = \sup_{w\neq 0} \frac{\|Aw\|_2}{\|w\|_2} = \max_{1\le i\le r} \sigma_i.$$

1. In this exercise you will fill in some of the details from the proof of the upper bound for sparse PCA under a spike covariance model.

(a) For $p \geq 1$ the Schatten p-norm of a $n \times m$ matrix A is the ℓ_p norm of its singular values:

$$\|A\|_p = \left(\sum_{i=1}^{r} \sigma_i^p\right)^{1/p},$$

where $\sigma \geq \sigma_2 \geq \ldots \sigma_r \geq 0$ are the singular values of A and $r = \min\{m, n\}$. Prove the non-commutative Hölder inequality for conformal matrices A and B:

$$|\mathrm{tr}(A^\top B)| \leq \|A\|_p \|B\|_q,$$

for all $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

(b) Let u and v be two unit norm vectors in $\mathbb{R}^d$. Show that

$$\|uu^\top - vv^\top\|_{\mathrm{op}} = \|uu^\top - vv^\top\|_F = \sqrt{2 - 2(u^\top v)^2} = \sqrt{2\sin^2(\angle(u, v))}.$$

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

Points: 15 pts = 5 + 10.

Solution.

(a)

Let A be $n \times m_1$ matrix and let B be $n \times m_2$ matrix, and let $r = \min\{n, m_1\}$, $s = \min\{n, m_2\}$. Let $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ be singular values of A, then from singular value decomposition, there exists an orthonormal set of vectors $\{u_1, \ldots, u_r\} \subset \mathbb{R}^n$ and an orthonormal set of vectors $\{v_1, \ldots, v_r\} \subset \mathbb{R}^{m_1}$ such that $A = \sum_{i=1}^r \sigma_i u_i v_i^\top$. These $\{u_i\}$'s are first r columns of U and $\{v_i\}$'s are first r columns of V, where $A = U\Sigma V^\top$ is the singular value decomposition of A. Similarly, let $\lambda_1 \geq \cdots \geq \lambda_s \geq 0$ be singular values of B, then there exists an orthonormal set of vectors $\{w_1, \ldots, w_s\} \subset \mathbb{R}^n$ and an orthonormal set of vectors $\{t_1, \ldots, t_s\} \subset \mathbb{R}^s$ such that $B = \sum_{j=1}^s \lambda_j w_j t_j^\top$. Note that $\|A\|_p = (\sum_{i=1}^r \sigma_i^p)^{1/p}$ and $\|B\|_q = \left(\sum_{j=1}^s \lambda_j^q\right)^{1/q}$. Meanwhile, LHS can be expanded as

$$\begin{aligned}
\left|\mathrm{tr}(A^\top B)\right| &= \left|\mathrm{tr}\left(\left(\sum_{i=1}^{r} \sigma_i v_i u_i^\top\right)\left(\sum_{j=1}^{s} \lambda_j w_j t_j^\top\right)\right)\right| \\
&= \left|\sum_{i=1}^{r}\sum_{j=1}^{s} \sigma_i \lambda_j \mathrm{tr}\left(v_i (u_i^\top w_j) t_j^\top\right)\right| \\
&\leq \sum_{i=1}^{r}\sum_{j=1}^{s} \sigma_i \lambda_j \left|u_i^\top w_j\right| \left|\mathrm{tr}\left(v_i t_j^\top\right)\right|.
\end{aligned}$$

Now, note that for any $u, w \in \mathbb{R}^n$ with $\|u\| = \|v\| = 1$, Cauchy-Schwarz inequality yields $\left|u^\top w\right| \leq \|u\|\|w\| = 1$. Also, for any $v \in \mathbb{R}^{m_1}$ and $t \in \mathbb{R}^{m_2}$ with $\|v\| = \|t\| = 1$, Cauchy-Schwarz inequality yields

$$\left|\mathrm{tr}\left(v_i t_j^\top\right)\right| = \sum_{i=1}^{\min\{m_1, m_2\}} v_i t_i \leq \sqrt{\sum_{i=1}^{\min\{m_1, m_2\}} v_i^2} \sqrt{\sum_{i=1}^{\min\{m_1, m_2\}} t_i^2} \leq \|v\|\|t\| = 1.$$

Hence applying this and further applying Hölder's inequality gives

$$\begin{aligned}\left|\operatorname{tr}(A^\top B)\right| &\leq \sum_{i=1}^{r}\sum_{j=1}^{s}\sigma_i\lambda_j \\ &\leq \left(\sum_{i=1}^{r}\sigma_i^p\right)^{\frac{1}{p}}\left(\sum_{j=1}^{s}\lambda_j^q\right)^{\frac{1}{q}} = \|A\|_p\|B\|_q.\end{aligned}$$

Note that above inequality holds for $p = 1, q = \infty$ or $p = \infty, q = 1$ as well, where $\|A\|_\infty = \max_{1\leq i\leq r}\sigma_i$.

(b)

The rest of the inequality follows from HW4 Problem 4(c), hence we are only left to show the first inequality $\sqrt{2}\|uu^\top - vv^\top\|_{\text{op}} = \|uu^\top - vv^\top\|_F$.

Note that the rank of $uu^\top - vv^\top$ is at most 2, so nonzero eigenvalues can be at most 2. Hence $uu^\top - vv^\top$ can be expressed as

$$uu^\top - vv^\top = \lambda_1 w_1 w_1^\top + \lambda_2 w_2 w_2^\top,$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$ and $w_1, w_2 \in \mathbb{R}^d$ with $w_1 \perp w_2$. Then from

$$\lambda_1 + \lambda_2 = tr(uu^\top - vv^\top) = u^\top u - v^\top v = 0,$$

$\lambda_2 = -\lambda_1$, i.e. there exists $\lambda > 0$ such that

$$uu^\top - vv^\top = \lambda w_1 w_1^\top - \lambda w_2 w_2^\top.$$

Then since $uu^\top - vv^\top$ is real symmetric, all the singular values are absolute values of eigenvalues. Hence singular values of $uu^\top - vv^\top$ are $\lambda, \lambda, 0, \ldots, 0$. Hence from Lemma 0.1,

$$\|uu^\top - vv^\top\|_{\text{op}} = \lambda \quad \text{and} \quad \|uu^\top - vv^\top\|_F = \sqrt{\lambda^2 + \lambda^2} = \sqrt{2}\lambda,$$

and hence $\sqrt{2}\|uu^\top - vv^\top\|_{\text{op}} = \|uu^\top - vv^\top\|_F$ holds.

2. This is a result that I cited when discussing spectra clustering for stochastic block models. A random matrix A of dimension $n \times m$ is sub-Gaussian with parameter σ^2, written as $A \in SG_{m,n}(\sigma^2)$, when $y^\top Ax$ is $SG(\sigma^2)$ for any $y \in \mathbb{S}^{n-1}$ and $x \in \mathbb{S}^{m-1}$. You may assume that $\mathbb{E}[A] = 0$ (or otherwise replace A by $A - \mathbb{E}[A]$).

   (a) Suppose that the entries of A are independent variables that are $SG(\sigma^2)$. Show that $A \in SG_{m,n}(\sigma^2)$.

   (b) Let $A \in SG_{n,m}(\sigma^2)$ and recall that the operator norm of A is

   $$\|A\|_{\text{op}} = \max_{x\in\mathbb{R}^m, x\neq 0}\frac{\|Ax\|}{\|x\|} = \max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax.$$

   Show that, for some $C > 0$,

   $$\mathbb{E}\left[\|A\|_{\text{op}}\right] \leq C\left(\sqrt{n} + \sqrt{m}\right).$$

   (c) Find a concentration inequality for $\|A\|_{\text{op}}$.

Hint: work with a 1/4 net for $\mathbb{S}^{n-1}$ and a 1/4 net for $\mathbb{S}^{m-1}$.

Points: 25 pts = 5 + 10 + 10.

Solution.
Assume that $\mathbb{E}[A] = 0$.
(a)
For all $y \in \mathbb{S}^{n-1}$, $x \in \mathbb{S}^{m-1}$, and $\lambda \in \mathbb{R}$, by using the entries of A being independent and $SG(\sigma^2)$, the mgf of $y^\top Ax$ is bounded as

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda y^\top Ax\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n}\sum_{j=1}^{m} y_i A_{ij} x_j\right)\right] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m} \mathbb{E}\left[\exp\left((\lambda y_i x_j) A_{ij}\right)\right] \quad \text{(using independence of } A_{ij}\text{)} \\
&\le \prod_{i=1}^{n}\prod_{j=1}^{m} \exp\left(\frac{1}{2}\sigma^2 (\lambda y_i x_j)^2\right) \\
&= \exp\left(\frac{1}{2}\sigma^2\lambda^2 \left(\sum_{i=1}^{n} y_i^2\right)\left(\sum_{j=1}^{m} x_j^2\right)\right) \\
&= \exp\left(\frac{1}{2}\sigma^2\lambda^2\right), \quad (y \in \mathbb{S}^{n-1},\, x \in \mathbb{S}^{m-1})
\end{aligned}$$

hence $A \in SG_{m,n}(\sigma^2)$.
(b)
Let $\mathcal{B}_n \subset \mathbb{S}^{n-1}$ and $\mathcal{B}_m \subset \mathbb{S}^{m-1}$ be the $\frac{1}{4}$-net for $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ with respect to l_2 distances. Then from Lecture note 5(Sep 18),

$$|\mathcal{B}_n| \le \frac{Vol\left(\frac{2}{1/4}\mathcal{B}(0,1) + \mathcal{B}(0,1)\right)}{Vol(\mathcal{B}(0,1))} = \frac{Vol\left(9\mathcal{B}(0,1)\right)}{Vol(\mathcal{B}(0,1))} = 9^n,$$

and similarly $|\mathcal{B}_m| \le 9^m$ holds. Also for all $y \in \mathbb{S}^{n-1}$ and $x \in \mathbb{S}^{m-1}$, there exists $y' \in \frac{1}{4}\mathbb{S}^{n-1}$, $x' \in \frac{1}{4}\mathbb{S}^{m-1}$, $w \in \mathcal{B}_n$, $z \in \mathcal{B}_m$ with $y = w + y'$ and $x = z + x'$. Hence

$$\begin{aligned}
&\max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax \\
&\le \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az + \max_{w\in\mathcal{B}_n, x\in\frac{1}{4}\mathbb{S}^{m-1}} w^\top Ax' + \max_{y'\in\frac{1}{4}\mathbb{S}^{n-1}, z\in\mathcal{B}_m} y'^\top Az + \max_{y'\in\frac{1}{4}\mathbb{S}^{n-1}, x'\in\frac{1}{4}\mathbb{S}^{m-1}} y'^\top Ax' \\
&= \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az + \frac{1}{4}\max_{w\in\mathcal{B}_n, x\in\mathbb{S}^{m-1}} w^\top Ax + \frac{1}{4}\max_{y\in\mathbb{S}^{n-1}, z\in\mathcal{B}_m} y^\top Az + \frac{1}{16}\max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax.
\end{aligned} \tag{1}$$

Now, note that

$$\begin{aligned}
\max_{w\in\mathcal{B}_n, x\in\mathbb{S}^{m-1}} w^\top Ax &\le \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az + \max_{w\in\mathcal{B}_n, x'\in\frac{1}{4}\mathbb{S}^{m-1}} w^\top Ax' \\
&\max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az + \frac{1}{4}\max_{w\in\mathcal{B}_n, x\in\mathbb{S}^{m-1}} w^\top Ax,
\end{aligned}$$

so $\max_{w\in\mathcal{B}_n, x\in\mathbb{S}^{m-1}} w^\top Ax \leq \frac{4}{3} \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az$, and similarly $\max_{y\in\mathbb{S}^{n-1}, z\in\mathcal{B}_m} y^\top Az \leq \frac{4}{3} \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az$. And applying these to (1) yields

$$\max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax \leq \frac{16}{9} \max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az. \tag{2}$$

Hence these yields the upper bound of $\mathbb{E}\left[\|A\|_{\text{op}}\right]$ as

$$\begin{aligned}
\mathbb{E}\left[\|A\|_{\text{op}}\right] &= \mathbb{E}\left[\max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax\right] \\
&\leq \frac{16}{9}\mathbb{E}\left[\max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az\right] \quad \text{(using (2))} \\
&\leq \frac{16\sigma}{9}\sqrt{2\log\left(|\mathcal{B}_m|\,|\mathcal{B}_n|\right)} \\
&\leq \frac{16\sigma}{9}\sqrt{2(m+n)\log 9} \\
&< \frac{32\sigma\sqrt{\log 3}}{9}\left(\sqrt{m}+\sqrt{n}\right) \quad \text{(using } \sqrt{m+n} < \sqrt{m}+\sqrt{n} \text{ for } m,n>0\text{)}.
\end{aligned}$$

(c)
From (2), the tail probability $\mathbb{P}\left(\|A\|_{op} \geq t\right)$ can be bounded as

$$\begin{aligned}
\mathbb{P}\left(\|A\|_{op} \geq t\right) &= \mathbb{P}\left(\max_{y\in\mathbb{S}^{n-1}, x\in\mathbb{S}^{m-1}} y^\top Ax \geq t\right) \\
&\leq \mathbb{P}\left(\frac{16}{9}\max_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} w^\top Az \geq t\right) \quad \text{(using (2))} \\
&\leq \sum_{w\in\mathcal{B}_n, z\in\mathcal{B}_m} \mathbb{P}\left(w^\top Az \geq \frac{9}{16}t\right).
\end{aligned} \tag{3}$$

Then for each $w \in \mathcal{B}_n$ and $z \in \mathcal{B}_m$, since $w^\top Az \in SG(\sigma^2)$ with $\mathbb{E}\left[w^\top Az\right] = 0$, so by applying Hoeffding's inequality,

$$\mathbb{P}\left(w^\top Az \geq \frac{9}{16}t\right) \leq \exp\left(-\frac{81t^2}{512\sigma^2}\right).$$

Hence applying this to (3) gives the bound for the tail probability $\mathbb{P}\left(\|A\|_{op} \geq t\right)$ as

$$\begin{aligned}
\mathbb{P}\left(\|A\|_{op} \geq t\right) &\leq |\mathcal{B}_m|\,|\mathcal{B}_n| \exp\left(-\frac{81t^2}{512\sigma^2}\right) \\
&\leq 9^{m+n}\exp\left(-\frac{81t^2}{512\sigma^2}\right).
\end{aligned}$$

3. Exercise 8.4:

   Show that the orthogonal matrix $V \in \mathbb{R}^{d\times r}$ maximizing the criterion (8.2) has columns formed by the top r eigenvectors of $\Sigma = cov(X)$:

$$\mathbb{E}\left\|V^\top X\right\|_2^2 = \sum_{j=1}^{r} \mathbb{E}\left[\langle v_j, X\rangle^2\right].$$

Points: 15 pts.

Solution.

$$
\begin{aligned}
\mathbb{E}\left\|V^\top X\right\|_2^2 &= \mathbb{E}\left[X^\top V V^\top X\right] = \mathbb{E}\left[tr\left(X^\top V V^\top X\right)\right] \\
&= \mathbb{E}\left[tr\left(V V^\top X X^\top\right)\right] = tr\left(V V^\top \mathbb{E}\left[X X^\top\right]\right) \\
&= tr\left(V V^\top \Sigma\right).
\end{aligned}
$$

Let eigenvalues and corresponding eigenvectors of Σ be $\lambda_1 \geq \cdots \geq \lambda_d$ and $u_1, \ldots, u_d$. Then

$$
\begin{aligned}
\mathbb{E}\left\|V^\top X\right\|_2^2 &= tr\left(V V^\top \sum_{i=1}^d \lambda_i u_i u_i^\top\right) = \sum_{i=1}^d \lambda_i tr\left(V V^\top u_i u_i^\top\right) \\
&= \sum_{i=1}^d \lambda_i tr\left(u_i^\top V V^\top u_i\right) = \sum_{i=1}^d \lambda_i \left\|V^\top u_i\right\|_2^2.
\end{aligned}
$$

Now, note that $(VV^\top)V = V(V^\top V) = VI_r = V$, and hence for any $w \in Col(V)$, $(VV^\top)w = w$, where $Col(V)$ is the column space of V. Also, for any $w^\perp \in Col(V)^\perp$, $V^\top w^\perp = 0$, which implies $VV^\top w^\perp = 0$. Hence for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$, let $u = w + w^\perp$ with $w \in Col(V)$ and $w^\perp \in Col(V)^\perp$, then

$$
\begin{aligned}
\left\|V^\top u\right\|_2^2 &= (w + w^\perp)^\top V V^\top (w + w^\top) \\
&= (w + w^\perp)^\top w = \|w\|_2^2 \leq \|u\|_2^2 = 1,
\end{aligned}
$$

hence $V^\top u_i$ satisfy

$$
0 \leq \left\|V^\top u_i\right\|_2^2 \leq 1. \tag{4}
$$

Also, $\left\|V^\top u_i\right\|_2^2 = 1$ holds if and only if $u_i \in Col(V)$ (i.e. $Col(V)^\perp$ part is 0) and $\left\|V^\top u_i\right\|_2^2 = 0$ if and only if $u_i \in Col(V)^\perp$ (i.e. $Col(V)$ part is 0). Also, note that

$$
\begin{aligned}
\sum_{i=1}^d \left\|V^\top u_i\right\|_2^2 &= \sum_{i=1}^d tr\left(V V^\top u_i u_i^\top\right) = tr\left(V V^\top \sum_{i=1}^d u_i u_i^\top\right) \\
&= tr\left(V V^\top I_d\right) = tr\left(V^\top V\right) = tr\left(I_r\right) = r.
\end{aligned} \tag{5}
$$

Now, consider the following optimization problem:

$$
\begin{aligned}
&\text{maximize } \sum_{i=1}^d \lambda_i x_i, \\
&\text{subject to } \forall i,\ 0 \leq x_i \leq 1,\ \sum_{i=1}^d x_i = r.
\end{aligned} \tag{6}
$$

Then, the constraints imply that $1 \leq \forall i \leq d$, $\sum_{j=1}^i x_j \leq \min\{i, r\}$. Let $\lambda_{d+1} = 0$ for convenience, and note that

$$\begin{aligned}
\sum_{i=1}^{d} \lambda_i x_i &= \sum_{i=1}^{d} (\lambda_i - \lambda_{i+1}) \left(\sum_{j=1}^{i} x_j \right) \\
&\leq \sum_{i=1}^{r} (\lambda_i - \lambda_{i+1}) i + \sum_{i=r+1}^{d} (\lambda_i - \lambda_{i+1}) r \\
&= \sum_{i=1}^{r} \lambda_i,
\end{aligned}$$

and the equality holds if and only if $\sum_{j=1}^{i} x_j = \min \{i, r\}$ whenever $\lambda_i > \lambda_{i+1}$. Hence

$$x_1 = \cdots = x_r = 1,\ x_{r+1} = \cdots = x_d = 0 \tag{7}$$

maximizes the objective function $\sum_{i=1}^{d} \lambda_i x_i$, and if we further have $\lambda_r > \lambda_{r+1}$, it is the unique maximizer.

Hence, applying (4) and (5) to (6) implies that

$$\mathbb{E} \left\| V^\top X \right\|_2^2 = \sum_{i=1}^{d} \lambda_i \left\| V^\top u_i \right\|_2^2 \leq \sum_{i=1}^{r} \lambda_i.$$

And from (7), the equality holds if

$$\left\| V^\top u_1 \right\|_2^2 = \cdots = \left\| V^\top u_r \right\|_2^2 = 1 \text{ and } \left\| V^\top u_{r+1} \right\|_2^2 = \cdots = \left\| V^\top u_d \right\|_2^2 = 0, \tag{8}$$

and only if holds under $\lambda_r > \lambda_{r+1}$. Then from equality conditions from (4), (8) is equivalent to

$$Col(V) = \langle u_1, \ldots, u_r \rangle,$$

where $\langle u_1, \ldots, u_r \rangle \subset \mathbb{R}^d$ is the linear subspace spanned by $u_1, \ldots, u_r$. In conclusion,

$$\mathbb{E} \left\| V^\top X \right\|_2^2 \leq \sum_{i=1}^{r} \lambda_i,$$

and the equality holds if $Col(V) = \langle u_1, \ldots, u_r \rangle$, and under $\lambda_r > \lambda_{r+1}$, it is necessary as well. In particular, when top r eigenvectors $\{u_1, \ldots, u_r\}$ form the columns of V, $\mathbb{E} \left\| V^\top X \right\|_2^2$ is maximized.

4. Suppose we observe an i.i.d. sample $X_1, \ldots, X_n$ from the mixture distribution

$$\frac{1}{2} P_1 + \frac{1}{2} P_2,$$

where $P_1 = N_d(\mu, I_d)$ and $P_2 = N_d(-\mu, I_d)$, with $\mu \in \mathbb{R}^d$ a non-zero vector. Ours task is to cluster the sample points into two groups, where points in the same group originated from the same component of the mixture (i.e. Either P_1 or P_2)

We will use spectral clustering: compute the leading eigenvector $\hat{\nu}$ of the empirical covariance matrix and cluster the points depending on the sign of $X_i^\top \hat{\nu}$, $i = 1, \ldots, n$.

Use the Davis-Kahan theorem to derive an upper bound on the proportion of misclustered nodes.

Points: 20 + 1 pts.

Solution.

X has the same distribution as $W\mu + Y$, where $W \sim \mathit{Rademacher}$, $Y \sim N_d(0, I_d)$. Then $\mathbb{E}[X] = 0$ and

$$\Sigma := Var\,[X] = \mathbb{E}\left[XX^\top\right] = \mu\mu^\top + I_d.$$

Hence by letting $v := \frac{\mu}{\|\mu\|}$, then v is the leading eigenvector of Σ with eigenvalue $1 + \|\mu\|^2$. Hence from Davis-Kahan theorem,

$$\sin\angle(v, \hat{v}) \leq \frac{2\left\|\hat{\Sigma} - \Sigma\right\|_{op}}{\|\mu\|^2}. \tag{9}$$

Then note that the number of misclustered nodes is upper bounded by

$$\sum_{i=1}^{n} I\left(\left\{X_i^\top \hat{v} \leq 0, W_i = 1\right\} \cup \left\{X_i^\top \hat{v} > 0, W_i = -1\right\}\right).$$

Hence the misclustering proportion R is bounded as

$$\begin{aligned} R &\leq \mathbb{P}\left(X_i^\top \hat{v} \leq 0, W_i = 1\right) + \mathbb{P}\left(X_i^\top \hat{v} > 0, W_i = -1\right) \\ &\leq \mathbb{P}\left((\mu + Y_i)^\top \hat{v} \leq 0\right) + \mathbb{P}\left((-\mu + Y_i)^\top \hat{v} > 0\right). \end{aligned}$$

Now, consider the first event $\left\{(\mu + Y_i)^\top \hat{v} \leq 0\right\}$. Note that

$$\begin{aligned} \langle \mu + Y_i, \hat{v} \rangle &= \langle \mu, \hat{v} \rangle + \langle Y_i, \hat{v} \rangle \\ &\geq \langle \mu, \hat{v} \rangle - \|Y_i\|, \end{aligned}$$

Hence $\langle \mu + Y_i, \hat{v} \rangle \leq 0$ implies

$$\|Y_i\| \geq \langle \mu, \hat{v} \rangle - \langle \mu + Y_i, \hat{v} \rangle \geq \langle \mu, \hat{v} \rangle.$$

Its geometric interpretation is in Figure 1. And applying $-Y_i$ instead of Y_i gives that $\langle -\mu + Y_i, \hat{v} \rangle > 0$ implies $\|Y_i\| > \langle \mu, \hat{v} \rangle$ as well. Since $\hat{v}$ is aligned with μ so that $\langle \mu, \hat{v} \rangle \geq 0$, and hence the misclustering proportion R is bounded as

$$R \leq 2\mathbb{P}\left(\|Y_i\|^2 \geq (\langle \mu, \hat{v} \rangle)^2\right).$$

Then applying the result of Davis-Kahan Theorem in (9) implies

$$\begin{aligned} (\langle \mu, \hat{v} \rangle)^2 &= \|\mu\|^2 \cos^2\angle(v, \hat{v}) = \|\mu\|^2 \left(1 - \sin^2\angle(v, \hat{v})\right) \\ &\geq \|\mu\|^2 \left(1 - \frac{4\left\|\hat{\Sigma} - \Sigma\right\|_{op}^2}{\|\mu\|^4}\right). \end{aligned}$$

Hence the misclustering proportion R is bounded as

$$R \leq 2\mathbb{P}\left(\|Y_i\|^2 \geq \|\mu\|^2 \left(1 - \frac{4\left\|\hat{\Sigma} - \Sigma\right\|_{op}^2}{\|\mu\|^4}\right)\right).$$

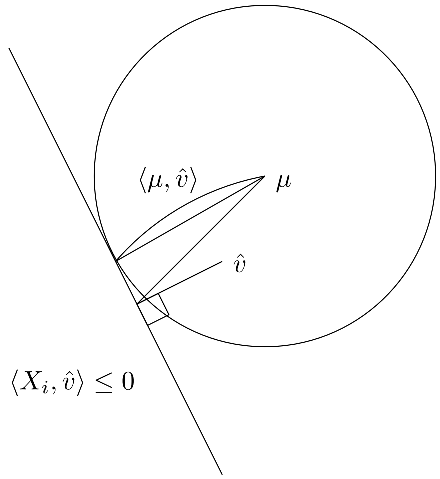

Figure 1: Geometrical interpretation of $\langle \mu + Y_i, \hat{v} \rangle \leq 0$ implying $\|Y_i\| \geq \langle \mu, \hat{v} \rangle$. $\langle \mu + Y_i, \hat{v} \rangle \leq 0$ is equivalent to $X_i = \mu + Y_i$ lying on the bottom-left region of the hyperplane which is perpendicular to $\hat{v}$. Then X_i lies outside of a sphere centered at μ and of radius $\langle \mu, \hat{v} \rangle$. Hence $\|X_i - \mu\| = \|Y_i\| \geq \langle \mu, \hat{v} \rangle$.

Now, note that for any $v \in \mathbb{R}^d$ with $\|v\| = 1$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda v^\top X_i\right)\right] &= \frac{1}{2}\mathbb{E}\left[\exp\left(\lambda v^\top(\mu + Y_i)\right)\right] + \frac{1}{2}\mathbb{E}\left[\exp\left(\lambda v^\top(-\mu + Y_i)\right)\right] \\
&= \frac{1}{2}\exp\left(\lambda v^\top \mu + \frac{1}{2}\lambda^2\right) + \frac{1}{2}\exp\left(-\lambda v^\top \mu + \frac{1}{2}\lambda^2\right) \\
&= \exp\left(\frac{1}{2}\lambda^2\right)\cosh\left(\lambda v^\top \mu\right) \\
&\leq \exp\left(\frac{1}{2}\lambda^2(1 + (v^\top \mu)^2)\right) \quad (\because \cosh x \leq \exp(x^2/2)) \\
&\leq \exp\left(\frac{1}{2}\lambda^2(1 + \|\mu\|^2)\right), \quad (\because \text{Cauchy-Schwarz})
\end{aligned}$$

Hence $X_i \in SG(1 + \|\mu\|^2)$. Hence from Lecture note 6 on Sep 20, for all $\delta \in (0, 1)$, there exists some $C > 0$ such that

$$\mathbb{P}\left(\left\|\hat{\Sigma} - \Sigma\right\|_{op} \leq C(1 + \|\mu\|^2)\max\left\{\sqrt{\frac{d + \log(2/\delta)}{n}}, \frac{d + \log(2/\delta)}{n}\right\}\right) \geq 1 - \delta. \tag{10}$$

Let E be the event that $\left\|\hat{\Sigma} - \Sigma\right\|_{op} \leq C(1 + \|\mu\|^2)\max\left\{\sqrt{\frac{d+\log(2/\delta)}{n}}, \frac{d+\log(2/\delta)}{n}\right\}$ happens. Then $\|Y_i\|^2 \geq \|\mu\|^2\left(1 - \frac{4\|\hat{\Sigma}-\Sigma\|_{op}^2}{\|\mu\|^4}\right)$ and E implies

$$\begin{aligned}
\|Y_i\|^2 &\geq \|\mu\|^2\left(1 - \frac{4C^2\left(1 + \|\mu\|^2\right)^2}{\|\mu\|^4}\max\left\{\frac{d + \log(2/\delta)}{n}, \left(\frac{d + \log(2/\delta)}{n}\right)^2\right\}\right) \\
&\geq \|\mu\|^2\left(1 - 8C^2\left(1 + \frac{1}{\|\mu\|^4}\right)\max\left\{\frac{d + \log(2/\delta)}{n}, \left(\frac{d + \log(2/\delta)}{n}\right)^2\right\}\right) \\
&:= K_{n,d,\delta,\|\mu\|,C},
\end{aligned}$$

where $K_{n,d,\delta,\|\mu\|,C} = \|\mu\|^2 \left(1 - 8C^2\left(1 + \frac{1}{\|\mu\|^4}\right)\max\left\{\frac{d+\log(2/\delta)}{n}, \left(\frac{d+\log(2/\delta)}{n}\right)^2\right\}\right)$. And hence

$$\begin{aligned}
R &\le 2\mathbb{P}\left(\left\{\|Y_i\|^2 \ge \|\mu\|^2\left(1 - \frac{4\left\|\hat{\Sigma} - \Sigma\right\|_{op}^2}{\|\mu\|^4}\right)\right\} \cap E\right) \\
&\quad + 2\mathbb{P}\left(\left\{\|Y_i\|^2 \ge \|\mu\|^2\left(1 - \frac{4\left\|\hat{\Sigma} - \Sigma\right\|_{op}^2}{\|\mu\|^4}\right)\right\} \cap E^c\right) \\
&\le 2\mathbb{P}\left(\|Y_i\|^2 \ge K_{n,d,\delta,\|\mu\|,C}\right) + 2\mathbb{P}(E^c) \\
&\le 2\mathbb{P}\left(\|Y_i\|^2 \ge K_{n,d,\delta,\|\mu\|,C}\right) + 2\delta.
\end{aligned}$$

Now, since $Y_i \sim N_d(0, I_d)$, $\|Y_i\|^2 \sim \chi_d^2$, chi-square distribution with degree d. Hence

$$\mathbb{P}\left(\|Y_i\|^2 \ge K_{n,d,\delta,\|\mu\|,C}\right) = 1 - F_{\chi_d^2}(K_{n,d,\delta,\|\mu\|,C}),$$

where $F_{\chi_d^2}$ is the CDF function for χ_d^2. To conclude, the misclustering proportion R is bounded as

$$R \le 2 \inf_{\delta\in(0,1)} \left\{1 - F_{\chi_d^2}(K_{n,d,\delta,\|\mu\|,C}) + \delta\right\},$$

where $K_{n,d,\delta,\|\mu\|,C} = \|\mu\|^2 \left(1 - 8C^2\left(1 + \frac{1}{\|\mu\|^4}\right)\max\left\{\frac{d+\log(2/\delta)}{n}, \left(\frac{d+\log(2/\delta)}{n}\right)^2\right\}\right)$.

Note that when getting the bound (10), matrix Bernstein inequality is a bit inconvenient to use since $\|X_i\|$ is not bounded. There is 1 bonus point for using the correct covariance bound for getting (10).

5. Exercise 4.3:

   (a) Recall the re-centered function class $\bar{\mathcal{F}} = \{f - \mathbb{E}[f] \,|\, f \in \mathcal{F}\}$. Show that

   $$\mathbb{E}_{X,\epsilon}\left[||R_n||_{\bar{\mathcal{F}}}\right] \ge \mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}[f]|}{\sqrt{n}}.$$

   (b) Use concentration results to complete the proof of Proposition 4.2.

Points: 15 pts = 10 + 5.

Solution.

(a)

Note that for any $\bar{f} \in \bar{\mathcal{F}}$, there exists $f \in \mathcal{F}$ such that $\bar{f}(x) = f(x) + \mathbb{E}[f]$. Hence

$$\begin{aligned}
\|R_n\|_{\bar{\mathcal{F}}} &= \sup_{\bar{f}\in\bar{\mathcal{F}}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \bar{f}(X_i)\right| \\
&= \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) - \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}[f]\right| \\
&\ge \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| - \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}[f]\right|. \qquad (11)
\end{aligned}$$

Now, note that by applying Cauchy Schwarz, $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|\right]$ can be bounded as

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|\right] \leq \frac{1}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{n}\epsilon_i\right)^2\right]} = \frac{1}{n}\sqrt{\mathbb{E}\left[\sum_{i=1}^{n}\epsilon_i^2 + \sum_{i\neq j}\epsilon_i\epsilon_j\right]}$$
$$= \frac{1}{n}\sqrt{n} = \frac{1}{\sqrt{n}}.$$

By plugging this to expectation of (11), we get

$$\begin{aligned}\mathbb{E}_{X,\epsilon}\left[||R_n||_{\bar{\mathcal{F}}}\right] &\geq \mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \sup_{f\in\mathcal{F}}|\mathbb{E}[f]|\,\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|\right] \\ &\geq \mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{\sqrt{n}}.\end{aligned}$$

(b)
Note that from (a) and Proposition 4.1 (a), we have

$$\begin{aligned}\mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] &\geq \mathbb{E}_{X,\epsilon}\left[\frac{1}{2}\|R_n\|_{\bar{\mathcal{F}}}\right] \\ &\geq \frac{1}{2}\mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}}.\end{aligned} \tag{12}$$

Now, for each $f\in\mathcal{F}$ let $\bar{f}(x) = f(x) - \mathbb{E}[f]$ as in (a), and let $G(x_1,\cdots,x_n) = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\bar{f}(x_i)\right|$ so that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = G(X_1,\cdots,X_n)$. Note that if $x,y\in\mathcal{X}^n$ satisfies $x_j = y_j$ for all $j\neq i$, then for any $f\in\mathcal{F}$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{j=1}^{n}\bar{f}(x_j)\right| - \sup_{h\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\bar{h}(y_j)\right| &\leq \left|\frac{1}{n}\sum_{j=1}^{n}\bar{f}(x_j)\right| - \left|\frac{1}{n}\sum_{j=1}^{n}\bar{f}(y_j)\right| \\ &\leq \left|\frac{1}{n}\sum_{j=1}^{n}\bar{f}(x_j) - \frac{1}{n}\sum_{j=1}^{n}\bar{f}(y_j)\right| \\ &= \frac{1}{n}\left|\bar{f}(x_j) - \bar{f}(y_j)\right| \\ &\leq \frac{2}{n}\|f\|_\infty \leq \frac{2b}{n}.\end{aligned}$$

Hence taking supremum over $f\in\mathcal{F}$ yields $G(x) - G(y) \leq \frac{2b}{n}$. Since the same argument can be applied to get $G(y) - G(x) \leq \frac{2b}{n}$, so $|G(x) - G(y)| \leq \frac{2b}{n}$. Hence from Bounded differences inequality,

$$\mathbb{P}\left(G(X) - \mathbb{E}\left[G(X)\right] \geq \delta\right) \leq \exp\left(-\frac{n\delta^2}{2b^2}\right).$$

Then applying $G(X) = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ and (12) to this gives

$$\begin{aligned}&\mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \left(\frac{1}{2}\mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}}\right) \geq \delta\right) \\ &\leq \mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] \geq \delta\right) \leq \exp\left(-\frac{n\delta^2}{2b^2}\right).\end{aligned}$$

Hence $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le \frac{1}{2}\mathbb{E}_{X,\epsilon}\left[||R_n||_{\mathcal{F}}\right] - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}} + \delta$ with $\mathbb{P}$-probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$.

6. **Massart's finite class Lemma** Let $\mathcal{A}$ be a finite subset of $\mathbb{R}^n$ and let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ be a vector of i.i.d. Rademacher variables. Show that

$$\mathbb{E}\left[\frac{1}{n}\sup_{a\in\mathcal{A}} a^\top \epsilon\right] \le D\frac{\sqrt{2\log|\mathcal{A}|}}{n}$$

where $D = \max_{a\in\mathcal{A}}\|a\|$. Use this result to prove Lemma 4.1 of Chapter 4. (In proving both claims it is OK if you get different constants; know however that the constants in Lemma 4.1 are sub-optimal).

Points: 10 pts.

Solution.
Note that $\mathbb{E}\left[e^{\lambda\epsilon_i}\right]$ can be bounded as

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\epsilon_i\right)\right] &= \frac{1}{2}(e^\lambda + e^{-\lambda}) = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}\\ &\le \sum_{k=0}^{\infty}\frac{(\lambda^2)^k}{k!}\times\frac{k!}{(2k)!} \le \sum_{k=0}^{\infty}\frac{(\lambda^2)^k}{k!}\times\left(\frac{1}{2}\right)^k = e^{\frac{1}{2}\sigma^2},\end{aligned}$$

hence $\epsilon_i \in SG(1)$. Since ϵ_i's are independent, from HW1 Problem 6 Details,

$$a^\top\epsilon \in SG\left(\|a\|_2^2\right) \subset SG\left(D^2\right).$$

Hence from maximum inequality for subgaussian random variables,

$$\mathbb{E}\left[\frac{1}{n}\sup_{a\in\mathcal{A}} a^\top\epsilon\right] \le D\frac{\sqrt{2\log|\mathcal{A}|}}{n},$$

and from HW2 Problem 1, we also have a bound for absolute value version,

$$\mathbb{E}\left[\frac{1}{n}\sup_{a\in\mathcal{A}}\left|a^\top\epsilon\right|\right] \le D\frac{\sqrt{2\log(2|\mathcal{A}|)}}{n}. \tag{13}$$

For proving Lemma 4.1 of Chapter 4, let $\mathcal{F}(x_1^n) := \{(f(x_1), \cdots, f(x_n)) \in \mathbb{R}^n | \, f \in \mathcal{F}\}$ as in (4.11) in Wainwright, then $\max_{a\in\mathcal{F}(x_1^n)}\|a\| = \sqrt{n}D(x_1^n)$. Hence by applying (13),

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(x_i)\right|\right] \le D(x_1^n)\sqrt{\frac{2(\log|\mathcal{F}(x_1^n)| + \log 2)}{n}}.$$

Then $\mathcal{F}$ being polynomial discrimination of order ν implies $|\mathcal{F}(x_1^n)| \le (n+1)^\nu$, hence

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(x_i)\right|\right] &\le D(x_1^n)\sqrt{\frac{2(\nu\log(n+1) + \log 2)}{n}}\\ &\le 2D(x_1^n)\sqrt{\frac{\nu\log(n+1)}{n}}.\end{aligned}$$