COMS 4772 Fall 2016 Homework 3
Due Monday, November 21

Si Kai Lee sl3950

Instructions:

- The required number of points for this assignment is 100. Any points you earn beyond this is extra credit.
- The usual homework policies (`http://www.cs.columbia.edu/~djhsu/coms4772-f16/about.html`) are, of course, in effect.
- Using this LaTeX template will be helpful for grading purposes.

Problem 1 (25 points)**.** Let $\boldsymbol{A}$ be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and corresponding orthonormal eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$—ordered so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Also let $\widetilde{\boldsymbol{A}}$ be a symmetric $n \times n$ matrix with $\hat{\boldsymbol{v}}_1 \in \arg\max_{\boldsymbol{v} \in S^{n-1}} |\boldsymbol{v}^\top \widetilde{\boldsymbol{A}} \boldsymbol{v}|$. Prove that

$$\langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \;\geq\; 1 - \big(2\epsilon/\gamma\big)^2 \;,$$

where $\gamma := \min_{i \neq 1} |\lambda_1 - \lambda_i|$ and $\epsilon := \|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_2$. Try to do this from first principles (i.e., do not invoke Davis-Kahan or Wedin's theorem).

Solution.
We know that $\langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \leq 1$ as $\langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle \leq 1$. We begin by manipulating $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{H}$

$$\begin{aligned}
\widetilde{\boldsymbol{A}} &= \boldsymbol{A} + \boldsymbol{H} \\
\widetilde{\boldsymbol{A}}\boldsymbol{v}_1 &= \boldsymbol{A}\boldsymbol{v}_1 + \boldsymbol{H}\boldsymbol{v}_1 \\
\hat{\boldsymbol{v}}_1 \widetilde{\boldsymbol{A}}\boldsymbol{v}_1 &= \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \hat{\boldsymbol{v}}_1 \boldsymbol{H}\boldsymbol{v}_1 \\
&= \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + (\boldsymbol{v}_1 + \boldsymbol{w}_1)\boldsymbol{H}\boldsymbol{v}_1 \\
&= \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \lambda_{\max}(H)\boldsymbol{w}_1^T \boldsymbol{H}\boldsymbol{v}_1 \\
&= \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{H}\boldsymbol{v}_1 \\
&= \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1
\end{aligned}$$

Then, we bound $\hat{\boldsymbol{v}}_1 \widetilde{\boldsymbol{A}} \boldsymbol{v}_1$ from above and below

$$\begin{aligned}
(\lambda_1 - \min_{i \neq 1} |\lambda_1 - \lambda_i|)\hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1 \leq \hat{\boldsymbol{v}}_1 \widetilde{\boldsymbol{A}}\boldsymbol{v}_1 &\leq \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \\
(\lambda_1 - \gamma)\hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1 &\leq \lambda_1 \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \boldsymbol{v}_1 \boldsymbol{H}\boldsymbol{v}_1 + \epsilon \\
(\lambda_1 + \gamma)\hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1 &\leq (\lambda_1 + \gamma)\hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \epsilon \\
\gamma \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + \epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1 &\leq \gamma \hat{\boldsymbol{v}}_1^T \boldsymbol{v}_1 + 2\epsilon \\
\gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 + 2\epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1 + (\epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1)^2 &\leq \gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 + 8\epsilon \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle + 4\epsilon^2 \\
\gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 + (\epsilon \boldsymbol{w}_1^T \boldsymbol{v}_1)^2 &\leq \gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 + 4\epsilon^2 \\
\gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 - 4\epsilon^2 &\leq \gamma^2 \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \\
\langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 - \frac{4\epsilon^2}{\gamma^2} &\leq \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \\
1 - (\frac{2\epsilon}{\gamma})^2 &\leq \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \text{ where } \langle \hat{\boldsymbol{v}}_1, \boldsymbol{v}_1 \rangle^2 \approx 1
\end{aligned}$$

$\square$

Problem 2 (25 points)**.** Let $\boldsymbol{A}$ be the adjacency matrix in $\{0,1\}^{n\times n}$ for a random undirected graph over n vertices, where the edges appear independently, each with probability at most p. Use Matrix Bernstein (Theorem 1, below) to prove that with probability at least 0.99,

$$\left\|\boldsymbol{A}-\mathbb{E}(\boldsymbol{A})\right\|_2 \;\le\; O\left(\sqrt{pn\log n}+\log n\right)\,.$$

Solution.
We begin by splitting the adjacency matrix $\boldsymbol{A}$ into n $n\times n$ symmetric matrices $\boldsymbol{A}_1,..,\boldsymbol{A}_n$ where $\sum_{i=1}^n \boldsymbol{A}_i=\boldsymbol{A}$. Since the edges appear independently with probability at most p, we can assume that $\boldsymbol{A}$ and its sub-matrices $\boldsymbol{A}_1,...,\boldsymbol{A}_n$ are also independent. Such a split would also yield n $n\times n$ expected matrices $\mathbb{E}(\boldsymbol{A}_1),...,\mathbb{E}(\boldsymbol{A}_n)$ and assuming that the expected matrices are unbiased, we have $\mathbb{E}(\boldsymbol{A}_i-\mathbb{E}(\boldsymbol{A}_i))=0$. Lastly, we know that $\lambda_{\max}(\boldsymbol{A}_i)\le 1$ $(=R)$ as the maximum value of any entry $\boldsymbol{A}_i$ is 1. With the above, we fulfil the conditions needed to satisfy the Matrix Bernstein inequality.

The problem requires the spectral norm of $||\boldsymbol{A}-\mathbb{E}(\boldsymbol{A})||_2=\sqrt{\lambda_{\max}((\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))*(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))}$ to be $\le O\left(\sqrt{pn\log n}+\log n\right)$ with probability 0.99. As $\boldsymbol{A}$ contains no complex numbers, its complex conjugate $\boldsymbol{A}*=\boldsymbol{A}$, $\sqrt{\lambda_{\max}((\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))*(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))}=\lambda_{\max}((\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))$. Manipulating the Matrix Bernstein inequality to fit the desired form, we have $\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^N \boldsymbol{X}_i\right)\le t\right)\ge\; d\cdot\exp\left(-\frac{t}{2(\sigma^2+Rt/3)}\right)$ and $\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^N \boldsymbol{X}_i\right)\le t\right)=0.99$, we are left with $d\cdot\exp\left(-\frac{t^2}{2(\sigma^2+Rt/3)}\right)\le 0.99$.

To find t, we first need to know σ^2. The value of each entry $\boldsymbol{A}_{ij}$ can be taken to be a Bernoulli random variable which has a variance of $p(1-p)=p-p^2=\mathbb{E}(\boldsymbol{A}_{ij}^2)-\mathbb{E}(\boldsymbol{A}_{ij})\mathbb{E}(\boldsymbol{A}_{ij})$. As we know that $\mathbb{E}\boldsymbol{A}_{ij}=p$, $\mathbb{E}(\boldsymbol{A}_{ij}^2)=p$ and $\sigma^2=p(n-1)$.

With that in hand, we work on $0.99\ge n\cdot\exp\left(-\frac{t^2}{2(p(n^2-n)+t/3)}\right)$. In the processing of deriving t, we add and drop multiplicative constants and constants when convenient as the end result would be still be within the bound.

$$\begin{aligned}
0.99 &\ge n\cdot\exp\left(-\frac{t^2}{2p(n-1)+2t}\right)\\
\log 0.99 &\ge \log n-\frac{t^2}{2pn+2t}\\
\frac{t^2}{2pn+2t} &\ge \log n\\
t^2-2t(\log n)-2pn(\log n) &\ge 0\\
(t-(\log n))^2 &\ge pn(\log n)-(\log n)^2\\
t &\ge \sqrt{pn(\log n)-(\log n)^2}+\log n\\
&\ge \sqrt{pn\log n}+\log n\\
&= O\left(\sqrt{pn\log n}+\log n\right)
\end{aligned}$$

□

Problem 3 (10+65=75 points)**.** In a crowdsourcing problem, there are m images that need to be labeled with either $+1$ or -1, and there are n workers available to do the labeling. Each worker provides $\{\pm 1\}$ labels for all images: the label provided by worker j on image i is $X_{i,j}$. The correct $\{\pm 1\}$ label for image i is Z_i.

Assume the following generative process for the correct labels and worker-provided labels. The process is governed by parameters $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \ldots, \gamma_m) \in [-1,+1]^m$ and $\boldsymbol{\delta} := (\delta_1, \delta_2, \ldots, \delta_n) \in [-1,+1]^n$. The data for images $\{(X_{i,1}, X_{i,2}, \ldots, X_{i,n}, Z_i)\}_{i=1}^m$ are independent. For each image i,

- the distribution of the correct label is given by

$$\mathbb{P}(Z_i = +1) \ = \ 1 - \mathbb{P}(Z_i = -1) \ = \ \frac{1+\gamma_i}{2} \,;$$

- the worker-provided labels $X_{i,1}, X_{i,2}, \ldots, X_{i,n}$ are conditionally independent given Z_i;
- worker j provides the correct label with probability $\frac{1+\delta_j}{2}$: for each $z \in \{\pm 1\}$,

$$\mathbb{P}(X_{i,j} = z \mid Z_i = z) \ = \ 1 - \mathbb{P}(X_{i,j} \neq z \mid Z_i = z) \ = \ \frac{1+\delta_j}{2} \,.$$

Suppose the random matrix $\boldsymbol{X}$ (whose (i,j)-th entry is $X_{i,j}$) is observed, and the correct labels $\boldsymbol{Z} := (Z_1, Z_2, \ldots, Z_m)$ are hidden.

(a) Write expressions for the largest singular value of $\mathbb{E}(\boldsymbol{X})$ and also for the corresponding (unit length) left and right singular vectors.

(b) Assume that $\boldsymbol{\gamma} \in \{\pm 1\}^m$ and $\delta_1 \geq 0.1$. Write a procedure for estimating $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ based on the singular value decomposition of $\boldsymbol{X}$. Prove bounds on the Euclidean norm errors of your estimates that hold with probability at least 0.99.

Hint: you may find (some of) the theorems given below to be useful.

Solution.

(a)

$$\begin{aligned}
\mathbb{E}X_{ij} &= \mathbb{P}(X_{ij} = 1|Z_i = 1)\mathbb{P}(Z_i = 1) + \mathbb{P}(X_{ij} = 1|Z_i = -1)\mathbb{P}(Z_i = -1) \\
&\quad - \mathbb{P}(X_{ij} = -1|Z_i = 1)\mathbb{P}(Z_i = 1) - \mathbb{P}(X_{ij} = -1|Z_i = -1)\mathbb{P}(Z_i = -1) \\
&= (\frac{1+\delta_j}{2})(\frac{1+\gamma_i}{2}) + (\frac{1-\delta_j}{2})(\frac{1-\gamma_i}{2}) - (\frac{1-\delta_j}{2})(\frac{1+\gamma_i}{2}) - (\frac{1+\delta_j}{2})(\frac{1-\gamma_i}{2}) \\
&= (\frac{1+\delta_j}{2})\gamma_i - (\frac{1-\delta_j}{2})\gamma_i \\
&= \gamma_i \delta_j
\end{aligned}$$

Hence $\mathbb{E}\boldsymbol{X}$ is a matrix containing $\gamma_i\delta_j$s. Taking SVD of $\mathbb{E}X$, we obtain a $m \times m$ matrix U containing a single column of $\frac{\gamma_1}{\sqrt{\sum_{i=1}^m \gamma_i^2}}, \ldots, \frac{\gamma_m}{\sqrt{\sum_{i=1}^m \gamma_i^2}}$ with the rest 0s, a $m \times n$ matrix containing only $\sqrt{\sum_{i=1}^m \gamma_i^2}\sqrt{\sum_{i=1}^n \delta_i^2}$ in the diagonals and a $n \times n$ matrix V containing a single column of $\frac{\delta_1}{\sqrt{\sum_{i=1}^n \delta_i^2}}, \ldots, \frac{\delta_n}{\sqrt{\sum_{i=1}^n \delta_i^2}}$ with the rest 0s. As the largest values within $\mathbb{E}X$ is at most 1, $\sigma_1 \leq 1$. $U_1 = \gamma_1, \ldots, \gamma_m$ and $V_1 = \delta_1, \ldots, \delta_n$.

(b) The algorithm is:

(a) Sample the random matrix $\boldsymbol{X}$ N times

(b) $\mathbb{E}\hat{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i$ which approaches $\mathbb{E}\boldsymbol{X}$ as $N \to \infty$

(c) Run SVD on $\mathbb{E}\hat{\boldsymbol{X}}$ to obtain $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$

(d) For values in the first column of $\boldsymbol{U}$ that are not 1 or -1, round them up to 1 if greater than 0 and vice versa. Set $\boldsymbol{\gamma}$ as the rounded values of the first column of $\boldsymbol{U}$

(e) Divide a diagonal entry in Σ by $\sqrt{m}$ and set δ as the result multiplied with the first column of $\boldsymbol{V}$

We use the Matrix-Bernstein inequality to bound the errors of $\boldsymbol{U}$ and $\boldsymbol{V}$ by taking the SVD of each random $\boldsymbol{X}$ we draw and set the $\boldsymbol{X}_i$ in the inequality to be $\boldsymbol{U} - \boldsymbol{U}_i$ and $\boldsymbol{V} - \boldsymbol{V}_i$ respectively. For both cases, we have $\lambda_{\max}(\boldsymbol{X}_i) \leq 1$. $\sigma^2_{\boldsymbol{U}-\boldsymbol{U}_i} \leq Nm$ and $\sigma^2_{\boldsymbol{V}-\boldsymbol{V}_i} \leq Nn$ as it is reasonable to assume all entries within $\mathbb{E}[\boldsymbol{U} - \boldsymbol{U}_i]$ and $\mathbb{E}[\boldsymbol{V} - \boldsymbol{V}_i]$ are smaller than 1. Hence the error bound for $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are $O\left(\frac{1}{m}(\sqrt{Nm\log m} + \log m)\right) + O\left(\frac{1}{n}(\sqrt{Nn\log n} + \log n)\right)$ as we are only estimating the error of a column for each matrix. The derivations are not shown as they are almost identical to question 2.

$\square$

Some theorems

Theorem 1 (Matrix Bernstein). *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ be independent, random symmetric matrices in $\mathbb{R}^{d \times d}$. Assume each $\boldsymbol{X}_i$ satisfies $\mathbb{E}(\boldsymbol{X}_i) = \boldsymbol{0}$ and $\lambda_{\max}(\boldsymbol{X}_i) \le R$ almost surely. For all $t \ge 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^{N} \boldsymbol{X}_i\right) \ge t\right) \;\le\; d \cdot \exp\left(-\frac{t^2}{2(\sigma^2 + Rt/3)}\right) \quad \textit{where} \quad \sigma^2 \;:=\; \left\|\sum_{i=1}^{N} \mathbb{E}\boldsymbol{X}_i^2\right\|_2 .$$

Theorem 2 (Weyl). *For any symmetric $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$\lambda_i(\boldsymbol{A}) + \lambda_n(\boldsymbol{H}) \;\le\; \lambda_i(\boldsymbol{A} + \boldsymbol{H}) \;\le\; \lambda_i(\boldsymbol{A}) + \lambda_1(\boldsymbol{H}), \quad 1 \le i \le n,$$

where $\lambda_i(\cdot)$ denotes the i-th largest eigenvalue of its argument.

Theorem 3 (Weyl (again)). *For any $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{E}$,*

$$\left|\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{A} + \boldsymbol{E})\right| \;\le\; \|\boldsymbol{E}\|_2 , \quad 1 \le i \le \min\{m, n\},$$

where $\sigma_i(\cdot)$ denotes the i-th largest singular value of its argument.

Theorem 4 (Davis-Kahan). *Let $\boldsymbol{A} = \boldsymbol{E}_0 \boldsymbol{A}_0 \boldsymbol{E}_0{}^\top + \boldsymbol{E}_1 \boldsymbol{A}_1 \boldsymbol{E}_1{}^\top$ and $\boldsymbol{A} + \boldsymbol{H} = \boldsymbol{F}_0 \boldsymbol{\Lambda}_0 \boldsymbol{F}_0{}^\top + \boldsymbol{F}_1 \boldsymbol{\Lambda}_1 \boldsymbol{F}_1{}^\top$ be symmetric matrices with $[\boldsymbol{E}_0, \boldsymbol{E}_1]$ and $[\boldsymbol{F}_0, \boldsymbol{F}_1]$ orthogonal. If the eigenvalues of $\boldsymbol{A}_0$ are contained in an interval (a, b), and the eigenvalues of $\boldsymbol{\Lambda}_1$ are excluded from the interval $(a - \delta, b + \delta)$ for some $\delta > 0$, then*

$$\left\|\boldsymbol{F}_1{}^\top \boldsymbol{E}_0\right\|_2 \;\le\; \frac{\left\|\boldsymbol{F}_1{}^\top \boldsymbol{H} \boldsymbol{E}_0\right\|_2}{\delta} .$$

Theorem 5 (Wedin). *Suppose matrices $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{m \times n}$ may be written as*

$$\boldsymbol{A} \;=\; \boldsymbol{U}_1 \boldsymbol{S}_1 \boldsymbol{V}_1^\top + \boldsymbol{U}_2 \boldsymbol{S}_2 \boldsymbol{V}_2^\top , \qquad \widetilde{\boldsymbol{A}} \;=\; \widetilde{\boldsymbol{U}}_1 \widetilde{\boldsymbol{S}}_1 \widetilde{\boldsymbol{V}}_1^\top + \widetilde{\boldsymbol{U}}_2 \widetilde{\boldsymbol{S}}_2 \widetilde{\boldsymbol{V}}_2^\top ,$$

where $\boldsymbol{U}_1^\top \boldsymbol{U}_1 = \boldsymbol{V}_1^\top \boldsymbol{V}_1 = \boldsymbol{I}$, $\boldsymbol{U}_2^\top \boldsymbol{U}_2 = \boldsymbol{V}_2^\top \boldsymbol{V}_2 = \boldsymbol{I}$, $\widetilde{\boldsymbol{U}}_1^\top \widetilde{\boldsymbol{U}}_1 = \widetilde{\boldsymbol{V}}_1^\top \widetilde{\boldsymbol{V}}_1 = \boldsymbol{I}$, and $\widetilde{\boldsymbol{U}}_2^\top \widetilde{\boldsymbol{U}}_2 = \widetilde{\boldsymbol{V}}_2^\top \widetilde{\boldsymbol{V}}_2 = \boldsymbol{I}$; and $\boldsymbol{S}_1, \boldsymbol{S}_2, \widetilde{\boldsymbol{S}}_1, \widetilde{\boldsymbol{S}}_2$ are diagonal and non-negative. If there exists $\alpha > 0$ and $\delta > 0$ such that the smallest singular value in $\boldsymbol{S}_1$ is at least $\alpha + \delta$, and the largest singular value in $\widetilde{\boldsymbol{S}}_2$ is at most α, then

$$\max\left\{\left\|\boldsymbol{U}_2^\top \tilde{\boldsymbol{U}}_1\right\|_2 , \left\|\boldsymbol{V}_2^\top \tilde{\boldsymbol{V}}_1\right\|_2\right\} \;\le\; \frac{\max\left\{\left\|(\widetilde{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{V}_1\right\|_2 , \left\|\boldsymbol{U}_1^\top (\widetilde{\boldsymbol{A}} - \boldsymbol{A})\right\|_2\right\}}{\delta} .$$