STAT 37710: Machine Learning	Autumn 2022

Homework 0

Due date: XXX

You are allowed to drop 1 subproblem without penalty. But you cannot drop problems on simulation.

1 Probability and statistics

1. Probability and Statistics (*20 points*)

a.(*10 points*) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

b.(*10 points*) Let $X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d random variables. Compute the following:

1. a, b such that $aX_1 + b \sim \mathcal{N}(0, 1)$.
2. $\mathbb{E}X_1 + 2X_2, \operatorname{Var} X_1 + 2X_2$.
3. Setting $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the mean and variance of $\sqrt{n}(\widehat{\mu}_n - \mu)$.

o **2. F** (*0 points*) r any function $g\colon \mathbb{R} \to \mathbb{R}$ and random variable X with PDF $f(x)$, recall that the expected value of $g(X)$ is defined as $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(y)f(y)y$. For a boolean event A, define $\mathbb{I}\{A\}$ as 1 if A is true, and 0 otherwise. Fix $F(x) = \mathbb{E}\mathbb{I}\{X \leq x\}$. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with CDF $F(x)$. Define $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \leq x\}$. Note, for every x, that $\widehat{F}_n(x)$ is an empirical estimate of $F(x)$.

1. For any x, what is $\mathbb{E}\widehat{F}_n(x)$?
2. For any x, the variance of $\widehat{F}_n(x)$ is $\mathbb{E}\Big(\widehat{F}_n(x) - F(x)\Big)^2$. Show that $\operatorname{Var} \widehat{F}_n(x) = \frac{F(x)(1-F(x))}{n}$.
3. Using your answer to b, show that for all $x \in \mathbb{R}$, we have $\mathbb{E}\Big(\widehat{F}_n(x) - F(x)\Big)^2 \leq \frac{1}{4n}$.

3. Linear algebra (*20 points*) Consider two orthonormal matrices $\boldsymbol{U}, \boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$, satisfying $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}^{\star\top}\boldsymbol{U}^\star = \boldsymbol{I}_r$ with $r < n$. We have discussed extensively the distance using projection matrices

$$\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|, \quad \text{and} \quad \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|_{\mathrm{F}}.$$

Also, our default choice of distance is the one using optimal rotation matrix:

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \big\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\big\|, \quad \text{and} \quad \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}}\,.$$

Here $\mathbb{O}^{r\times r} := \{\boldsymbol{R} \in \mathbb{R}^{r\times r} \mid \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}_r\}$ is the set of all $r \times r$ orthonormal matrices.

a.(*10 points*) Show that

$$\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\| \le \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\right\| \le \sqrt{2}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|.$$

b.(*10 points*) Show that

$$\tfrac{1}{\sqrt{2}}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|_{\mathrm{F}} \le \min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}} \le \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|_{\mathrm{F}}.$$

4. Variant of Wedin's theorem (*10 points*) Consider the setting and notation used in class. Wedin's $\sin\boldsymbol{\Theta}$ theorem tells us that if $\|\boldsymbol{E}\| < \sigma_r^\star - \sigma_{r+1}^\star$, then there exist two orthonormal matrices $\boldsymbol{R}_{\boldsymbol{U}}, \boldsymbol{R}_{\boldsymbol{V}} \in \mathbb{R}^{r\times r}$ such that

$$\max\left\{\|\boldsymbol{U}\boldsymbol{R}_{\boldsymbol{U}} - \boldsymbol{U}^\star\|_{\mathrm{F}}, \|\boldsymbol{V}\boldsymbol{R}_{\boldsymbol{V}} - \boldsymbol{V}^\star\|_{\mathrm{F}}\right\} \le \frac{\sqrt{2}\max\left\{\|\boldsymbol{E}^\top\boldsymbol{U}^\star\|_{\mathrm{F}}, \|\boldsymbol{E}\boldsymbol{V}^\star\|_{\mathrm{F}}\right\}}{\sigma_r^\star - \sigma_{r+1}^\star - \|\boldsymbol{E}\|}.$$

However, in some cases, we hope for a single rotation matrix that could align both $(\boldsymbol{U}, \boldsymbol{U}^\star)$ and $(\boldsymbol{V}, \boldsymbol{V}^\star)$. It turns out that this is achievable. Show that if $\|\boldsymbol{E}\| < \sigma_r^\star - \sigma_{r+1}^\star$, there exists a single orthonormal matrix $\boldsymbol{R} \in \mathcal{O}^{r\times r}$ such that

$$\left(\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star\|_{\mathrm{F}}^2 + \|\boldsymbol{V}\boldsymbol{R} - \boldsymbol{V}^\star\|_{\mathrm{F}}^2\right)^{1/2} \le \frac{\sqrt{2}\left(\|\boldsymbol{E}^\top\boldsymbol{U}^\star\|_{\mathrm{F}}^2 + \|\boldsymbol{E}\boldsymbol{V}^\star\|_{\mathrm{F}}^2\right)^{1/2}}{\sigma_r^\star - \sigma_{r+1}^\star - \|\boldsymbol{E}\|}.$$

You are allowed to invoke the general Davis-Kahan $\sin\boldsymbol{\Theta}$ theorem given in class.

5. Quadratic systems of equations (*10 points*) Suppose that our goal is to estimate an unknown vector $\boldsymbol{x}^\star \in \mathbb{R}^n$ (obeying $\|\boldsymbol{x}^\star\|_2 = 1$) based on m i.i.d. samples of the form

$$y_i = (\boldsymbol{a}_i^\top\boldsymbol{x}^\star)^2, \qquad i = 1, \dots, m,$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$ are independent vectors (known *a priori*) obeying $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

Suggest a spectral method for estimating $\boldsymbol{x}^\star$ that is consistent with either $\boldsymbol{x}^\star$ or $-\boldsymbol{x}^\star$ in the limit of infinite data, i.e., as m goes to infinity.

6. Matrix completion (*20 points*) Suppose that the ground-truth matrix is given by

$$\boldsymbol{M}^\star = \boldsymbol{u}^\star\boldsymbol{v}^{\star\top} \in \mathbb{R}^{n\times n},$$

where $\boldsymbol{u}^\star = \tilde{\boldsymbol{u}}/\|\tilde{\boldsymbol{u}}\|_2$ and $\boldsymbol{v}^\star = \tilde{\boldsymbol{v}}/\|\tilde{\boldsymbol{v}}\|_2$, with $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ generated independently. Each entry of $\boldsymbol{M}^\star = [M_{i,j}^\star]_{1\le i,j\le n}$ is observed independently with probability p. In the lecture, we have constructed a matrix $\boldsymbol{M} = [M_{i,j}]_{1\le i,j\le n}$, where

$$M_{i,j} = \begin{cases} \frac{1}{p}M_{i,j}^\star, & \text{if } M_{i,j}^\star \text{ is observed;} \\ 0, & \text{else.} \end{cases}$$

We have shown in class that with high probability, the leading left singular vector $\boldsymbol{u}$ of $\boldsymbol{M}$ is a reliable estimate of $\boldsymbol{u}^\star$, provided that $p \gg \frac{\log^3 n}{n}$.

Now, consider a new matrix $\boldsymbol{M}^{(1)} = [M^{(1)}_{i,j}]_{1\le i,j\le n}$ obtained by zeroing out the 1st column and 1st row of $\boldsymbol{M}$. More precisely, for any $1 \le i, j \le n$,

$$M^{(1)}_{i,j} = \begin{cases} M_{i,j}, & \text{if } i \neq 1 \text{ and } j \neq 1; \\ 0, & \text{else.} \end{cases}$$

Let $\boldsymbol{u}^{(1)}$ (resp. $\boldsymbol{v}^{(1)}$) be the leading left (resp. right) singular vector of $\boldsymbol{M}^{(1)}$.

a.(*10 points*) Recall that Wedin's $\sin\boldsymbol{\Theta}$ Theorem states that: for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, their leading left singular vectors (denoted by $\boldsymbol{u}_A$ and $\boldsymbol{u}_B$ respectively) satisfy

$$\mathsf{dist}(\boldsymbol{u}_A, \boldsymbol{u}_B) \le \frac{\|\boldsymbol{A}-\boldsymbol{B}\|}{\sigma_1(\boldsymbol{A}) - \sigma_2(\boldsymbol{A}) - \|\boldsymbol{A}-\boldsymbol{B}\|}.$$

Use it to derive an upper bound on $\mathsf{dist}(\boldsymbol{u}^{(1)}, \boldsymbol{u})$ in terms of n and p.

b.(*10 points*) Recall that a more refined version of Wedin's $\sin\boldsymbol{\Theta}$ Theorem states that: for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, their leading left singular vectors (denoted by $\boldsymbol{u}_A$ and $\boldsymbol{u}_B$ respectively) satisfy

$$\mathsf{dist}(\boldsymbol{u}_A, \boldsymbol{u}_B) \le \frac{\max\left\{\|(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{v}_A\|, \|(\boldsymbol{A}-\boldsymbol{B})^\top \boldsymbol{u}_A\|\right\}}{\sigma_1(\boldsymbol{A}) - \sigma_2(\boldsymbol{A}) - \|\boldsymbol{A}-\boldsymbol{B}\|}$$

where $\boldsymbol{v}_A$ is the leading right singular vector of $\boldsymbol{A}$. Can you use this refined version to derive a sharper upper bound on $\mathsf{dist}(\boldsymbol{u}^{(1)}, \boldsymbol{u})$? Here, you can assume without proof that $\|\boldsymbol{u}\|_\infty, \|\boldsymbol{u}^{(1)}\|_\infty, \|\boldsymbol{v}\|_\infty, \|\boldsymbol{v}^{(1)}\|_\infty \lesssim \sqrt{\frac{\log n}{n}}$ with high probability.

7. Community detection experiments (*20 points*) Consider the SBM model discussed in class. Fix the number n of nodes in a graph to be 100. Set $p = \frac{1+\varepsilon}{2}$ and $q = \frac{1-\varepsilon}{2}$ for some quantity $\varepsilon \in [0, 1/2]$. Generate a random graph and then use the spectral method to cluster the nodes. Please plot the mis-clustering rate vs. the probability gap ε. At the minimum, you should take 50 different values of ε (with linear spacing) in $[0, 1/2]$. For each value of ε, you need to run the experiment with at least 200 Monte-Carlo trials to calculate the average mis-clustering rate across trials.