STAT 37797: Mathematics of Data Science	Autumn 2021

Homework 2

Due date: 11:59pm on Nov. 19th

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

1. Concentration of Gaussian random variables (*20 points*)

a.(*5 points*) Let X be a standard normal random variable. Prove that

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/2).$$

b.(*5 points*) Let $X_1, X_2, \ldots, X_n$ be n i.i.d. standard normal random variables. Prove that with probability at least $1 - O(n^{-10})$, one has

$$\max_{1 \leq i \leq n} |X_i| \leq 5\sqrt{\log n}.$$

c.(*10 points*) Let $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector where each coordinate is an independent standard normal random variable. Using the the conclusion above, one can show that $\|\boldsymbol{x}\|_2 \lesssim \sqrt{n \log n}$ with high probability. However, this falls short in two aspects. First, the upper bound on $\|x\|_2$ is not tight. Second, it doesn't provide a high-probability lower bound of $\|\boldsymbol{x}\|_2$. In this part, prove that for all $t \in (0, 1)$, one has

$$\mathbb{P}(|\|\boldsymbol{x}\|_2^2 - n| \geq nt) \leq 2\exp(-nt^2/8).$$

(Hint: Laplace transform method)

2. Norm of Gaussian random matrices (*40 points*)

Recall that in class, we have used the bound $\|\boldsymbol{E}\| \lesssim \sqrt{n}$ where $\boldsymbol{E} \in \mathbb{R}^{n \times n}$ is composed of i.i.d. standard normal random variables.

a.(*10 points*) Use matrix Bernstein's inequality to show that with high probability $\|\boldsymbol{E}\| \lesssim \sqrt{n \log n}$. (Hint: truncation)

As before, the bound proved in part (a) is off by a $\sqrt{\log n}$ factor. In the following, we will prove a tighter bound. Recall the definition of $\|\boldsymbol{E}\|$:

$$\|\boldsymbol{E}\| = \sup_{\|\boldsymbol{v}\|_2 = 1} \|\boldsymbol{E}\boldsymbol{v}\|_2$$

Hence it suffices to show that with high probability $\sup_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{E}\boldsymbol{v}\|_2 \lesssim \sqrt{n}$.

b.(*5 points*) Let's first focus on a fixed vector $\|\boldsymbol{v}\|_2 = 1$. Prove that for any fixed $\boldsymbol{v} \in \mathbb{R}^n$, one has

$$\mathbb{P}(\|\boldsymbol{E}\boldsymbol{v}\|_2 \geq 10\sqrt{n}) \leq 2\exp(-100n).$$

It is tempting to apply "union bound" and "obtain"

$$\mathbb{P}(\sup_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{E}\boldsymbol{v}\|_2 \geq 10\sqrt{n}) \leq \sum_{\|\boldsymbol{v}\|_2=1} \mathbb{P}(\|\boldsymbol{E}\boldsymbol{v}\|_2 \geq 10\sqrt{n}).$$

However this argument is ABSOLUTELY wrong as one cannot apply union bound to a uncountable set. Therefore, to properly apply union bound, one needs to restrict attention to a finite subset of the unit sphere in $\mathbb{R}^n$ that well approximates the unit sphere. This motivates the construction of the ε-net.

c.(*10 points*) Let $\mathcal{N}_\varepsilon$ be a subset of $\{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_2 = 1\}$ such that for any point $\boldsymbol{v}$ in $\{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_2 = 1\}$, one can find an element $\boldsymbol{u} \in \mathcal{N}_\varepsilon$ such that $\|\boldsymbol{u} - \boldsymbol{v}\|_2 \leq \varepsilon$. In particular, set $\mathcal{N}_\varepsilon$ be such a set with smallest cardinality. Prove that

$$|\mathcal{N}_\varepsilon| \leq (1 + \tfrac{2}{\varepsilon})^n,$$

where $|\mathcal{N}_\varepsilon|$ denotes the cardinality of $\mathcal{N}_\varepsilon$.

d.(*5 points*) Fix some $\varepsilon \in (0, 1)$. Prove that for any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$

$$\|\boldsymbol{A}\| \leq \frac{1}{1-\varepsilon} \cdot \max_{\boldsymbol{v} \in \mathcal{N}_\varepsilon} \|\boldsymbol{A}\boldsymbol{v}\|_2.$$

This shows the usefulness of $\mathcal{N}_\varepsilon$ in terms of approximating $\|\boldsymbol{A}\|$.

e.(*10 points*) Combine the previous steps to show that with high probability $\|\boldsymbol{E}\| \lesssim \sqrt{n}$.

3. Matrix concentration in matrix completion (*10 points*) Consider the matrix completion problem introduced in class where $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \in \mathbb{R}^{n \times n}$ is a rank-r matrix. Let μ be its incoherence parameter. Prove that with high probability

$$\|\boldsymbol{M} - \boldsymbol{M}^\star\| \lesssim \sqrt{\frac{\mu r \log n}{np}} \|\boldsymbol{M}^\star\|$$

as long as $np \geq C\mu r \log n$ for some sufficiently large constant $C > 0$.

4. Matrix completion experiments (*20 points*) Consider the matrix completion problem introduced in class. Let $\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ be the underlying groundtruth matrix. Here $\boldsymbol{U}^\star, \boldsymbol{V}^\star \in \mathbb{R}^{n \times r}$ are two independent random orthonormal matrices. For simplicity consider $\boldsymbol{\Sigma}^\star = \boldsymbol{I}_r$. Let p be the observation probability for each entry. Let $\hat{\boldsymbol{M}}$ be the spectral estimate of the matrix $\boldsymbol{M}^\star$.

Fix $n = 200, r = 5$, and vary p from 0.2 to 1. Please report the relative Euclidean error $\frac{\|\hat{\boldsymbol{M}} - \boldsymbol{M}^\star\|_{\mathrm{F}}}{\|\boldsymbol{M}^\star\|_{\mathrm{F}}}$ and the relative entrywise error $\frac{\|\hat{\boldsymbol{M}} - \boldsymbol{M}^\star\|_\infty}{\|\boldsymbol{M}^\star\|_\infty}$ vs. the sampling probability p. Please choose at least 20 different p's and for each p, use at least 50 Monte-Carlo simulations.

5. Community detection experiments (*30 points*)

Consider the SBM model discussed in class, where $p = \alpha \log n/n$ and $q = \beta \log n/n$. Throughout this exercise, we will use the second eigenvector of the adjacency matrix $\boldsymbol{A}$, which does not rely on the knowledge of either p or q.

a.(*15 points*) In the first part, we are going to investigate the phase transition behavior we discussed in class. Fix $n = 300$. Vary β from 0 to 10, and α from 0 to 30, with increments 0.1 and 0.3 respectively. For each α, β, run spectral method for 100 random trials and report the success rate plot. On the same plot, please also add the curves that correspond to $(\sqrt{\alpha} - \sqrt{\beta})^2 = 2$. You should be able to see a sharp transition in terms of success rate around these curves.

b.(*5 points*) In the second part, we will take a closer look at the entrywise behavior of $\hat{\boldsymbol{u}}_2$—the second eigenvector of the adjacency matrix $\boldsymbol{A}$. Fix $n = 5000, \alpha = 4.5, b = 0.25$. Check that based on our theory, spectral method should succeed in exact recovery with high probability for this configuration. To verify this, plot the histogram of the entries in $\sqrt{n}\hat{\boldsymbol{u}}_2$. Are those uniformly close to ± 1?

c.(*10 points*) In class, to prove exact recovery, we actually compare $\hat{\boldsymbol{u}}_2$ with the linearization $\boldsymbol{A}\boldsymbol{u}_2^\star/\lambda_2^\star$, instead of $\boldsymbol{u}_2^\star$. Here we investigate the reason underlying this. Use the same configuration as above, and run 100 Monte-Carlo simulations. Report the boxplots for $\sqrt{n}\|\hat{\boldsymbol{u}}_2 - \boldsymbol{u}_2^\star\|_\infty$, $\sqrt{n}\|\hat{\boldsymbol{u}}_2 - \boldsymbol{A}\boldsymbol{u}_2^\star/\lambda_2^\star\|_\infty$, and $\|\boldsymbol{A}\boldsymbol{u}_2^\star/\lambda_2^\star - \boldsymbol{u}_2^\star\|_\infty$. Which one is the smallest among the three?