STAT 37797: Mathematics of Data Science	Autumn 2021
Homework 1	
Due date: 11:59pm on Oct. 21st	

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

1. Concentration of gaussian random variable (*20 points*)

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

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

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

$$\max_i |X_i| \leq 5\sqrt{\log n}.$$

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 $\|x\|_2$. 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 with control of moment generating function)

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 the tighter version. 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 TOTALLY 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 can approximate the unit sphere well.

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 $\|\boldsymbol{u} - \boldsymbol{v}\|_2 \le \varepsilon$. In particular, set $\mathcal{N}_\varepsilon$ be such a set with smallest cardinality. Prove that

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

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 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$

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

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

4. 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.

5. Matrix completion 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.