36-710, Fall 2018
Homework 3

Due Friday Oct 19 by 5:00pm in JaeHyeok's mailbox

1. Let m be the median of P, a probability distribution in $\mathbb{R}$, assumed unique for convenience. Assume also that there exists an $\eta > 0$ such that the c.d.f. F of P is differentiable at all $x \in I = (m-\eta, m+\eta)$, with $\inf_{x \in I} F'(x) \geq C > 0$. Compute a high probability bound on the length of the confidence interval found in problem 6(a) of HW2. You may use the following result, known as the DKW inequality. If $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution over the real line with c.d.f. F and F_n denotes the corresponding empirical c.d.f. (i.e. for all $x \in \mathbb{R}$, $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leq x\}}$), then

$$\mathbb{P}\left(\|F - F_n\|_\infty \geq t\right) \leq 2e^{-2nt^2}.$$

What happens when η or C gets smaller?

2. 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 $\mathbb{S}^{m-1}$.

3. Exercise 6.10.

4. Exercise 5.1.

5. Exercise 5.2.

6. Let A be a $n \times n$ symmetric matrix with zero diagonal and off-diagonal entries consisting of $\binom{n}{2}$ independent Bernoulli's. Specifically, for any $i < j$,

$$A_{i,j} = A_{j,i} \sim \text{Bernoulli}(p_{i,j}),$$

where each $p_{i,j} \in [0,1]$. Then A is the adjacency matrix of an *inhomogeneous Bernoulli network*, a random simple graph whose edges are independent Bernoulli's. In particular, if $p_{i,j} = p$ for all $i < j$, A is the adjacency matrix of an Erdö-Renyi random graph.

In many problems – for example when analyzing the performance of spectral clustering algorithms for community detection – we need a high probability bound for the quantity

$$\|A - \mathbb{E}[A]\|_{\text{op}}$$

Let $\alpha = \max_{i<j} p_{i,j}$ and assume that $\alpha = \alpha_n$ is allowed to vanish with n in such a way that $\alpha_n = C_1 \frac{\log n}{n}$, for some C_1. Notice that $\alpha_n n$ is a bound on the maximal degree of the graph.

Show that there exists a constant C' such that, with probability at least $\frac{1}{n}$,

$$\|A - \mathbb{E}[A]\|_{\text{op}} \leq \sqrt{C' n \alpha_n \log n}.$$

Thus, as long as α_n is of larger order than $\frac{\log n}{n}$ (so that the graph may be *sparse,* in the sense that the maximal degree is of smaller order than n), $\|A - \mathbb{E}[P]\|_{\text{op}}$ converges in probability to zero. *Write* $A - \mathbb{E}[P] = \sum_{i<j}(A_{i,j} - p_{i,j})(E^{i,j)} + E^{(j,i)})$, *where* $E^{(i,j)}$ *is the* $n \times n$ *matrix whose entries are all zeros, except for the* (i,j)*th entry, which is* 1*. Use Bernsteion matrix inequality.*

For the current state-of-the art on bounds for this type of problems see, *Afonso S. Bandeira and Ramon van Handel, (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries, Ann. Probab. Volume 44, Number 4, 2479-2506.*

7. Consider the linear regression model

$$Y = X\theta^* + \epsilon$$

where $\theta \in \mathbb{R}^d$, X is fixed and $\epsilon \in \mathbb{R}^n$ consists of independent zero-mean variables with finite variance. The ridge estimator is defined as

$$\hat{\theta}_{\text{ridge}} = \hat{\theta}_{\text{ridge}}(\lambda) = \text{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\theta\|^2 \right\},$$

where $\lambda > 0$.

(a) Show that $\hat{\theta}_{\text{ridge}}$ is uniquely defined for any $\lambda > 0$ and find a closed-form expression. Will the solution exist and be unique if $d > n$?

(b) Compute the bias of $\hat{\theta}_{\text{ridge}}$.

8. **Hard thresholding in the sub-gaussian many means problem.** Suppose we observe the vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, where

$$X = \theta^* + \epsilon,$$

with $\theta^* \in \mathbb{R}^d$ unknown and $\epsilon \in SG_d(\sigma^2)$. We would like to estimate θ^* using the hard thresholding estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ with parameter $\tau > 0$, given by:

$$\hat{\theta}_i = \begin{cases} X_i & \text{if } |X_i| > \tau \\ 0 & \text{if } |X_i| \leq \tau. \end{cases}$$

This estimator either keeps or kills each coordinate of X.

For $\delta \in (0,1)$, set

$$\tau = 2\sigma\sqrt{2\log(2d/\delta)}.$$

Notice that $\mathbb{P}(\max_i |\epsilon_i| > \tau/2) \leq \delta$ (If this surprises you, refresh your memory on maximal inequalities).

(a) Prove that the hard-thresholding estimator is the solution the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|X - \theta\|^2 + \tau^2\|\theta\|_0.$$

(b) Prove that if $\|\theta^*\|_0 = k$, with probability at least $1-\delta$,

$$\|\hat{\theta} - \theta^*\|^2 \leq C\sigma^2 k \log(2d/\delta),$$

for some universal constant $C > 0$. *Hint: show that, for each* $i = 1, \ldots, d$

$$|\hat{\theta}_i - \theta_i^*| \leq C' \min\{|\theta_i^*|, \tau\}$$

for some $C' > 0$*, with probability at least* $1-\delta$*.*

(c) Compare with the oracle estimator $\hat{\theta}^{\mathrm{or}}$, with coordinates given by

$$\hat{\theta}_i^{\mathrm{or}} = \begin{cases} X_i & \text{if } i \in \operatorname{supp}(\theta^*) \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, d$. This estimator is of course not computable, as it requires knwoledge of $\operatorname{supp}(\theta^*)$. It is an estimator that an oracle, who has access to this additional knowledge, would be able to compute. Oracle estimators are idealized estimators, which perform at least as well as any computable estimators. Thus, in rder to show that a given estimator performs well, it is enoygh to show that it mimicks closely the performance of an oracle estimator.

(d) Show that if $\min_{i \in \operatorname{supp}(\theta^*)} |\theta_i| > \frac{3}{2}\tau$, then, with probability at least $1-\delta$,

$$\operatorname{supp}(\hat{\theta}) = \operatorname{supp}(\theta^*).$$

How does $\hat{\theta}$ compare now to the oracle estimator?