Final

SDS384

Spring 2021

This exam has 4 short and 4 long questions. You will have to answer **all short questions**, **three long questions**. The assigned points are noted next to each question; the total number of points is 50. Please upload your answers in latex by 11:59 pm Sunday May 16th. Use the latex file format provided.

Read each question carefully, **show your work** and **clearly present your answers**. Note, the exam is printed two-sided - please don't forget the problems on the even pages!

Good Luck!

Name: ____

UTeid: ____

1 Short questions (17 points)

Please answer all of the short questions.

1. (5 pts) Suppose $X_1, \ldots, X_n$ are i.i.d random variables with mean μ and variance σ^2. Let $T_n = \sum_{j=1}^n z_{nj} X_j$ where z_{nj} are given numbers. Let $\mu_n = E[T_n]$ and $\sigma_n^2 = \mathrm{var}(T_n)$. Show that

$$\frac{T_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1),$$

provided $\max_{j \le n} \frac{z_{nj}^2}{\sum_{j=1}^n z_{nj}^2} \to 0$ as $n \to \infty$.

2. (5 pts) Let $X_1, \ldots, X_n$ be independent and suppose that $X_n = \sqrt{n}$ with probability 1/2 and $-\sqrt{n}$ with probability 1/2, for $n = 1, 2, \ldots$. Find the asymptotic distribution of $\bar{X}_n$.

3. (4 pts) Consider a function class with functions of the following form:

$$f_\alpha(x) = \begin{cases} 1 & \text{If } \sin(\alpha x) > 0 \\ 0 & \text{o.w.} \end{cases} \tag{1}$$

Consider a set of datapoints $\{10^{-i}, i = 1, \ldots, n\}$. Show that any set of labeling $y_i, i = 1, \ldots, n$ can be achieved by using

$$\alpha = \pi \left(1 + \sum_{i=1}^n (1 - y_i) 10^i \right).$$

Using this, what do you think the VC dimension of this function class is?

4. (3 pts) Consider a r.v. X such that for all $\lambda \in \Re$

$$E[e^{\lambda X}] \le e^{\frac{\lambda^2 \sigma^2}{2} + \lambda \mu}$$

Prove that $E[X] = \mu$.

2 Long questions (33 points)

Please answer any three of the long questions.

1. (11 pts) Look at the seminar paper "Probability Inequalities for Sums of Bounded Random Variables" by Wassily Hoeffding. It should be available via `lib.utexas.edu`. You can assume that n is a multiple of m (the degree of the kernel). Assume that the kernel is bounded, i.e. $|h(X_1, \ldots, X_m) - \theta| \leq b$, where $\theta = E[h(X_1, \ldots, X_m)]$.

   (a) (4 pts) Read and reproduce the proof of equation 5.7 for large sample deviation of order m U statistics.

   (b) (7 pts) Also prove Bernstein's inequality (see below) for U statistics. This is buried in the paper, you will have to find the bits and pieces and put them together. The Bernstein inequality is given by:

   $$P(|U_n - \theta| \geq \epsilon) \leq a \exp\left(-\frac{n\epsilon^2/m}{c_1\sigma^2 + c_2\epsilon}\right),$$

   where $\sigma^2 = \mathrm{var}(h(X_1, \ldots, X_m))$ and a, c_1, c_2 are universal constants.

2. (11 pts) Consider a random undirected network, where $A_{ij} = A_{ji} \overset{iid}{\sim} Bernoulli(p_n)$ for $1 \leq i < j \leq n$. $A_{ii} = 0$ for $1 \leq i \leq n$. The degree of a node is defined as $d_i = \sum_j A_{ij}$. Consider the regime where $np_n/\log n \to \infty$. *Hint: remember, not all concentration inequalities work in this regime.*

   (a) (5 pts) Show that the degree of a fixed node concentrates around its expectation $(n-1)p_n$. Obtain the tail bound explicitly.

   (b) (4 pts) Can you obtain a uniform error bound on the degrees? That is, can you show that $\max_i \dfrac{|d_i - (n-1)p_n|}{(n-1)p_n}$ goes to zero in probability? If not, show why not. If yes, obtain the tail bound.

   (c) (2 pts) Denote by $d_{(n)}$ the maximum degree. Using the last two questions, show that the maximum degree also concentrates. Obtain the tail bound explicitly.

3. (11 pts) We will go back to finding the covering number of infinite dimensional ellipses in this problem. Given a collection of positive numbers $\{\mu_j, j = 1 \ldots d\}$, consider the ellipse

   $$\mathcal{E}_d = \{\theta \in \mathcal{R}^d : \sum_i \theta_i^2/\mu_i^2 \leq 1\},$$

   specified by the sequence $\mu_j = j^{-2\beta}$ for some parameter $\beta > 1/2$.

(a) (5 pts) Obtain an upper bound on the ϵ packing number of $\mathcal{E}_d$ under an appropriate distance metric. What metric do you think you should use?

(b) (6 pts) Now consider an infinite-dimensional ellipse $\mathcal{E}$, specified by the sequence $\mu_j = j^{-2\beta}$ for some parameter $\beta > 1/2$. Show that

$$\log N(\epsilon; \mathcal{E}, \|.\|_{\ell_2}) \leq C \left(\frac{1}{\epsilon}\right)^{1/\beta},$$

where $\|\theta - \theta'\|_{\ell_2}^2 = \sum_{i=1}^{\infty}(\theta_i - \theta_i')^2$ is the squared ℓ_2-norm on the space of square summable sequences.

4. (11 pts) Consider a random undirected network, where $A_{ij} = A_{ji} \overset{iid}{\sim} Bernoulli(p)$ for $1 \leq i < j \leq n$. $A_{ii} = 0$ for $1 \leq i \leq n$. Let T denote the number of triangles in this graph.

(a) (6 pts) Show that the variance of T is

$$\binom{n}{3}(p^3 - p^6) + c_1\binom{n}{4}(p^5 - p^6),$$

where c_1 is a universal constant.

(b) (5 pts) Now use the Efron Stein inequality to obtain an upper bound on the variance. Use the true variance as a guideline to get a tight upper bound.