Practice Final

SDS384

May 16, 2019

This exam three short and 4 long questions. You will have to answer **all short questions, and 3 out of 4 long questions**. The assigned points are noted next to each question; the total number of points is 50. You have 180 minutes to answer the questions.

Please answer all problems in the space provided on the exam. Use extra pages if needed. Of course, please put your name on extra pages.

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)

1. (5 pts) Let X be the set of binary strings of length 4 (i.e. the set of 4-character strings consisting only of 0 and 1). Let the class H be the set of schemas over X, where a schema consists of the symbols 1, 0, and $*$, where $*$ matches either 0 or 1. For example, h =1*** returns true for any string that starts with 1, and false for everything else. Similarly, $h = * * **$ returns true for all strings. Is X shattered by H? No. Consider the labeling $(0000, 1111), +$ and everything else is labeled $-$. This cannot be shattered by a schema.

2. (6 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 $\sigma_n^2 = \text{var}(T_n)$. Using the Lindeberg Feller theorem, show that

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

provided $\max\limits_{j \le n} \dfrac{z_{nj}^2}{\sum_{j=1}^n z_{nj}^2} \to 0$ as $n \to \infty$

Check Q3 in HW 1.

3. (6 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 $\bar{X}_n$.

   Check $E[X_n] = 0$ and $\text{var}(X_n) = n$. Let $Z_n = \sum_i X_i$ and $B_n^2 = \text{var}(Z_n) = \sum_i i = n(n+1)/2$ and $\text{var}(X_n)/B_n^2 \to 0$, so the Lindeberg condition is necessary and sufficient. The Lindeberg condition involves the following quantity.

$$\begin{aligned}\frac{1}{B_n}\sum_i E[X_i^2 1(|X_i| \geq \epsilon B_n)] &= \frac{\sqrt{2}}{\sqrt{n(n+1)}}\sum_i E[X_i^2 1(|X_i| \geq \epsilon\sqrt{n(n+1)/2})] \\ &\leq \frac{\sqrt{2}}{n}\sum_i i1(i \geq \epsilon^2 n(n+1)/2)\end{aligned}$$

   For all $j \leq n$ when $n+1 > 2/\epsilon^2$, summands in the above are zero. So we have:

$$\frac{1}{B_n}\sum_i E[X_i^2 1(|X_i| \geq \epsilon B_n)] = 0$$

   for all $n \geq 2/\epsilon^2$. So LC is satisfied and so $\sum_i X_i/B_n \xrightarrow{d} N(0,1)$, i.e. $\bar{X}_n \xrightarrow{d} N(0, 1/2)$

2 Long questions (33 points)

There are 4 long questions. Please answer any three of them.

1. (11 pts) Let $X_1, \dots, X_n$ be iid random variables. Consider the V statistic $V_n = \dfrac{\sum_{i=1}^n \sum_{j=1}^n h(x_i, x_j)}{n^2}$, where h is a symmetric kernel such that $E[h^2(X_i, X_j)] < \infty$. (Solution due to Garvesh Raskutti)

   (a) (5 pts) Let U_n be the corresponding U statistic. Show that $V_n - U_n \xrightarrow{P} 0$.

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} h(X_i, X_j) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j) \\ V_n &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1} h(X_i, X_j) = \frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) + \frac{1}{n^2} \sum_{i} h(X_i, X_I) \end{aligned}$$

So

$$\begin{aligned} V_n - U_n &= \left(\frac{1}{n^2} - \frac{1}{n(n-1)} \right) \sum_{i \neq j} h(X_i, X_j) + \frac{1}{n^2} \sum_{i} h(X_i, X_I) \\ &= -\frac{1}{n} \left(\frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j) \right) + \frac{1}{n} \left(\frac{1}{n} \sum_{i} h(X_i, X_i) \right) \\ &= -\frac{1}{n} U_n + O_P\left(\frac{1}{n}\right) \end{aligned}$$

For the last term we used the LLN to get $\frac{1}{n} \sum h(X_i, X_i) = O_P(1)$. By the asymptotic normality of U-statistics, we know[2] $\sqrt{n}(U_n - \theta) = O_p(1)$, which implies $U_n = \theta + O_p(1/\sqrt{n})$ and $n^{-1} U_n = O_p(1/n)$. This gives $V_n - U_n = O_p(1/n) = o_p(1)$. The result now follows from Slutsky's theorem.

(b) (2 pts) What is the asymptotic distribution of V_n? Why?

V_n has the same distribution as U_n which is normal.

(c) (4 pts) Can we write V_n as a U statistic U_n? That is, can we find a symmetric function $g(x_i, x_j)$ such that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(x_i, x_j) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} g(x_i, x_j)$$

. Why or why not?

No. For any fixed n we can find a g that satisfies the equation, but we need it to be true for all n. The form of $g(x_1, x_2)$ is already determined in the case $n = 2$:

$$\frac{1}{4}\left[h(x_1, x_1) + 2h(x_1, x_2) + h(x_2, x_2)\right] \quad = \quad g(x_1, x_2)$$

Does this work for $n = 3$? Plugging in, we get

$$\begin{aligned}
\frac{1}{\binom{3}{2}} \sum_{1 \le i < j \le 3} g(x_i, x_j) &= \frac{1}{3} \cdot \frac{1}{4} \sum_{1 \le i < j \le 3} \left[h(x_i, x_i) + 2h(x_i, x_j) + h(x_j, x_j)\right] \\
&= \frac{1}{12}\left[\sum_{i=1}^{3}\sum_{j=1}^{3} h(x_i, x_j) + \sum_{k=1}^{3} h(x_k, x_k)\right] \\
&= \frac{9}{12}\left(\frac{1}{3^2}\sum_{i=1}^{n}\sum_{j=1} h(x_i, x_j)\right) + \frac{1}{12}\sum_{k=1}^{3} h(x_k, x_k) \\
&\neq \frac{1}{3^2}\sum_{i=1}^{n}\sum_{j=1} h(x_i, x_j)
\end{aligned}$$

2. (11 pts) Let $X_1, \dots X_n$ be i.i.d $N(0, 1)$ random variables. Let $X_{(n)}$ be the largest order statistic.

   (a) (6 pts) Show that $X_{(n)}$ concentrates around $E[X_{(n)}]$. Obtain the tail bound explicitly. If you use Lipschitz continuity, please provide a proof of that. Use Gaussian Lipschitz theorem. First show that $X_{(n)}$ is 1-Lipschitz (you have done this in your homework).

(b) (4 pts) Obtain an upper bound on $E[X_{(n)}]$.

Use the finite class lemma. $E[X_{(n)}] \leq \sqrt{2 \log n}$.

(c) (1 pts) Using the above two answers, can you upper bound $X_{(n)}$?

$X_{(n)} \leq \sqrt{2 \log n} + \epsilon$ with probability $1 - \exp(-2\epsilon^2)$.

3. (11 pts) Let $\mathcal{F}_d$ be the set of L Lipschitz functions

$$\mathcal{F}_d = \{g : [0,1] \to [-1,1] : g(0) = 0, \text{ and } |g(x) - g(y)| \leq L|x-y|, \forall x, y \in [0,1]\}.$$

Recall the fact that $\log N(\epsilon; \mathcal{F}_d, \|.\|_\infty) = O((L/\epsilon)^d)$.

(a) (1 pts) Consider the process $Y_f := \langle \epsilon, f(X_1^n) \rangle$, where ϵ are n iid Rademacher random variables and $X_1^n = (X_1, \ldots, X_n)$ are n iid random variables in $[0,1]$. Is $Y_f - Y_g$ a subgaussian process? If so, under what distance metric? $d_X = \|f(X_1^n) - g(X_1^n)\|$

(b) (2 pts) What is the diameter of the function class $\mathcal{F}_d$ under this distance metric? $\|f(X_1^n) - g(X_1^n)\| \leq \sqrt{n} \max_i |f(X_i) - g(X_i)| \leq 2\sqrt{n}$

(c) (6 pts) Using the one step discretization bound, prove that

$$\mathcal{R}_{\mathcal{F}_d} \leq c_{d,L} \left(\frac{1}{n}\right)^{C_d},$$

where $c_{d,L}$ is a constant which depends only on d and L and C_d is a constant which depends on d. Derive C_d. Bonus points for deriving $c_{d,L}$.

We have $\log N(\epsilon, \mathcal{F}(X_1^n), d_X) \leq \log N(\epsilon/\sqrt{n}, \mathcal{F}(X_1^n), \|.\|_\infty) = (L\sqrt{n}/\epsilon)^d$. Plugging in

$$\mathcal{R}_F \leq 2\frac{\epsilon}{\sqrt{n}} + c\sqrt{\frac{L^d n^{d/2-1}}{\epsilon^d}}$$

Optimizing over ϵ we get:

$$\epsilon = \left(\frac{cd\sqrt{L}}{4}\right)^{2/(d+2)} n^{d/(2(d+2))}$$

Plugging in, we get

$$\mathcal{R}_{\mathcal{F}} \leq c_d \left(\frac{L}{n}\right)^{1/(d+2)}$$

(d) (2 pts) Recall the smoothly parametrized function class we studied in class. To remind you, in this case, $\mathcal{F}$ is a class of parametric functions $\mathcal{F}_L := \{f(\theta, .) : \theta \in B_2\}$, where B_2 is the unit L_2 ball in $\mathbb{R}^d$. Assume that $\mathcal{F}_L$ is closed under negation. f is L Lipschitz w.r.t. the Euclidean distance on Θ, i.e. $|f(\theta, .) - f(\theta', .)| \leq L\|\theta - \theta'\|_2$. Briefly compare the Rademacher complexity you obtained in the last part with $\mathcal{R}_{\mathcal{F}_L}$. You can cite your lecture notes to get $\mathcal{R}_{\mathcal{F}_L}$.

Rademacher complexity of a parametric function class is $\sqrt{d/n}$ (up-to logarithmic factors). Here the dependence on d is far worse, which is not surprising since the nonparametric function class is a lot richer than a particular parametrized function class.

4. (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.

   Use Bernstein's inequality to get

$$P(|d_n - (n-1)p_n| \geq t) \leq 2\exp(-\frac{t^2/2}{np_n + t/3})$$

   Using $t = \sqrt{4cnp_n \log n}$, the above tail becomes n^{-c}.

(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.

Using the union bound, $P(\max_i |d_i - (n-1)p_n| \geq \sqrt{4cnp_n \log n}) \leq n^{1-c}$. Using $c = 2$, the above tail becomes n^{-1}.

(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.

With probability at least $1 - 1/n$,

$|d_{(n)} - (n-1)p_n| \leq \max_i |d_i - (n-1)p_n| \leq \sqrt{8np_n \log n}$.