Final

SDS384

Spring 2022

This exam has 5 short and 4 long questions. You will have to answer **four 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 May 14th. Use the latex file format provided.

Read each question carefully, **show your work** and **clearly present your answers**. If you just show the final answer without any reasoning, you will **not** get partial credit. You will also **not** get credit for incorrect explanation of the correct answer.

You can use a fact or result in the book/lecture notes/homework. But if I gave you a problem which you have seen in the book and/or worked on already as part of your HW problems, you will still need to provide an answer **in your own words**.

This exam is open book (Martin Wainwright's book). You can look at your class notes and HW solutions. But you cannot use *any* other material.

Good Luck!

Name: _____

UTeid: _____

1 Short questions (20 points)

Please answer any four of the short questions. If you do all 5, I will do best 4 out of 5.

1. (5 pts) Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of random variable with density f on the real line. A standard estimate of f is the kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where $K : \Re \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)dt = 1$, and h is a bandwidth parameter. We will measure the quality of $\hat{f}$ using $\|\hat{f} - f\|_1 := \int_{-\infty}^{\infty} |\hat{f}(t) - f(t)|dt$. Prove that:

$$P(\|\hat{f} - f\|_1 \geq E\|\hat{f} - f\|_1 + \delta) \leq e^{-cn\delta^2},$$

where c is some constant.

2. (5 pts) Let $X_1, \ldots, X_n$ be IID uniform$(0, \tau)$ random variables. Consider the U statistic U_n corresponding to the kernel $h(x, y) = |x - y|$. What is the limiting distribution of this U statistic? Write your answer in the following form:

$$a_n(U_n - \theta_1) \xrightarrow{d} X, \qquad X \sim f_{\theta_2}$$

where a_n is some deterministic sequence, $\theta_1 = EU_n$ and f is the limiting distribution (e.g. normal/exponential etc) parametrized by θ_2. Provide expressions for a_n, θ_1, θ_2 and specify what the limiting distribution is.

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

4. (5 pts) Consider datapoints in a two dimensional unit circle centered at the origin. Now consider a function class $\mathcal{F}_x$ which is a set of linear classifiers in $\mathbb{R}^2$ such that the distance of any datapoint from the classifier is at least x, where x is some non-negative number. What is the VC-dimension of $\mathcal{F}_{3/4}$?

5. (5 pts) Consider the autoregressive sequence:

$$X_n = \beta X_{n-1} + \epsilon_n, \quad \text{for } n = 1, 2, \ldots$$

where $\epsilon_1, \epsilon_2, \ldots$ are IID random variables with $E[\epsilon_i] = \mu$ and $\text{var}(\epsilon_i) = \sigma^2$. $X_0 = 0$ and $-1 \leq \beta < 1$. Show that

(a) For $\beta \in (-1, 1)$,

$$\sqrt{n}\left(\bar{X}_n - \frac{\mu}{1-\beta}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{(1-\beta)^2}\right)$$

(b) For $\beta = -1$,

$$\sqrt{n}\left(\bar{X}_n - \frac{\mu}{2}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{2}\right)$$

2 Long questions (30 points)

Please answer any three of the long questions. If you do all 4, I will do best 3 out of 4.

1. (10 pts) Let $X_i \in \mathbb{R}^p, i = 1 \dots n$ be IID random variables such that $X_i \sim N(0, I_{p\times p})$ where $I_{p\times p}$ is the $p \times p$ identity matrix. Define the function class $\mathcal{F} = \{f : \mathbb{R}^p \to \mathbb{R} | f(x_1, \dots, x_p) = \beta^T x; \|\beta\|_1 \leq R\}$, where $\beta^T x = \sum_{i=1}^p \beta_i x_i$. We will prove that $\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \xrightarrow{P} 0$.

   (a) (5pts) Show that

   $$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \leq \frac{R}{\sqrt{n}} \max_{1\leq j\leq p} |Z_j|,$$

   where Z_j's are IID standard normal random variables.

   (b) (5pts) Now show that, as long as $R\sqrt{\log p/n} \to 0$,

   $$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \xrightarrow{P} 0.$$

2. (10 pts) Consider the set with s sparse vectors in the Euclidean unit ball. The $\|u\|_0$ norm counts the number of non-zero elements in a vector u.

   $$\mathcal{S}_d(s) = \{\theta \in \mathbb{R}^d | \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\}.$$

   We will prove the following bound on the Gaussian complexity

   $$\mathcal{G}(\mathcal{S}_d(s)) \leq C\sqrt{s\log(ed/s)}, \tag{1}$$

   for some constant C.

   (a) (Bonus points:) Show that $\mathcal{G}(\mathcal{S}_d(s)) = E\left[\max_{|S|=s} \|w_S\|_2\right]$ where $w_S \in \mathbb{R}^{|S|}$ is the sub-vector of $w_1, \dots, w_d$ indexed by the subset $S \subset \{1, \dots, d\}$.

   (b) (4 pts) Show that for any fixed subset S of size s,

   $$P\left(\|w_S\|_2 \geq \sqrt{s} + \delta\right) \leq \exp(-\delta^2/2),$$

   where C is some positive constant.

   (c) (6 pts) Use (a) and (b) to prove Eq 1.

3. (10 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) (5 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.

4. (10 pts) Let $X_1, \ldots, X_n$ be IID random variables. Consider the following statistic

$$Y_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j),$$

where h is a bounded symmetric kernel with $Eh^2(X_1, X_2) < \infty$, and $\mathrm{var}(E[h(X_1, X_2)|X_1]) > 0$.

(a) (4 pts) Is Y_n an unbiased estimator of $\theta := Eh(X_1, X_2)$? Prove or provide a counter example.

(b) (5 pts) Establish the asymptotic distribution of $\sqrt{n}(Y_n - \theta)$ (you need to provide the limiting distribution and the parameter of the limiting distribution, like short question 2). Clearly state what results you are using for full credit. *Hint: You can use existing results on convergence of U statistics.*

(c) (1 pt) Do you think we can find a symmetric kernel g such that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j) = \frac{1}{\binom{n}{2}}\sum_{i<j} g(X_i, X_j).$$

Why or why not?