Math 281A **2019 Fall Final** **Dec 12, 2019**

Name: ____________________ **PID:** ____________________

Do not turn the page until told to do so.

1. No calculators, tablets, phones, or other electronic devices are allowed during this exam.
2. Read each question carefully and answer each question completely.
3. Show all of your work. No credit will be given for unsupported answers, even if correct.
4. **Please answer questions within the spaces provided.** If you do need some more room, use the back side of the same piece of paper and clearly label the question.
5. If you are unsure of what a question is asking for, **do not hesitate to ask an instructor or course assistant for clarification**.
6. This exam has 9 pages.

Question	*Points Available*	*Points Earned*
1	10	
2	10	
3	10	
4	10	
5	15	
6	15	
TOTAL	70	

1. [10 points] Let X_n be the maximum of a random sample $Y_1, \ldots, Y_n$ from the density $p(x) = 2(1-x)I(0 \leq x \leq 1)$. Find constants a_n and b_n such that $b_n(X_n - a_n)$ converges in distribution to a non-degenerate limit.

Solution: $a_n = 1$ and $b_n = \sqrt{n}$. For any $x \leq 0$, we have

$$\begin{aligned}\mathbb{P}(\sqrt{n}(X_n - 1) \leq x) &= \mathbb{P}(X_n \leq x/\sqrt{n} + 1) = [\mathbb{P}(Y \leq x/\sqrt{n} + 1)]^n \\ &= (1 - x^2/n)^n \to e^{-x^2}.\end{aligned}$$

2. [10 points] Let $Z_1, \ldots, Z_n$ be independent standard normal variables. Show that the vector $U = (Z_1, \ldots, Z_n)^\intercal / N$, where $N^2 = \sum_{i=1}^n Z_i^2$, is uniformly distributed over the unit sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$ in the sense that U and OU are identically distributed for every orthogonal transformation O of $\mathbb{R}^n$.

   Solution: Denote $Z = (Z_1, \ldots, Z_n)^\intercal$, so $U = Z/||Z||_2$. Given that $Z \sim \mathcal{N}(0, I_n)$, it's easy to see that for any orthogonal transformation O of $\mathbb{R}^n$, $OZ \sim \mathcal{N}(0, I_n)$ and $||OZ||_2 = ||Z||_2$. So

$$OU = \frac{OZ}{||Z||_2} = \frac{OZ}{||OZ||_2} \sim \frac{Z}{||Z||_2} = U.$$

3. [10 points] Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed with mean μ and finite variance σ^2. Find the asymptotic distribution of $\bar{X}_n^2$ (after it is properly normalized), where $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$.

   Solution: By central limit theorem, we have

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

   To get the asymptotic distribution of $\bar{X}_n^2$, we discuss two cases.

   <u>Case 1</u>: $\mu \neq 0$. Applying delta method gives

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{d} \mathcal{N}(0, 4\mu^2\sigma^2).$$

   <u>Case 2</u>: $\mu = 0$. Then

$$\sqrt{n}\frac{\bar{X}_n}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

   Taking square on both sides with continuous mapping gives

$$n\frac{\bar{X}_n^2}{\sigma^2} \xrightarrow{d} \chi_1^2.$$

4. [10 points] Suppose $X_n \sim \text{binomial}(n, p)$, where $0 < p < 1$. (a). Find the asymptotic distribution of $g(X_n/n) - g(p)$, where $g(x) = \min\{x, 1-x\}$. (b) Show that $h(x) = \sin^{-1}(\sqrt{x})$ is a variance-stabilizing transformation for X_n/n. This is called the *arcsine transformation* of a sample proportion. **Hint**: $\frac{d}{du}\sin^{-1}(u) = 1/\sqrt{1-u^2}$.

Solution:

(a). $g(x)$ is not differentiable at $1/2$, so we need to discuss two cases.

Case 1: $p \neq 1/2$. Central limit theorem gives us

$$\sqrt{n}(X_n/n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Since $g(x)$ is differentiable at p, and $g'(p) = \pm 1$. Applying delta method yields

$$\sqrt{n}(g(X_n/n) - g(p)) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Case 2: $p = 1/2$. Denote

$$Y_n = \frac{X_n}{n} - \frac{1}{2},$$

then by central limit theorem, we have $\sqrt{n}Y_n \to \mathcal{N}(0, 1/4)$. Notice that

$$\begin{aligned} g(X_n/n) - g(1/2) &= g(Y_n + 1/2) - 1/2 = \min\{1/2 + Y_n, 1/2 - Y_n\} - 1/2 \\ &= 1/2 - |Y_n| - 1/2 = -|Y_n|. \end{aligned}$$

Since the absolute value function is continuous, by continuous mapping, we have

$$\sqrt{n}(g(X_n/n) - g(1/2)) = -\sqrt{n}|Y_n| \xrightarrow{d} -|Z|,$$

where $Z \sim \mathcal{N}(0, 1/4)$.

The absolute value of normal distribution is called folded normal distribution.

(b). It can be found that

$$h'(x) = \frac{1}{2\sqrt{x(1-x)}}.$$

So $h'(p)^2 p(1-p) = 1/4$ is a constant, and $h(x)$ is a variance-stabilizing transformation for X_n/n.

5. [15 points] Suppose that we observe data in pairs $(X, Y) \in \mathbb{R}^d \times \{\pm 1\}$, where the data come from a logistic model with $X \sim P_0$ and $p_{Y|X}(y|x) = 1/(1 + e^{-y \cdot x^\intercal \theta_0})$. Define the log-loss function $\ell_\theta(y|x) = \log(1 + e^{-y \cdot x^\intercal \theta})$. Let $\hat{\theta}_n$ minimize the empirical logistic loss $L_n(\theta) = (1/n) \sum_{i=1}^n \ell_\theta(Y_i|X_i) = (1/n) \sum_{i=1}^n \log(1 + e^{-Y_i X_i^\intercal \theta})$ from pairs (X_i, Y_i) drawn from the logistic model with parameter θ_0. Assume that the covaraites $X_i \in \mathbb{R}^d$ are i.i.d. and satisfy $\mathbb{E}(X_i X_i^\intercal) = \Sigma \succ 0$ and $\mathbb{E}\|X_i\|_2^4 < \infty$.

(a) Let $L(\theta) = \mathbb{E}_{\theta_0}\{\ell_\theta(Y|X)\}$ be the population logistic loss. Show that the second order derivative evaluated at θ_0 is positive definite.

Solution: We can interchange expectation and derivative since $\ell_\theta(\cdot)$ is smooth enough.

$$\nabla L(\theta) = \mathbb{E}\big[(-1/(1 + e^{YX^\intercal \theta}))YX\big],$$
$$\nabla^2 L(\theta) = \mathbb{E}\big[(e^{YX^\intercal \theta}/(1 + e^{YX^\intercal \theta})^2)XX^\intercal\big].$$

For any $u \in \mathbb{R}^d$, we have

$$u^\intercal \nabla^2 L(\theta) u = \mathbb{E}\big[(e^{YX^\intercal \theta}/(1 + e^{YX^\intercal \theta})^2)(u^\intercal X)^2\big] \geq 0.$$

To further show that $\nabla^2 L(\theta) \succ 0$, if there is a $u \in \mathbb{R}^d$ such that $u^\intercal \nabla^2 L(\theta) u = 0$, then $u^\intercal \mathbb{E}(XX^\intercal) u = 0$. This is a contradiction with $\mathbb{E}(X_i X_i^\intercal) = \Sigma \succ 0$.

(b) Under these assumptions show that $\hat{\theta}_n$ is consistent estimator of θ_0 as $n \to \infty$. Provide details of your work.

Solution: By Taylor expansion of $L_n(\theta)$ around θ_0,

$$L_n(\theta) = L_n(\theta_0) + \nabla L_n(\theta_0)^\intercal(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\intercal \nabla^2 L_n(\tilde{\theta})(\theta - \theta_0), \tag{1}$$

where $\tilde{\theta}$ is between θ_0 and θ.

For the <u>gradient</u>, notice that $\mathbb{E}\nabla L_n(\theta_0) = \nabla L(\theta_0) = 0$, and by weak law of large numbers, we have $\nabla L_n(\theta_0) \xrightarrow{p} 0$. Thus, for any $\epsilon > 0$,

$$\mathbb{P}(||\nabla L_n(\theta_0)||_2 \leq \epsilon) \to 1. \tag{2}$$

For the <u>Hessian</u> matrix, consider general value of θ,

$$\nabla^2 L_n(\theta) - \nabla^2 L_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal \left[\frac{\exp(Y_i X_i^\intercal \theta)}{(1 + \exp(Y_i X_i^\intercal \theta))^2} - \frac{\exp(Y_i X_i^\intercal \theta_0)}{(1 + \exp(Y_i X_i^\intercal \theta_0))^2}\right].$$

If we define $\phi(t) = e^t/(1 + e^t)^2$, then it satisfies $-1 \leq \phi'(t) \leq 1$, so $\phi(\cdot)$ is 1-Lipschitz continuous. Using this notation, the above display becomes

$$\nabla^2 L_n(\theta) - \nabla^2 L_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal(\phi(Y_i X_i^\intercal \theta) - \phi(Y_i X_i^\intercal \theta_0)).$$

Consider any $u \in \mathbb{R}^p$ with $||u||_2 = 1$, by the above Lipschitz continuity and Cauchy-

Schwarz inequality, we have

$$\begin{aligned}\left|u^\intercal\{\nabla^2 L_n(\theta)-\nabla^2 L_n(\theta_0)\}u\right| &= \frac{1}{n}\sum_{i=1}^n (X_i^\intercal u)^2(\phi(Y_iX_i^\intercal\theta)-\phi(Y_iX_i^\intercal\theta_0))\\ &\le \frac{1}{n}\sum_{i=1}^n (X_i^\intercal u)^2|Y_iX_i^\intercal(\theta-\theta_0)|\\ &\le ||\theta-\theta_0||_2\times\frac{1}{n}\sum_{i=1}^n ||X_i||_2(X_i^\intercal u)^2,\end{aligned}$$

which further implies the ℓ_2-operator norm / spectral norm

$$||\nabla^2 L_n(\theta)-\nabla^2 L_n(\theta_0)||_2 \le ||\theta-\theta_0||_2\times\left|\left|\frac{1}{n}\sum_{i=1}^n ||X_i||_2 X_iX_i^\intercal\right|\right|_2. \tag{3}$$

Now since $\mathbb{E}\|X_i\|_2^4<\infty$, applying weak law of large numbers on the matrix on the RHS of (3) gives

$$\frac{1}{n}\sum_{i=1}^n ||X_i||_2X_iX_i^\intercal \xrightarrow{p} \mathbb{E}\big[||X_i||_2X_iX_i^\intercal\big].$$

Combining this with (3) means there exists a constant C, such that

$$\mathbb{P}\big(||\nabla^2 L_n(\theta)-\nabla^2 L_n(\theta_0)||_2\le C||\theta-\theta_0||_2\big)\to 1. \tag{4}$$

Now denote $\lambda=\lambda_{\min}(\nabla^2 L(\theta_0))$, from part (a), we know $\lambda>0$. Because of (4) and weak law of large numbers $\nabla^2 L_n(\theta_0)\xrightarrow{p}\nabla^2 L(\theta_0)$, there exists $\delta>0$ sufficiently small such that for any $\theta\in\{\theta:||\theta-\theta_0||_2\le\delta\}$,

$$\mathbb{P}\left(\nabla^2 L_n(\theta)\succeq\frac{\lambda}{2}I_p\right)\to 1. \tag{5}$$

Applying (2) and (5) to (1) yields that for any $\theta\in\{\theta:||\theta-\theta_0||_2\le\delta\}$, with probability tending to 1,

$$L_n(\theta)\ge L_n(\theta_0)-\epsilon||\theta-\theta_0||_2+\frac{\lambda}{4}||\theta-\theta_0||_2^2,$$

and if $||\theta-\theta_0||>4\epsilon/\lambda$, then $-\epsilon||\theta-\theta_0||_2+\lambda||\theta-\theta_0||_2^2/4>0$, which means θ cannot minimize $L_n(\theta)$, so $||\hat\theta_n-\theta_0||\le\min\{4\epsilon/\lambda,\delta\}$. Finally, by taking $\epsilon>0$ arbitrarily small, we therefore have

$$\hat\theta_n\xrightarrow{p}\theta_0.$$

(c) Find the asymptotic distribution of $\sqrt{n}(\hat\theta_n-\theta_0)$, provided that it is consistent. You may assume $d=1$.

Solution: When $d=1$, the Fisher information is

$$I_{\theta_0}=\mathbb{E}\big[X^2e^{X\theta_0}/(1+e^{X\theta_0})^2\big],$$

and under regularity conditions, we have

$$\sqrt{n}(\hat\theta_n-\theta_0)\xrightarrow{d}\mathcal{N}(0,1/I_{\theta_0}).$$

6. [15 points] Let $X_1, \dots, X_n$ be a data sample of a continuous random variable X with distribution function F and density f. From the kernel density estimator $\hat{f}_h(x) = (1/n)\sum_{i=1}^n K_h(x - X_i)$, where $K_h(u) = K(u/h)/h$, one can construct a kernel estimator for the distribution function as $\hat{F}_h(x) = \int_{-\infty}^x \hat{f}_h(t)\mathrm{d}t$. Equivalently, we have

$$\hat{F}_h(x) = \frac{1}{n}\sum_{i=1}^n H\left(\frac{x - X_i}{h}\right), \quad \text{where } H(x) = \int_{-\infty}^x K(t)\mathrm{d}t.$$

Assume K is non-negative, symmetric around 0 and integrates to 1. Under smoothness conditions, find the leading term of the mean integrated square error (MISE) of $\hat{F}_h$, that is, $\mathrm{MISE}(\hat{F}_h) = \int_{-\infty}^{\infty} \mathbb{E}_f\{\hat{F}_h(x) - F(x)\}^2\mathrm{d}x$. What is the order of the optimal bandwidth?

Solution: With integration by parts, change of variable and Taylor expansion, the bias is

$$\begin{aligned}
\mathbb{E}_f\hat{F}_h(x) - F(x) &= \int H\left(\frac{x-t}{h}\right)\mathrm{d}F(t) - F(x) \\
&= H\left(\frac{x-t}{h}\right)F(t)\Big|_{-\infty}^{\infty} - \int F(t)\mathrm{d}H\left(\frac{x-t}{h}\right) - F(x) \\
&= \frac{1}{h}\int F(t)K\left(\frac{x-t}{h}\right)\mathrm{d}t - F(x) \\
&= \frac{1}{h}\int F(t)K\left(\frac{t-x}{h}\right)\mathrm{d}t - F(x) \\
&= \int F(x+hy)K(y)\mathrm{d}y - F(x) \\
&\sim \int \{F(x) + hyf(x) + h^2y^2f'(x)/2\}K(y)\mathrm{d}y - F(x) \\
&= F(x)\int K(y)\mathrm{d}y + hf(x)\int yK(y)\mathrm{d}y + \frac{h^2}{2}f'(x)\int y^2K(y)\mathrm{d}y - F(x) \\
&\sim \frac{h^2}{2}f'(x)\int y^2K(y)\mathrm{d}y := h^2B_f(x). \qquad (6)
\end{aligned}$$

The variance is

$$\mathrm{Var}_f\hat{F}_h(x) = \frac{1}{n}\left[\int H^2\left(\frac{x-t}{h}\right)\mathrm{d}F(t) - \left(\int H\left(\frac{x-t}{h}\right)\mathrm{d}F(t)\right)^2\right] \qquad (7)$$

For the first term,

$$\begin{aligned}
\int H^2\left(\frac{x-t}{h}\right)\mathrm{d}F(t) &= H^2\left(\frac{x-t}{h}\right)F(t)\Big|_{-\infty}^{\infty} - \int F(t)\mathrm{d}H^2\left(\frac{x-t}{h}\right) \\
&= \frac{2}{h}\int F(t)H\left(\frac{x-t}{h}\right)K\left(\frac{x-t}{h}\right)\mathrm{d}t \\
&= \frac{2}{h}\int F(t)\left[1 - H\left(\frac{t-x}{h}\right)\right]K\left(\frac{t-x}{h}\right)\mathrm{d}t \\
&= 2\int F(x+hy)(1-H(y))K(y)\mathrm{d}y \\
&\sim 2\int \{F(x) + hyf(x)\}(K(y) - H(y)K(y))\mathrm{d}y \\
&= F(x) - 2hf(x)\int yH(y)K(y)\mathrm{d}y. \qquad (8)
\end{aligned}$$

For the second term, it's been calculated in (6),

$$\left(\int H\left(\frac{x-t}{h}\right)\mathrm{d}F(t)\right)^2 \sim F(x)^2 \tag{9}$$

Combining (7), (8) and (9) gives

$$\begin{aligned}\mathrm{Var}_f \hat{F}_h(x) &\sim \frac{1}{n}F(x)(1-F(x)) - \frac{2h}{n}f(x)\int yH(y)K(y)\mathrm{d}y \\ &:= \frac{1}{n}F(x)(1-F(x)) - \frac{h}{n}V_f(x). \end{aligned} \tag{10}$$

Finally, with bias-variance decomposition and (6), (10), we thus have

$$\mathrm{MISE}(\hat{F}_h) \sim h^4 \int B_f^2(x)\mathrm{d}x + \frac{1}{n}\int F(x)(1-F(x))\mathrm{d}x - \frac{h}{n}\int V_f(x)\mathrm{d}x,$$

and the optimal choice of h is

$$h = \left(\frac{\int V_f(x)\mathrm{d}x}{4n\int B_f^2(x)\mathrm{d}x}\right)^{1/3} \sim n^{-1/3}.$$