Math 281C **2020 Spring Final** **June 9, 2020**

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. If you are unsure of what a question is asking for, **do not hesitate to ask an instructor or course assistant for clarification**.
5. This exam has 7 pages.

Question	*Points Available*	*Points Earned*
1	50	
2	50	
TOTAL	100	

1. [50 points] Let $\theta \in \mathbb{R}^p$ and define

$$f(\theta) = \mathbb{E}\{F(\theta; X)\} = \int_{\mathcal{X}} F(\theta; x)\mathrm{d}P(x),$$

where $F(\cdot; x)$ is convex in its first argument (in θ) for all $x \in \mathcal{X}$, and P is a probability distribution. We assume $F(\theta; \cdot)$ is integrable for all θ. Recall that a function h is convex if

$$h(t\theta + (1-t)\theta') \le th(\theta) + (1-t)h(\theta') \quad \text{for all} \quad \theta, \theta' \in \mathbb{R}^p, \quad t \in [0,1].$$

Let $\theta_0 \in \arg\min_\theta f(\theta)$, and assume that f satisfies the following ν-strong convexity property:

$$f(\theta) \ge f(\theta_0) + \frac{\nu}{2}\|\theta - \theta_0\|^2 \quad \text{for all } \theta \text{ satisfying } \|\theta - \theta_0\| \le \beta,$$

where $\beta > 0$ is some constant. We also assume that $F(\cdot; x)$ is L-Lipschitz with respect to the norm $\|\cdot\|$, the Euclidean norm in $\mathbb{R}^p$.

Let $X_1, \ldots, X_n$ be an iid sample from P, and define $f_n(\theta) = (1/n)\sum_{i=1}^n F(\theta; X_i)$. Let

$$\hat{\theta}_n \in \arg\min_\theta f_n(\theta).$$

(a) Show that for any convex function $h : \mathbb{R}^p \to \mathbb{R}$, if there is some $r > 0$ and a point θ_0 such that $h(\theta) > h(\theta_0)$ for all θ such that $\|\theta - \theta_0\| = r$, then $h(\theta') > h(\theta_0)$ for all θ' with $\|\theta' - \theta_0\| > r$.

Solution: For any θ' with $\|\theta' - \theta_0\| > r$, there exists a $t \in (0,1)$, such that $\theta = t\theta_0 + (1-t)\theta'$, and $||\theta - \theta_0|| = r$. By the definition of convexity,

$$h(\theta_0) < h(\theta) \le th(\theta_0) + (1-t)h(\theta'),$$

which implies $h(\theta') > h(\theta_0)$.

(b) Show that f and f_n are convex.

Solution: For any $\theta, \theta' \in \mathbb{R}^p$, and $t \in [0,1]$,

$$\begin{aligned} f(t\theta + (1-t)\theta') &= \int_{\mathcal{X}} F(t\theta + (1-t)\theta'; x)\mathrm{d}P(x) \\ &\le t\int_{\mathcal{X}} F(\theta; x)\mathrm{d}P(x) + (1-t)\int_{\mathcal{X}} F(\theta'; x)\mathrm{d}P(x) \\ &= tf(\theta) + (1-t)f(\theta'). \end{aligned}$$

The proof for f_n is similar.

(c) Show that θ_0 is unique.

Solution: For any $\theta \ne \theta_0$, if $||\theta - \theta_0|| \le \beta$, then by the local strong convexity,

$$f(\theta) \ge f(\theta_0) + \frac{\nu}{2}\|\theta - \theta_0\|^2 > f(\theta_0),$$

and if $||\theta - \theta_0|| > \beta$, by the property of part (a), $f(\theta) > f(\theta_0)$. Hence, θ_0 is the unique minimizer of $f(\cdot)$.

(d) Let

$$\Delta(\theta, x) = \{F(\theta; x) - f(\theta)\} - \{F(\theta_0; x) - f(\theta_0)\}.$$

Show that $\Delta(\theta, X)$ (with $X \sim P$) is $4L^2\|\theta - \theta_0\|^2$-sub-Gaussian. [We say a random variable X with mean μ is ν^2-sub-Gaussian if $\log \mathbb{E} e^{\lambda(X-\mu)} \le \lambda^2\nu^2/2$ for all $\lambda \in \mathbb{R}$.]

Solution: We can check that $\mathbb{E}[\Delta(\theta, x)] = 0$, and

$$|\Delta(\theta, x)| \le |F(\theta; x) - F(\theta_0; x)| + |f(\theta) - f(\theta_0)| \le 2L||\theta - \theta_0||.$$

The sub-Gaussianity follows from Homework 1, question 1.

(e) Show that for some constant $\sigma < \infty$, which may depend on the parameters of the problem (you should specify this dependence in your solution),

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| \ge \sigma \cdot \frac{1+t}{\sqrt{n}} \right) \le C e^{-t^2}$$

for all $t \ge 0$, where $C < \infty$ is a numerical constant. [Hint: The quantity $\Delta_n(\theta) := (1/n)\sum_{i=1}^n \Delta(\theta, X_i)$ may be helpful, as may be the bounded differences inequality.]

Solution: Define

$$\Delta_n(\theta) := \frac{1}{n}\sum_{i=1}^n \Delta(\theta, X_i)$$

as in the hint. The following argument is mainly based on the global convexity and local strong convexity.

Let $r \le \beta$ be the deviation we want to show, and define a local ball $\Theta_r = \{\theta : ||\theta - \theta_0|| \le r\}$. If $\hat{\theta}_n \notin \Theta_r$, then there is a $t \in (0, 1)$ and $\theta' = t\theta_0 + (1-t)\hat{\theta}_n \in \partial\Theta_r$ such that $||\theta' - \theta_0|| = r$ and

$$f_n(\theta') \le t f_n(\theta_0) + (1-t) f_n(\hat{\theta}_n) \le f_n(\theta_0).$$

Combining this result with the local strong convexity gives

$$\begin{aligned} \frac{\nu}{2} r^2 = \frac{\nu}{2}||\theta' - \theta_0||^2 &\le f(\theta') - f(\theta_0) \\ &\le (f_n(\theta_0) - f(\theta_0)) - (f_n(\theta') - f(\theta')) \le \sup_{\theta \in \Theta_r} |\Delta_n(\theta)|. \end{aligned} \tag{1}$$

This means the event $\{\hat{\theta}_n \notin \Theta_r\}$ implies (1), then we derive a probabilistic bound for the local fluctuation $\sup_{\theta\in\Theta_r} |\Delta_n(\theta)|$.

Let $\Delta'_n(\theta)$ be the counterpart of $\Delta_n(\theta)$ with X_i replaced by X'_i, for some $i \in [n]$. By the Lipschitz continuity of F, it can be found that

$$\begin{aligned} \left| \sup_{\theta\in\Theta_r} |\Delta_n(\theta)| - \sup_{\theta\in\Theta_r} |\Delta'_n(\theta)| \right| &\le \sup_{\theta\in\Theta_r} |\Delta_n(\theta) - \Delta'_n(\theta)| \\ &\le \frac{2L}{n} \sup_{\theta\in\Theta_r} ||\theta - \theta_0|| = \frac{2Lr}{n}. \end{aligned}$$

Applying the one-sided bounded difference inequality gives

$$\mathbb{P}\left(\sup_{\theta\in\Theta_r} |\Delta_n(\theta)| \ge \mathbb{E}\left[\sup_{\theta\in\Theta_r} |\Delta_n(\theta)| \right] + t \right) \le \exp\left(-\frac{nt^2}{2L^2r^2} \right) \tag{2}$$

for any $t > 0$. In the next step, we give an upper bound for the expected supremum. Using similar technique as Homework 6, question 2(c), we can show that $\sqrt{n}\Delta_n(\theta)/(2L)$ is a sub-Gaussian process, and there is a constant c such that

$$\mathbb{E}\left[\sup_{\theta\in\Theta_r} |\Delta_n(\theta)|\right] \le c'\frac{L}{\sqrt{n}}\int_0^r \sqrt{\log(N(\Theta_r, ||\cdot||, \epsilon))}\mathrm{d}\epsilon \le c\frac{Lr\sqrt{p}}{\sqrt{n}}.$$

Combining this display and (2) with some algebra, we obtain

$$\mathbb{P}\left(\sup_{\theta\in\Theta_r} |\Delta_n(\theta)| \ge c\frac{Lr}{\sqrt{n}}(\sqrt{p}+t)\right) \le \exp(-t^2) \tag{3}$$

for any $t > 0$. Choosing $r = 2cL(\sqrt{p}+t)/(\nu\sqrt{n})$ and combining (1), (3) gives

$$\begin{aligned}\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \ge r) &\le \mathbb{P}\left(\sup_{\theta\in\Theta_r} |\Delta_n(\theta)| \ge \frac{\nu}{2}r^2\right) \\ &= \mathbb{P}\left(\sup_{\theta\in\Theta_r} |\Delta_n(\theta)| \ge c\frac{Lr}{\sqrt{n}}(\sqrt{p}+t)\right) \le \exp(-t^2)\end{aligned}$$

under the scaling condition $\sqrt{p}+t \lesssim \sqrt{n}$. Finally, taking

$$\sigma = \sigma(L, \nu, p) := \frac{2cL\sqrt{p}}{\nu}$$

completes the proof.

2. [50 points] In the phase retrieval problem, the goal is to recover a signal $\theta^* \in \mathbb{R}^p$ based on noisy observations of the magnitudes of inner products $\langle X_i, \theta^* \rangle$ with a sample of n vectors $X_1, \ldots, X_n \in \mathbb{R}^p$. In physical detectors, we observe a number of photons $Y_i \in \mathbb{N}$ (here $\mathbb{N}$ denotes the collection of all non-negative integers) that scale roughly with $\langle X_i, \theta^* \rangle^2$. This association is usually characterized via a Poisson regression model, that is, the distribution of Y_i given X_i is

$$Y_i | X_i \sim \text{Poisson}(\langle X_i, \theta^* \rangle^2).$$

Recall that $Y \sim \text{Poisson}(\lambda)$ if the probability mass function of Y is

$$p_\lambda(k) = \frac{e^{-\lambda} \lambda^k}{k!}, k = 0, 1, \ldots.$$

Consider the (conditional) expectation of negative log-likelihood

$$\varphi_i(\theta) = \mathbb{E}_{\theta^*}\{-\log p_{\langle X_i, \theta \rangle^2}(Y_i)\},$$

where the expectation is taken over $Y_i \sim \text{Poisson}(\langle X_i, \theta^* \rangle^2)$.

(a) Suppose that $Y \sim \text{Poisson}(\lambda_0)$ for some $\lambda_0 > 0$. Show that

$$\mathbb{E}\{-\log p_\lambda(Y)\} - \mathbb{E}\{-\log p_{\lambda_0}(Y)\} \geq \frac{1}{4} \min\left\{ |\lambda - \lambda_0|, \frac{(\lambda - \lambda_0)^2}{\lambda_0} \right\}.$$

Solution: It can be calculated that

$$\mathbb{E}\{-\log p_\lambda(Y)\} - \mathbb{E}\{-\log p_{\lambda_0}(Y)\} = \lambda - \lambda_0 + \lambda_0 \log \frac{\lambda_0}{\lambda}.$$

The desired result can be established by discussing two cases (1) $\lambda \geq 2\lambda_0$ and (2) $\lambda < 2\lambda_0$ separately.

(b) Let $g : \mathbb{R}^p \to \mathbb{R}$ be a twice-differentiable convex function and satisfy $\nabla^2 g(\theta) \succeq \lambda I_p$ (I_p is the $p \times p$ identity matrix) for all θ satisfying $\|\theta - \theta_0\| \leq c$. Show that

$$g(\theta) \geq g(\theta_0) + \nabla g(\theta_0)^\intercal (\theta - \theta_0) + \frac{\lambda}{2} \min\{\|\theta - \theta_0\|^2, c\|\theta - \theta_0\|\}.$$

Solution: If $||\theta - \theta_0|| \leq c$, the result follows directly from a Taylor expansion to the second order. If $||\theta - \theta_0|| > c$, there is a $t \in (0, 1)$ such that $\theta' = t\theta_0 + (1 - t)\theta$ and $||\theta' - \theta|| = c$. By the convexity,

$$\begin{aligned} tg(\theta_0) + (1-t)g(\theta) \geq g(\theta') &\geq g(\theta_0) + \nabla g(\theta_0)^\intercal (\theta' - \theta_0) + \frac{\lambda}{2} \|\theta' - \theta_0\|^2 \\ &= g(\theta_0) + (1-t)\nabla g(\theta_0)^\intercal (\theta - \theta_0) + \frac{\lambda}{2}(1-t)^2 \|\theta - \theta_0\|^2. \end{aligned}$$

Rearranging the above result gives

$$g(\theta) \geq g(\theta_0) + \nabla g(\theta_0)^\intercal (\theta - \theta_0) + \frac{\lambda}{2} c \|\theta - \theta_0\|.$$

(c) Show that

$$\varphi_i(\theta) - \varphi_i(\theta^*) \geq \frac{1}{4} \min\left\{ |\langle X_i, \theta - \theta^* \rangle \langle X_i, \theta + \theta^* \rangle|, \frac{|\langle X_i, \theta - \theta^* \rangle \langle X_i, \theta + \theta^* \rangle|^2}{\langle X_i, \theta^* \rangle^2} \right\}.$$

Solution: This is trivial from part (a) by taking $\lambda = \langle X_i, \theta \rangle^2$ and $\lambda_0 = \langle X_i, \theta^* \rangle^2$.

(d) Suppose that $X_i \in \mathbb{R}^p$ are random vectors satisfying

$$\mathbb{P}\big(|\langle X_i, v \rangle| \geq \epsilon \|v\|_2\big) \geq 1 - \epsilon \quad \text{and} \quad \mathbb{E}\langle X_i, \theta^* \rangle^2 \leq M^2 \|\theta^*\|_2^2$$

for all $\epsilon \geq 0$ and all vectors $v \in \mathbb{R}^p$. Show that for (numerical) constants c_0, c_1, for any $\delta \in (0, 1)$, if

$$\sqrt{\frac{p + \log(1/\delta)}{n}} \leq c_0,$$

then with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^{n} \{\varphi_i(\theta) - \varphi_i(\theta^*)\} \geq c_1 \min\left\{ d(\theta, \theta^*) \cdot \max\{\|\theta\|_2, \|\theta^*\|_2\}, \frac{d^2(\theta, \theta^*)}{M^2} \right\}$$

holds simultaneously for all $\theta \in \mathbb{R}^p$, where $d(\theta, \theta^*) := \min_{s \in \{1,-1\}} \|\theta + s\theta^*\|_2$ is the distance (ignoring sign) between θ and θ^*.

Solution: For any $\epsilon > 0$, define a boolean function class

$$\mathcal{F} = \{x : I\{|\langle x, u \rangle| \geq \epsilon, |\langle x, v \rangle| \geq \epsilon, \langle x, \theta^* \rangle^2 \epsilon \leq M^2 ||\theta^*||_2^2\} | u \in \mathbb{S}^{p-1}, v \in \mathbb{S}^{p-1}\}.$$

By Markov inequality,

$$\mathbb{P}(\langle x, \theta^* \rangle^2 \epsilon \geq M^2 ||\theta^*||_2^2) \leq \frac{\mathbb{E}\langle x, \theta^* \rangle^2}{M^2 ||\theta^*||_2^2 / \epsilon} \leq \epsilon,$$

so $Pf \geq 1 - 3\epsilon$, for any $f \in \mathcal{F}$. Notice that the event $\{|\langle x, u \rangle| \geq \epsilon\}$ is the union of two events defined with closed half-spaces, conducting similar analysis as Homework 3, question 1 with discussions in review notes, the VC-dimension of $\mathcal{F}$ can be bounded by

$$\mathcal{V}(\mathcal{F}) \leq cp$$

for some constant c. For any $t > 0$, applying bounded difference inequality with Theorem 1.3 of Lecture 8 gives us

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq C \sqrt{\frac{p + t}{n}} \right) \leq e^{-t},$$

where C is a constant. The above inequality can be equivalently stated as with probability at least $1 - \exp(-t)$,

$$P_n f \geq (1 - 3\epsilon) - C \sqrt{\frac{p + t}{n}} \tag{4}$$

holds for any $f \in \mathcal{F}$. Now, combining (4) with the lower bound from part (c), with probability at least $1 - \exp(-t) - 3\epsilon$,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{\varphi_i(\theta) - \varphi_i(\theta^*)\} \\
&\geq \frac{1}{4n}\sum_{i=1}^{n}\min\left\{|\langle X_i, \theta - \theta^*\rangle\langle X_i, \theta + \theta^*\rangle|, \frac{|\langle X_i, \theta - \theta^*\rangle\langle X_i, \theta + \theta^*\rangle|^2}{\langle X_i, \theta^*\rangle^2}\right\} \\
&\geq \frac{1}{4n}\left[(1-3\epsilon) - C\sqrt{\frac{p+t}{n}}\right]\min\left\{\epsilon^2||\theta - \theta^*||_2||\theta + \theta^*||_2, \frac{\epsilon^6||\theta - \theta^*||_2^2||\theta + \theta^*||_2^2}{M^2||\theta^*||_2^2}\right\} \\
&\geq \frac{1}{4n}\left[(1-3\epsilon) - C\sqrt{\frac{p+t}{n}}\right]\min\left\{\epsilon^2 d(\theta, \theta^*)\cdot\max\{\|\theta\|_2, \|\theta^*\|_2\}, \frac{\epsilon^6 d^2(\theta, \theta^*)}{M^2}\right\},
\end{aligned}
$$

where the last inequality comes from the fact

$$
\max\{\|\theta\|_2, \|\theta^*\|_2\} \leq \max\{||\theta - \theta^*||_2, ||\theta + \theta^*||_2\}, \tag{5}
$$

so that

$$
\min\{\|\theta - \theta^*\|_2, ||\theta + \theta^*||_2\}\cdot\max\{\|\theta\|_2, \|\theta^*\|_2\} \leq ||\theta - \theta^*||_2||\theta + \theta^*||_2.
$$

(5) can be easily verified by drawing a parallelogram.

Finally, taking $\epsilon > 0$ small enough and $t > 0$ satisfying the scaling condition $p + t \lesssim n$ completes the proof.