Statistics 300B Winter 2018 Final Exam

Due 24 Hours after receiving it

Directions: This test is open book and open internet, but *must* be done without consulting other students. Any consultation of other students or people is an honor code violation. Cite any results you use from the literature that we did not explicitly prove in class or homework.

Question 1 (Asymptotics and an integrated delta method)**:** Let T_n be an estimator satisfying $r_n(T_n - \theta) \xrightarrow{d} T \in \mathbb{R}^d$ for some random variable T and a non-random sequence $r_n \to \infty$. Assume $T_n \xrightarrow{a.s.} \theta$. Let μ be a finite measure on a set $\mathcal{X}$ and for $z : \mathcal{X} \to \mathbb{R}$, let

$$F(z) = \int f(x, z(x)) d\mu(x),$$

where $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$. Let $\phi : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ be twice differentiable in its first argument, where for $t, \theta \in \mathbb{R}^d$,

$$\phi(t, x) = \phi(\theta, x) + \langle \nabla\phi(\theta, x), t - \theta \rangle + \frac{1}{2}(t - \theta)^T \nabla^2 \phi(\widetilde{\theta}, x)(t - \theta)$$

for some $\widetilde{\theta} \in [t, \theta] = \{\lambda t + (1 - \lambda)\theta \mid \lambda \in [0, 1]\}$. Assume that f satisfies the Lipschitz condition

$$|f(x, a) - f(x, b)| \le L(x)|a - b| \quad \text{for } a, b \in \mathbb{R}$$

where L is square integrable, that is, $\int L(x)^2 d\mu(x) < \infty$. Define the process

$$Z_n(x) := r_n(\phi(T_n, x) - \phi(\theta, x)),$$

so that $Z_n : \mathcal{X} \to \mathbb{R}$. Assume that ϕ is smooth enough around θ that

$$\int \sup_{\|t - \theta\| \le \epsilon} \left\| \nabla^2 \phi(t, x) \right\|_{\text{op}}^2 d\mu(x) < \infty$$

for all small enough $\epsilon > 0$, and that $\int \|\nabla\phi(\theta, x)\|^2 d\mu(x) < \infty$. Show that

$$F(Z_n) \xrightarrow{d} F_{\text{fo}}(T)$$

where $F_{\text{fo}} : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$F_{\text{fo}}(t) := \int f(x, \langle \nabla\phi(\theta, x), t \rangle) d\mu(x).$$

Hint: You do *not* need any of the empirical process results we have proved to answer this question.

Question 2 (Logistic regression)**:** 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

$$p_\theta(y \mid x) = \frac{1}{1 + e^{-yx^T\theta}},$$

with log loss $\ell_\theta(y \mid x) = \log(1 + \exp(-yx^T\theta))$. Let $\widehat{\theta}_n$ minimize the empirical logistic loss,

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell_\theta(Y_i \mid X_i) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-Y_i X_i^T\theta})$$

for pairs X, Y drawn from the logistic model with parameter θ_0. Let $\widehat{\theta}_n = \operatorname{argmin}_\theta L_n(\theta)$. Assume in addition that the data $X_i \in \mathbb{R}^d$ are i.i.d. and satisfy

$$\mathbb{E}[X_i X_i^T] = \Sigma \succ 0 \;\text{ and }\; \mathbb{E}[\|X_i\|_2^4] < \infty.$$

That is, the second moment matrix of the X_i is positive definite.

(a) Let $L(\theta) = \mathbb{E}_0[\ell_\theta(Y \mid X)]$ denote the population logistic loss. Show that

$$\nabla^2 L(\theta_0) \succ 0.$$

that is, the Hessian of L at θ_0 is positive definite. You may assume that the order of differentiation and integration may be exchanged.

(b) Under these assumptions, argue that $\widehat{\theta}_n$ is consistent for θ_0, that is, that

$$\widehat{\theta}_n \stackrel{a.s.}{\to} \theta_0.$$

For the remainder of the question, assume that data are 1-dimensional and satisfy $x \in \{-1, 1\}$, as it makes things simpler.

(c) Give the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$. You may assume that $\widehat{\theta}_n$ is consistent. In one sentence or so, describe the effect of the parameter value θ_0 on the efficiency of $\widehat{\theta}_n$.

(d) In many scenarios—especially large-scale prediction problems—we do not particularly care about the value of the parameter, but we wish to make accurate predictions—with confidence—of a label y given data x. For example, we might compare our predicted logistic model

$$p_{\widehat{\theta}_n}(y \mid x) = \frac{1}{1 + \exp(-y\widehat{\theta}_n x)}$$

to the true logistic prediction $p_{\theta_0}(y \mid x)$. With this in mind, define the risk

$$R(\theta) := \mathbb{E}_0\left[|p_\theta(Y \mid X) - p_{\theta_0}(Y \mid X)|\right]$$

the expected absolute error in our predictions of labels (when the true distribution is P_{θ_0}). What is the asymptotic distribution of $\sqrt{n}R(\widehat{\theta}_n)$? In one sentence or so, describe the effect of the parameter value θ_0 on the efficiency of $\widehat{\theta}_n$.

Question 3 (1 bit estimators of location)**:** Let f be a symmetric continuous density with $f(x) > 0$ for all $x \in \mathbb{R}$. Suppose we receive data $X_i \in \mathbb{R}$ drawn i.i.d. from the location family of distributions with densities $f(\cdot - \theta)$, $\theta \in \mathbb{R}$, where $f : \mathbb{R} \to \mathbb{R}_+$ is Lipschitz continuous, and the goal is to estimate θ. However, there is an additional wrinkle: for reasons of communication and reducing memory and power usage, we may only observe a single bit associated with each X_i, that is, we observe $Z_i \in \{0, 1\}$, where each Z_i is a function of X_i.

To make this a bit more concrete, we assume that each Z_i is in the form of a threshold, that is,

$$Z_i := \mathbf{1}\left\{X_i \le t_i\right\} \tag{1}$$

where $t_i \in \mathbb{R}$ is real-valued.

(a) Assume that for each i, $t_i \equiv t$ is a constant and that we observe $Z_1, \ldots, Z_n$ of the form (1). Give a $\sqrt{n}$-consistent estimator $T_n = T_n(Z_1, \ldots, Z_n; t)$ of θ based on $Z_1, \ldots, Z_n$ and t.

(b) What is the asymptotic distribution of your estimator T_n?

(c) Now, suppose that we have a sample $\{X_i\}_{i=1}^N$ of size N, and we divide it into two parts: $S^{(1)} = \{X_1, \ldots, X_n\}$ and $S^{(2)} = \{X_{n+1}, \ldots, X_N\}$. On the first sample $S^{(1)}$, let $Z_i = \mathbf{1}\left\{X_i \le t\right\}$ as in part (a). Then define

$$\widetilde{t}_n := T_n(Z_1, \ldots, Z_n; t)$$

where $T_n(\cdot)$ is your estimator from part (a). Now, let

$$Z_i = \mathbf{1}\left\{X_i \le \widetilde{t}_n\right\}$$

for $i \in \{n+1, \ldots, N\}$. Assume that $n/N \to 0$ as $N \to \infty$. Construct an estimator $\widehat{\theta}_N$, a function of $Z_{n+1}, \ldots, Z_N$ and $\widetilde{t}_n$, such that

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{4f(0)^2}\right)$$

when the data X_i are i.i.d. with density $f(\cdot - \theta_0)$. Prove that you achieve this efficiency.

Hint: The Berry-Esseen theorem may be useful. To remind (or define this) for you, the Berry-Esseen theorem is as follows: if Y_i are i.i.d. and $\mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^3] < \infty$, then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}(\overline{Y}_n - \mathbb{E}[Y])}{\sqrt{\mathrm{Var}(Y)}} \le t \right) - \Phi(t) \right| \le \frac{\mathbb{E}[|Y - \mathbb{E}[Y]|^3]}{\mathrm{Var}(Y)^{3/2}} \cdot \frac{1}{\sqrt{n}}$$

where Φ is the standard normal CDF and $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ is the sample mean.

Question 4 (Stochastic convex optimization)**:** Let $\theta \in \mathbb{R}^d$ and define

$$f(\theta) := \mathbb{E}[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x)$$

be a function, 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

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

Let $\theta_0 \in \operatorname{argmin}_\theta f(\theta)$, and assume that f satisfies the following *ν-strong convexity* guarantee:

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

where $\beta > 0$ is some constant. We also assume that the instantaneous functions $F(\cdot; x)$ are $L(x)$-Lipschitz continuous over the set $\{\theta \in \mathbb{R}^d \mid \|\theta - \theta_0\| \le \beta\}$, meaning that

$$|F(\theta; x) - F(\theta'; x)| \le L(x) \left\|\theta - \theta'\right\| \quad \text{when } \|\theta - \theta_0\| \le \beta, \ \left\|\theta' - \theta_0\right\| \le \beta.$$

We assume the (local) Lipschitz constant is sub-Gaussian, that is,

$$\mathbb{E}[L(X)] < \infty \text{ and } \mathbb{E}\left[\exp\left(\lambda(L(X) - \mathbb{E}[L(X)])\right)\right] \le \exp\left(\frac{\sigma_{\text{Lip}}^2 \lambda^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}.$$

We also make the following assumption on the relative errors in $F(\theta; X)$ and $F(\theta_0; X)$. Define $\Delta(\theta, x) = [F(\theta; x) - f(\theta)] - [F(\theta_0; x) - f(\theta_0)]$. Then

$$\log\left(\mathbb{E}\left[\exp\left(\lambda \Delta(\theta, X)\right)\right]\right) \le \frac{\lambda^2 \sigma^2}{2} \|\theta - \theta_0\|^2 \quad \text{for all } \lambda \in \mathbb{R} \text{ if } \|\theta - \theta_0\| \le \beta.$$

Let $X_1, \ldots, X_n$ be an i.i.d. sample according to P, and define $f_n(\theta) := \frac{1}{n} \sum_{i=1}^n F(\theta; X_i)$ and let

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} f_n(\theta).$$

Show that there exists a numerical constant C such that for all suitably large n,

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\|^2 \ge C \frac{\sigma^2}{\nu^2 n} \left[\log \frac{1}{\delta} + d \cdot \widetilde{O}(1)\right]\right) \le \delta + \exp\left(-n \frac{\mathbb{E}[L(X)]^2}{C \sigma_{\text{Lip}}^2}\right)$$

for all $\delta \in (0,1)$, where $\widetilde{O}(1)$ hides constants logarithmic in n, $\mathbb{E}[L(X)]$, and ν^{-1}. (You do not need to show this, but n such that $C \frac{\sigma^2}{\nu^2 n} [\log \frac{1}{\delta} + d \cdot \widetilde{O}(1)] \le \beta^2$ is sufficiently large.)

Hint 1: You may use the following facts, which are proved in Question 7.10.

i. For *any* convex function h, 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 $\|\theta' - \theta_0\| > r$.

ii. The functions f and f_n are convex.

iii. The point θ_0 is unique.

Hint 2: The bounded differences inequality will not work here. You should show that f_n is locally Lipschitz with high probability.