Statistics 300B Winter 2017 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)**:** Suppose that $Y \in \mathcal{Y}$ follows a generalized linear model (GLM) conditional on $X \in \mathbb{R}^d$. That is, $Y \mid X = x$ has density (with respect to a base measure μ) parameterized by $\theta \in \mathbb{R}^d$ defined by

$$p_\theta(y \mid x) = \exp\left(T(y)x^T\theta - A(\theta; x)\right) \quad \text{where} \quad A(\theta; x) = \log \int \exp(T(y)x^T\theta) d\mu(y) \tag{1}$$

and $T : \mathcal{Y} \to \mathbb{R}$ is the sufficent statistic for Y. You may assume that $A(\theta; x) < \infty$ for all $\theta \in \mathbb{R}^d$. Recall that for each x, the function $A(\theta; x)$ is convex in θ and infinitely differentiable (in θ) by standard exponential family results. Suppose you are given an i.i.d. sample from the model (1) with parameter θ_0, where $X_i \overset{\text{iid}}{\sim} P$ for some distribution P and $Y_i \mid X_i$ follows the model (1). You choose the maximum likelihood estimator

$$\widehat{\theta}_n = \operatorname*{argmax}_\theta \left\{ \sum_{i=1}^n \log p_\theta(Y_i \mid X_i) \right\}.$$

Assume the consistency result that $\widehat{\theta}_n \xrightarrow{p} \theta_0$ (though you could prove this using the techniques you know from the class and Q4 from Problem Set 2), that $\widehat{\theta}_n$ exists for all suitably large n, and that you may always swap the order of integration and differentiation. In addition, assume that $\theta \mapsto \nabla^2 A(\theta, x)$ is $H(x)$-Lipschitz continuous, meaning that $\left|\left|\left| \nabla^2 A(\theta; x) - \nabla^2 A(\theta'; x) \right|\right|\right|_{\text{op}} \leq H(x) \left\|\theta - \theta'\right\|$, where $\mathbb{E}[H(X)^2] < \infty$. Assuming that $\mathbb{E}[\nabla^2 A(\theta_0; X)]$ exists and is full rank, what is the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$?

Question 2 (Asymptotics and efficiency)**:** Consider the following classification problem, where we attempt to classify a vector $X \in \mathbb{R}^d$ as belonging to class $Y = 0$ or $Y = 1$. We have data drawn i.i.d. from a standard multivariate normal model, where

$$X \mid Y = 0 \sim \mathsf{N}\left(-\frac{1}{2}\theta_0, I_{d\times d}\right) \quad \text{and} \quad X \mid Y = 1 \sim \mathsf{N}\left(\frac{1}{2}\theta_0, I_{d\times d}\right), \tag{2}$$

where $I = I_{d\times d}$ is the $d \times d$ identity matrix and $\theta_0 \in \mathbb{R}^d$ is the unknown mean vector for each of the classes. We assume that $\mathbb{P}(Y = 1) = \frac{1}{2}$ and $\mathbb{P}(Y = 0) = \frac{1}{2}$. Assume you are given an i.i.d. sample $S := \{(X_i, Y_i)\}_{i=1}^n$ from the model (2), and would like to estimate θ_0.

(a) In order to estimate θ_0, the first obvious strategy is, given the sample S, to use the Gaussianity and estimate

$$\widehat{\theta}_{\mathrm{G},n} := \frac{1}{N_1} \sum_{i:Y_i=1} X_i - \frac{1}{N_0} \sum_{i:Y_i=0} X_i,$$

where $N_y = \mathrm{card}\{i \in [n] : Y_i = y\}$. Show that

$$\sqrt{n}(\widehat{\theta}_{\mathrm{G},n} - \theta_0) \xrightarrow{d} \mathsf{N}(0, \Sigma_{\mathrm{G}})$$

and specify Σ_{G}.

(b) Under the model (2), give the conditional distribution

$$p_\theta(y \mid x) = \mathbb{P}(Y = y \mid X = x; \theta).$$

(c) Because you like classification, you decide to fit a logistic regression model, defining the logistic loss for a pair (x, y) by

$$\ell(\theta; x, y) = \log(1 + \exp(x^T\theta)) - yx^T\theta.$$

Define the empirical risk and estimator

$$\widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i, Y_i) \quad \text{and} \quad \widehat{\theta}_{\mathrm{LR},n} := \operatorname*{argmin}_\theta \widehat{R}_n(\theta).$$

Show that

$$\sqrt{n}(\widehat{\theta}_{\mathrm{LR},n} - \theta_0) \xrightarrow{d} \mathsf{N}(0, \Sigma_{\mathrm{LR}})$$

and specify Σ_{LR}. [*Hint:* Use Question 1.]

(d) Show that if we take θ_0 far from 0, that is, $\|\theta_0\| \to \infty$, we have $\mathrm{tr}(\Sigma_{\mathrm{LR}}) \to \infty$. Does $\mathrm{tr}(\Sigma_{\mathrm{G}})$ have a limit as $\|\theta_0\| \to \infty$?

(e) Now consider the actual classification (log probability) risk, $R(\theta) = \mathbb{E}[\ell(\theta; X, Y)]$. Show that

$$n(R(\widehat{\theta}_{\mathrm{G},n}) - R(\theta_0)) \xrightarrow{d} W_{\mathrm{G}} \quad \text{and} \quad n(R(\widehat{\theta}_{\mathrm{LR},n}) - R(\theta_0)) \xrightarrow{d} W_{\mathrm{LR}}$$

for random variables W_{G} and W_{LR}, and specify their distributions. What is $\sup_{\theta_0} \frac{\mathbb{E}[W_{\mathrm{LR}}]}{\mathbb{E}[W_{\mathrm{G}}]}$? Is $\inf_{\theta_0} \frac{\mathbb{E}[W_{\mathrm{LR}}]}{\mathbb{E}[W_{\mathrm{G}}]} > 0$?

(f) Suppose that instead of a Gaussian model, the data X actually follows a different exponential family model, where X has densities determined by $\theta = (\theta_0, \theta_1)$ with

$$p_\theta(x \mid y) = \exp\left(\theta_y^T x - A(\theta_y)\right). \tag{3}$$

Show that the p.m.f. of $Y \in \{0, 1\}$ conditional on $X = x$ in this model has the sigmoid form

$$p(y \mid x) = \frac{e^{y\beta^T \phi(x)}}{1 + e^{\beta^T \phi(x)}}$$

for some function $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ and some $\beta \in \mathbb{R}^{d+1}$. Specify ϕ, and specify β in terms of θ and the prior probabilities $\pi(y) = \mathbb{P}(Y = y)$. Given that your data may come from model (3), which estimator $\widehat{\theta}_{\mathrm{G},n}$ or $\widehat{\theta}_{\mathrm{LR},n}$ do you prefer?

Question 3 (Moduli of continuity and high probability rates of convergence)**:** 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

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

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

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

where $\beta > 0$ is some constant. We also assume that the instantaneous functions $F(\cdot; x)$ are L-Lipschitz with respect to the norm $\|\cdot\|$.

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

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

(a) Show that 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 $\|\theta' - \theta_0\| > r$.

(b) Show that f and f_n are convex.

(c) Show that θ_0 is unique.

(d) Let

$$\Delta(\theta, x) := [F(\theta; x) - f(\theta)] - [F(\theta_0; x) - f(\theta_0)].$$

Show that $\Delta(\theta, X)$ (i.e. the random version where $X \sim P$) is $4L^2 \|\theta - \theta_0\|^2$-sub-Gaussian.

(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)

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \sigma \cdot \frac{1+t}{\sqrt{n}}\right) \leq C \exp\left(-t^2\right)$$

for all $t \geq 0$, where $C < \infty$ is a numerical constant. [*Hint:* The quantity $\Delta_n(\theta) := \frac{1}{n} \sum_{i=1}^n \Delta(\theta, X_i)$ may be helpful, as may be the bounded differences inequality from HW6.]

Question 4 (Non-asymptotic lower bounds on estimation)**:** We wish to estimate the mean θ from some distribution P belonging to a set $\mathcal{P}$ of distributions (each with at least two moments). Let $\theta(P) = \mathbb{E}_P[X]$ denote the mean of the distribution P, and consider the minimax absolute risk

$$\mathfrak{M}_n := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \widehat{\theta}(X_1, \ldots, X_n) - \theta(P) \right| \right],$$

where the expectation is taken over the n i.i.d. observations $X_i \stackrel{\text{iid}}{\sim} P$, and the infimum $\widehat{\theta}$ is taken over all estimators.

(a) Show that for any distributions $P_0, P_1 \in \mathcal{P}$, with $\theta_i = \theta(P_i)$ denoting their means, we have

$$\mathfrak{M}_n \geq \frac{1}{4} |\theta_0 - \theta_1| \left(1 - \|P_0^n - P_1^n\|_{\text{TV}}\right),$$

where P^n denotes the probability distribution of n i.i.d. observations from P.

(b) Let $\beta \in (0, \infty)$. Suppose that $\mathcal{P}$ is the family of exponential distributions with scale (also mean) $\theta \in [\beta, 2\beta]$, is, $\mathcal{P} = \{\mathsf{Exp}(\theta)\}_{\beta \leq \theta \leq 2\beta}$, where we recall that $X \sim \mathsf{Exp}(\theta)$ if X has density $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ for $x \geq 0$ (and density $f(x) = 0$ otherwise). As a reminder, $\mathbb{E}_\theta[X] = \theta$ and $\text{Var}_\theta(X) = \theta^2$. Using the result of part (a), show that

$$\frac{2\beta}{\sqrt{n}} \geq \mathfrak{M}_n \geq c \frac{\beta}{\sqrt{n}},$$

where $c > 0$ is a numerical constant.