Statistics 300B Winter 2019 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 0.1 (Distributed minimization and inference)**:** Consider a loss minimization problem with loss $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$, where we wish to minimize the population loss (risk)

$$L(\theta) := \mathbb{E}[\ell(\theta, Z)].$$

Assume that $\ell(\theta, z)$ is convex in θ for all $z \in \mathcal{Z}$ and that its first and second derivatives $\nabla_\theta \ell(\theta, z)$ and $\nabla^2 \ell(\theta, z)$ are $M_1(z)$ and $M_2(z)$-Lipschitz, respectively, where $\mathbb{E}[M_i(Z)^2] < \infty$.[1] At the point $\theta^\star = \operatorname{argmin}_\theta L(\theta)$, we have $\nabla^2 L(\theta^\star) \succ 0$, that is, the Hessian is positive definite.

You are given a large sample $Z_1, \ldots, Z_N$, split into equal-sized batches $\{B_k\}_{k=1}^m$ of size $n = N/m$ across m different computers (machines), where $\cup_k B_k = \{1, \ldots, N\}$. To save computation and communication you decide to construct an estimator $\overline{\theta}_N$ of $\theta^\star$ by aggregating independently computed local minimizers, for each k defining

$$\widehat{\theta}^k := \operatorname*{argmin}_\theta \widehat{L}_k(\theta), \quad \widehat{L}_k(\theta) = \frac{1}{|B_k|} \sum_{i \in B_k} \ell(\theta, Z_i)$$

and then setting $\overline{\theta}_N = \frac{1}{m} \sum_{k=1}^m \widehat{\theta}^k$. You may assume that m is fixed for this problem.

(a) Give the asymptotic distributions of

$$\sqrt{n} \begin{bmatrix} \widehat{\theta}^1 - \theta^\star \\ \vdots \\ \widehat{\theta}^m - \theta^\star \end{bmatrix} \quad \text{and} \quad \sqrt{N}(\overline{\theta}_N - \theta^\star).$$

(b) You decide you would like to use your subsampled estimators and $\overline{\theta}_N$ to construct confidence intervals for various functions of $\theta^\star$. For a fixed $v \in \mathbb{R}^d$, define $\sigma_v^2 := \frac{1}{m-1} \sum_{k=1}^m (v^T(\widehat{\theta}^k - \overline{\theta}_N))^2$. Give the asymptotic distribution of

$$\frac{v^T(\overline{\theta}_N - \theta^\star)}{\sigma_v}.$$

For $\alpha \in (0, 1)$, give a confidence interval $C_{N,v,\alpha}$ so that

$$\mathbb{P}\left(v^T \theta^\star \in C_{N,v,\alpha}\right) \to 1 - \alpha.$$

(*Hint:* If $W_1, \ldots, W_m \stackrel{\text{iid}}{\sim} \mathsf{N}(0, 1)$ and $P_m W = \frac{1}{m} \sum_{i=1}^m W_i$, what is the distribution of the normalized quantity $\sqrt{m} P_m W / \sqrt{\frac{1}{m-1} \sum_{i=1}^m (W_i - P_m W)^2}$?)

(c) Suppose that $m < d$, the dimension of the problem. For your confidence sets $C_{N,v,\alpha}$, compute

$$\liminf_{N \to \infty} \mathbb{P}(v^T \theta^\star \in C_{N,v,\alpha} \text{ for all } v).$$

[1] You may assume more moments if you wish; this is not important

Question 0.2 (Symmetrization, Lipschitz functions, and a multiclass ULLN)**:** Consider a multiclass classification problem with data pairs $(X, Y) \in \mathcal{X} \times [k]$, where $[k] = \{1, \ldots, k\}$. Let $L : \mathbb{R}^k \times [k] \to \mathbb{R}_+$ be a loss function, so that for a prediction $f(x) \in \mathbb{R}^k$ with true label y, we suffer loss $L(f(X), Y)$. For example, we might use the logistic loss

$$L_{\log}(\alpha, y) = \log\left(\sum_{i=1}^{k} \exp(\alpha_i - \alpha_y)\right).$$

The idea is that if $\alpha_y \gg \alpha_i$ for $i \neq y$, then $L(\alpha, y) \approx 0$. We assume throughout this question that

$$L(\mathbf{0}, y) = L(\mathbf{0}, y') =: L(\mathbf{0}) \text{ for all } y, y',$$

that is, the all-zeros prediction suffers constant loss. We assume that $L(\cdot, y)$ is 1-Lipschitz w.r.t. the ℓ_2-norm, which is the case for $L_{\log}$, that is, $|L(\alpha, y) - L(\beta, y)| \le \|\alpha - \beta\|_2$ for $\alpha, \beta \in \mathbb{R}^k$.

(a) Let $\mathcal{F}$ be a collection of functions mapping $\mathcal{X} \to \mathbb{R}$. Show that for any fixed $\{x_i, y_i\}_{i=1}^n$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \varepsilon_i L(f(x_i), y_i)\right|\right] \le \mathbb{E}\left[\left|\sum_{i=1}^{n} \varepsilon_i L(f_0(x_i), y_i)\right|\right] + C\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \langle w_i, f(x_i) - f_0(x_i)\rangle\right|\right],$$

where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{Uni}\{\pm 1\}$, $w_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_k)$, and f_0 is an arbitrary (fixed) element of $\mathcal{F}$.

(b) Let $\mathcal{F}$ be the function class consisting of

$$f_\theta(x) = (\langle \theta_1, x\rangle,\ \langle \theta_2, x\rangle,\ \ldots,\ \langle \theta_k, x\rangle) \in \mathbb{R}^k \tag{1}$$

where the vectors θ_i each belong to the ℓ_2-ball of radius r. That is, $\|\theta_i\|_2 \le r$ for $i = 1, \ldots, k$. Show that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \varepsilon_i L(f(x_i), y_i)\right|\right] \le \sqrt{n} L(\mathbf{0}) + Ckr\sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}.$$

(c) *Warning:* This is probably the most challenging question on the exam. If the number of classes k is very large, we may wish for somewhat sparser predictive sets. Suppose the functions f are linear (i.e. of the form (1)), but the θ_i belong to the set

$$\Theta_{\ell_1/\ell_2} := \left\{\theta_1, \ldots, \theta_k \in \mathbb{R}^d \mid \sum_{i=1}^{k} \|\theta_i\|_2 \le r\right\}.$$

Show that for $\mathcal{F} = \{f_\theta \mid \theta \in \Theta_{\ell_1/\ell_2}\}$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \varepsilon_i L(f(x_i), y_i)\right|\right] \le \sqrt{n} L(\mathbf{0}) + Cr\sqrt{\mathrm{tr}(X^T X)(2\log k + 1)},$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix with rows x_i^T. Note that $\mathrm{tr}(X^T X) = \sum_{i=1}^n \|x_i\|_2^2$.

Hint: In our solution it was useful to compute $\mathbb{E}[\exp(\lambda \|Xw\|_2^2)]$ for $w \sim \mathsf{N}(0, I_n)$.

(d) Suppose we receive data in pairs $(X_i, Y_i) \stackrel{\text{iid}}{\sim} P$, where $\mathbb{E}[\|X_i\|_2^2] \le b^2$ for all i, that $k \ge 2$, and that we use the logistic loss $L_{\log}$. Prove that for the same function class $\mathcal{F}$ as in part (c)

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|P_n L_{\log}(f(X), Y) - P L_{\log}(f(X), Y)\right|\right] \le C\left[\frac{\log k}{\sqrt{n}} + \frac{\sqrt{r^2 b^2 \log k}}{\sqrt{n}}\right].$$

Question 0.3 (Growth functions for non-convex problems)**:** In the phase retrieval problem[2], we wish to recover a signal $\theta^\star \in \mathbb{R}^n$ based on (noisy) observations of the magnitudes of its inner products $\langle X_i, \theta \rangle$ with a set of n vectors $X_1, \ldots, X_n$. In physical detectors, we observe a number of photons $Y_i \in \mathbb{N}$ that scale (roughly) with $\langle X_i, \theta^\star \rangle^2$, where in fact, the distribution of Y_i given $\langle X_i, \theta^\star \rangle^2$ is

$$Y_i \sim \mathsf{Poisson}(\langle X_i, \theta^\star \rangle^2).$$

Recall that $Y \sim \mathsf{Poisson}(\lambda)$ if the p.m.f. of Y is

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

Let us consider the (conditional) expectation of the log loss of our measurements, that is, define

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

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

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

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

(b) Now we consider the actual (conditional) losses for the parameter θ. Show that

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

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

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

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

$$\sqrt{\frac{d + \log \frac{1}{\delta}}{n}} \leq c_0$$

then with probability at least $1 - \delta$, simultaneously for all $\theta \in \mathbb{R}^d$,

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

where $\mathrm{dist}(\theta, \theta^\star) = \min_{s \in \{\pm 1\}} \|\theta + s\theta^\star\|_2$ is the distance (ignoring sign) between θ and $\theta^\star$.

[2]Technically, we do not work in the complex plane in this problem so we are actually addressing the sign retrieval problem.

Question 0.4 (Rates of convergence for sparse estimators without subgaussian noise)**:** Assume we have the sparse linear model

$$Y = X\theta^\star + \varepsilon$$

$X \in \mathbb{R}^{n\times d}$, where ε_i are independent, mean-zero, and satisfy

$$\mathbb{E}[|\varepsilon_i|^q] \le \sigma^q$$

for some $2 \le q < \infty$. Assume that the data matrix

$$X = [x_1 \ \cdots \ x_d] \in \mathbb{R}^{n\times d}$$

has normalized columns x_j satisfying $\frac{1}{n}\sum_{i=1}^n |x_{ji}|^q = 1$, and that it satisfies the restricted strong convexity condition

$$\frac{1}{n}\|X\Delta\|_2^2 \ge \mu\|\Delta\|_2^2 \quad \text{for all } \Delta \in \mathbb{C}_3(S) := \left\{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1\right\}.$$

Here $S \subset [d]$ denotes the support $S = \{j \in [d] : \theta_j^\star \ne 0\}$, and $|S| \le k \ll d$. Let $\widehat{\theta}_n$ be the Lasso estimator

$$\widehat{\theta}_n := \operatorname*{argmin}_\theta \left\{\frac{1}{2n}\|X\theta - Y\|_2^2 + \lambda_n\|\theta\|_1\right\}.$$

Let $\delta > 0$.

(a) Give a choice of λ_n and the tightest quantity

$$r_n = r_n(\delta, \mu, q, \sigma, n, k, d)$$

(that is, as a function of the probability δ, restricted strong convexity constant μ, moments σ and q of the noise, and dimension/sample sizes n, k, d) and you can such that

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta^\star\|_2 \ge r_n\right) \le \delta.$$

(b) How large must the moment q be so that you recover the standard Lasso guarantee (when ε_i are i.i.d. σ^2-sub-Gaussian) of

$$\|\widehat{\theta}_n - \theta^\star\|_2 \le \frac{c\sigma\sqrt{k\log\frac{d}{\delta}}}{\sqrt{n}}$$

in your rate r_n?

Question 0.5: Did you take a 5 minute walk during this exam to make sure to stretch your legs?