Problem Set 4: Minimax lower bounds

Stats 311/EE 377

Due: Thursday, March 17 (last day of quarter)

Question 4.1: In this question, we will show that the minimax rate of estimation for the parameter of a uniform distribution (in squared error) scales as $1/n^2$. In particular, assume that $X_i \stackrel{\text{i.i.d.}}{\sim} \mathsf{Uni}(\theta, \theta + 1)$, meaning that X_i have densities $p(x) = 1_{(x \in [\theta, \theta+1])}$. Let $X_{(1)} = \min_i\{X_i\}$ denote the first order statistic.

(a) Prove that

$$\mathbb{E}[(X_{(1)} - \theta)^2] = \frac{2}{(n+1)(n+2)}.$$

(Hint: the fact that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)dt$ for any positive Z may be useful.)

(b) Using Le Cam's two-point method, show that the minimax rate for estimation of $\theta \in \mathbb{R}$ for the uniform family $\mathcal{U} = \{\mathsf{Uni}(\theta, \theta + 1) : \theta \in \mathbb{R}\}$ in squared error has lower bound c/n^2, where c is a numerical constant.

Question 4.2: In this question, we explore estimation under a constraint known as differential privacy. In one version of private estimation, the collector of data is not trusted, so instead of seeing true data $X_i \in \mathcal{X}$ only a disguised version $Z_i \in \mathcal{Z}$ is viewed, where given $X = x$, we have $Z \sim Q(\cdot \mid X = x)$. We say that this Z_i is *α-differentially private* if for any subset $A \subset \mathcal{Z}$ and any pair $x, x' \in \mathcal{X}$,

$$\frac{Q(Z \in A \mid X = x)}{Q(Z \in A \mid X = x')} \leq \exp(\alpha). \tag{1}$$

The intuition here, from a privacy standpoint, is that no matter what the true data X is, any points x and x' are essentially equally likely to have generated the observed signal Z. We explore a few consequences of differential privacy in this question, including so-called quantitative data processing inequalities. We assume that $\alpha < 1$ for simplicity.

First, we show how differential privacy acts as a contraction on probability distributions. Let P_1 and P_2 be arbitrary distributions on $\mathcal{X}$ (with densities p_1 and p_2 w.r.t. a base measure μ) and define the *marginal* distributions

$$M_i(Z \in A) := \int_{\mathcal{X}} Q(Z \in A \mid X = x)p_i(x)d\mu(x), \quad i \in \{1, 2\}.$$

We will prove that there is a universal (numerical) constant $C < \infty$ such that for any P_1, P_2,

$$D_{\text{kl}}(M_1 \| M_2) + D_{\text{kl}}(M_2 \| M_1) \leq C(e^\alpha - 1)^2 \|P_1 - P_2\|_{\text{TV}}^2. \tag{2}$$

(a) Show that for any $a, b > 0$

$$\left|\log \frac{a}{b}\right| \le \frac{|a-b|}{\min\{a,b\}}.$$

(b) As discussed in HW 1, when considering $D_{\rm kl}(M_1 \| M_2)$, it is no loss of generality to assume that $\mathcal{Z} = \{1, \ldots, k\}$ for some finite k. Use the shorthands $q(z \mid x) = Q(Z = z \mid X = x)$ and $m_i(z) = \int q(z \mid x) p_i(x) d\mu(x)$. Show that there exists a universal constant $c < \infty$ such that

$$|m_1(z) - m_2(z)| \le c(e^\alpha - 1) \inf_{x \in \mathcal{X}} q(z \mid x) \left\|P_1 - P_2\right\|_{\rm TV}.$$

(c) Combining parts (a) and (b), show inequality (2).

We note in passing that, except for perhaps the constant factor C, inequality (2) cannot be improved generally. This can be shown by letting P_1 and P_2 be Bernoulli distributions, taking $\|P_1 - P_2\|_{\rm TV} \to 0$, and choosing a Bernoulli distribution for Q while taking $\alpha \to 0$. You do not need to prove this.

Question 4.3: In this question, we apply the results of Question 4.2 to a problem of estimation of drug use. Assume we interview a series of individuals $i = 1, \ldots, n$, asking each whether he or she takes illicit drugs. Let $X_i \in \{0, 1\}$ be 1 if person i uses drugs, 0 otherwise, and define $\theta^* = \mathbb{E}[X] = \mathbb{E}[X_i] = P(X = 1)$. To avoid answer bias, each answer X_i is perturbed by some channel Q, where Q is α-differentially private (recall definition (1)). That is, we observe independent Z_i where conditional on X_i, we have

$$Z_i \mid X_i = x \sim Q(\cdot \mid X_i = x).$$

To make sure everyone feels suitably private, we assume $\alpha < 1/2$ (so that $(e^\alpha - 1)^2 \le 2\alpha^2$). In the questions, let $\mathcal{Q}_\alpha$ denote the family of all α-differentially private channels, and let $\mathcal{P}$ denote the Bernoulli distributions with parameter $\theta(P) = P(X_i = 1) \in [0, 1]$ for $P \in \mathcal{P}$.

(a) Use Le Cam's method and the strong data processing inequality (2) to show that the minimax rate for estimation of the proportion θ^* in absolute value satisfies

$$\mathfrak{M}_n(\theta(\mathcal{P}), |\cdot|, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[|\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)|\right] \ge c\frac{1}{\sqrt{n\alpha^2}},$$

where $c > 0$ is a universal constant. Here the infimum is over channels Q and estimators $\widehat{\theta}$, and the expectation is taken with respect to both the X_i (according to P) and the Z_i (according to $Q(\cdot \mid X_i)$).

(b) Give a rate-optimal estimator for this problem. That is, define a channel Q that is α-differentially private and an estimator $\widehat{\theta}$ such that $\mathbb{E}[|\widehat{\theta}(Z_1^n) - \theta|] \le C/\sqrt{n\alpha^2}$, where $C > 0$ is a universal constant.

(c) Let $\mathcal{P}_k$, for $k \ge 2$, denote the family of distributions on $\mathbb{R}$ such that $\mathbb{E}_P|X|^k \le 1$ for $P \in \mathcal{P}_k$ (note that X is no longer restricted to have support $\{0, 1\}$). Show, similarly to part (a), that for $\theta(P) = \mathbb{E}_P[X]$

$$\mathfrak{M}_n(\theta(\mathcal{P}_k), |\cdot|, \alpha) := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}_k} \mathbb{E}\left[|\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)|\right] \ge c\frac{1}{(n\alpha^2)^{\frac{k-1}{2k}}}.$$

What does this say about $k = 2$?

(d) Download the dataset at `http://web.stanford.edu/class/stats311/Data/drugs.txt`, which consists of a sample of 100,000 hospital admissions and whether the patient was abusing drugs (a 1 indicates abuse, 0 no abuse). Use your estimator from part (b) to estimate the population proportion of drug abusers: give an estimated number of users for $\alpha \in \{2^{-k}, k = 0, 1, \ldots, 10\}$. Perform each experiment several times. Assuming that the proportion of users in the dataset is the true population proportion, how accurate is your estimator?

Question 4.4: In this question, we study the question of whether adaptivity can give better estimation performance for linear regression problems. That is, for $i = 1, \ldots, n$, assume that we observe variables Y_i in the usual linear regression setup,

$$Y_i = \langle X_i, \theta \rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2), \tag{3}$$

where $\theta \in \mathbb{R}^d$ is unknown. But now, based on observing $Y_1^{i-1} = \{Y_1, \ldots, Y_{i-1}\}$, we allow an adaptive choice of the next predictor variables $X_i \in \mathbb{R}^d$. Let $\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2)$ denote the family of linear regression problems under this adaptive setting (with n observations) where we constrain the Frobenius norm of the data matrix $X^\top = [X_1 \ \cdots \ X_n]$, $X \in \mathbb{R}^{n \times d}$, to have bound $\|X\|_{\text{Fr}}^2 = \sum_{i=1}^n \|X_i\|_2^2 \le \mathsf{F}^2$. We use Assouad's method to show that the minimax mean-squared error satisfies the following bound:

$$\mathfrak{M}(\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2), \|\cdot\|_2^2) := \inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \ge \frac{d\sigma^2}{n} \cdot \frac{1}{16 \frac{1}{dn} \mathsf{F}^2}. \tag{4}$$

Here the infimum is taken over all adaptive procedures satisfying $\|X\|_{\text{Fr}}^2 \le \mathsf{F}^2$.

In general, when we choose X_i based on the observations Y_1^{i-1}, we are taking $X_i = F_i(Y_1^{i-1}, U_1^i)$, where U_i is a random variable independent of ε_i and Y_1^{i-1} and F_i is some function. Justify the following steps in the proof of inequality (4):

(i) Assume that nature chooses $v \in \mathcal{V} = \{-1, 1\}^d$ uniformly at random and, conditionally on v, let $\theta = \theta_v$. Justify

$$\mathfrak{M}(\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2), \|\cdot\|_2^2) \ge \inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2].$$

Argue it is no loss of generality to assume that the choices for X_i are deterministic based on the Y_1^{i-1}. Thus, throughout we assume that $X_i = F_i(Y_1^{i-1}, u_1^i)$, where u_1^n is a fixed sequence, or, for simplicity, that X_i is a function of Y_1^{i-1}.

(ii) Fix $\delta > 0$. Let $v \in \{-1, 1\}^d$, and for each such v, define $\theta_v = \delta v$. Also let P_v^n denote the joint distribution (over all adaptively chosen X_i) of the observed variables $Y_1, \ldots, Y_n$, and define $P_{+j}^n = \frac{1}{2^{d-1}} \sum_{v: v_j = 1} P_v^n$ and $P_{-j}^n = \frac{1}{2^{d-1}} \sum_{v: v_j = -1} P_v^n$, so that $P_{\pm j}^n$ denotes the distribution of the Y_i when $v \in \{-1, 1\}^d$ is chosen uniformly at random but conditioned on $v_j = \pm 1$. Then

$$\inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2] \ge \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right].$$

(iii) We have

$$\frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right] \ge \frac{\delta^2 d}{2} \left[1 - \left(\frac{1}{d} \sum_{j=1}^d \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}^2\right)^{\frac{1}{2}}\right].$$

(iv) Let $P_{+j}^{(i)}$ be the distribution of the random variable Y_i conditioned on $v_j = +1$ (with the other coordinates of v chosen uniformly at random), and let $P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)$ denote the distribution of Y_i conditioned on $v_j = +1$, $Y_1^{i-1} = y_1^{i-1}$, and x_i. Justify

$$\begin{aligned}\left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}^2 &\le \frac{1}{2} D_{\mathrm{kl}}\left(P_{+j}^n \| P_{-j}^n\right) \\ &\le \frac{1}{2}\sum_{i=1}^n \int D_{\mathrm{kl}}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) dP_{+j}^{i-1}(y_1^{i-1}, x_i).\end{aligned}$$

(v) Then we have

$$\sum_{j=1}^d D_{\mathrm{kl}}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) \le \frac{2\delta^2}{\sigma^2} \left\|x_i\right\|_2^2.$$

(vi) We have

$$\sum_{j=1}^d \left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}^2 \le \frac{\delta^2}{\sigma^2} \mathbb{E}[\|X\|_{\mathrm{Fr}}^2],$$

where the final expectation is over V drawn uniformly in $\{-1, 1\}^d$ and all Y_i, X_i.

(vii) Show how to choose δ appropriately to conclude the minimax bound (4).

Question 4.5: Suppose under the setting of Question 4.4 that we may no longer be adaptive, meaning that the matrix $X \in \mathbb{R}^{n \times d}$ must be chosen ahead of time (without seeing any data). Assuming $n \ge d$, is it possible to attain (within a constant factor) the risk (4)? If so, give an example construction, if not, explain why not.