Advanced Topics in Artificial Intelligence: EE6180
Indian Institute of Technology Madras October 15, 2019
Instructor: Abhishek Sinha

Problem Set 4

- This problem set is due on **October 30, 2019** in the class.
- Each problem carries 10 points.
- You may work on the problems in groups of size at most **two**. However, **each student must write their own solution**. If you collaborate on the problems, clearly mention the name of your collaborator.

1. **(Minimax Lower Bound for the Uniform Location Family)** In this problem, we will show that the minimax rate of estimation for the parameter of a uniform distribution (in squared error) scales as $\frac{1}{n^2}$. In particular, assume that $X_i \overset{\text{i.i.d.}}{\sim} \mathsf{Uni}(\theta, \theta + 1)$. Let $X_{(1)} = \min_i\{X_i\}$ denote the first order statistic.

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

   (***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 $\frac{c}{n^2}$, where c is a numerical constant.

2. **(Regret Lower Bound for Parametric Classification)** In this problem, we will derive lower bounds for parametric classification problems. In the classification problem, the risk function is the expectation of the zero-one loss over the joint distribution. That is, if P_{XY} is the joint distribution, then the risk $R(f) = \mathbb{E}_{(X,Y)\sim P_{XY}} \mathbb{1}(f(X) \neq Y)$. We will establish minimax lower bounds on the *excess risk*, which is the risk minus the risk of the Bayes optimal estimator, which is the function minimizing the risk.

   ($\boldsymbol{a}$) Consider a classification setting where $X \in [0, 1]$ and $\mathbb{P}(Y = 1|X = x) = \eta(x)$ is the regression function. For a distribution P over X, Y with regression function $\eta(x)$, the Bayes optimal predictor is $f_B(x) = \mathbb{1}(\eta(x) \geq \frac{1}{2})$. Let $\mathcal{P}$ be the class of all distributions for which the Bayes optimal predictor is of the form $f_B(x) = \mathbb{1}(x \leq t)$. This means that the regression function $\eta(\cdot)$ crosses $\frac{1}{2}$ at most once at some point t (and it is above half on $[0, t)$). For such a distribution, the optimal estimator is the function $\mathbb{1}(x \leq t)$ and we say that its risk is R^*. Show that
   $$\inf_{\hat{f}_n} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{(X_i,Y_i)\sim P_{XY}}[R(\hat{f}_n) - R^*] \geq C\sqrt{\frac{1}{n}}.$$

   Hint: You can use Le Cam's method. It is convenient to choose the two distributions to have the same marginal $P_X = \mathsf{Uniform}[0, 1]$ and vary the kernel $P_{Y|X}(\cdot|\cdot)$.

(***b***) Consider now the same set up except where the data $X \in [0,1]^d$ and the class $\mathcal{P}$ corresponds to distributions with Bayes optimal predictors of the form $f_B(x) = \prod_{j=1}^d \mathbb{1}(x_j \le t_j)$. Use Fano's method to show that:

$$\inf_{\hat{f}_n} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{(X_i,Y_i)\sim P_{XY}}[R(\hat{f}_n) - R^*] \ge C\sqrt{\frac{d}{n}}.$$

3. **(KL Divergence and Differential Privacy)** In this problem, we explore estimation under a constraint known as differential privacy. The conclusion from this problem will be used in the next problem on detecting drug abuse with private data. 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 $Z \sim Q(\cdot|X = x)$. We say that this Z_i is differentially private if for any subset $A \subset \mathcal{Z}$ and any pair $x, x' \in \mathcal{X}$,

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

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

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

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

$$D(M_1||M_2) + D(M_2||M_1) \le C(e^\alpha - 1)^2||P_1 - P_2||^2. \tag{2}$$

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

$$|\ln\frac{a}{b}| \le \frac{|a-b|}{\min\{a,b\}}.$$

(***b***) Use the shorthands $q(z|x) = Q(Z = z|X = x)$ and $m_i(z) = \int q(z|x)p_i(x)dx$. Show that there exists a universal constant $c < \infty$ such that

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

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

4. **(Application of Le Cam's Method to Detecting Drug Abuse)** In this problem, we apply the results of the previous exercise to a problem of estimation of drug abuse. Assume we interview a series of individuals $i = 1, 2, \ldots, n$, 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 the definition in Eqn. (1)). That is, we observe independent Z_i where conditional on X_i, we have

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

To make sure everyone feels suitably private, we assume $\alpha < \frac{1}{2}$ (so that $(e^\alpha - 1)^2 \leq 2\alpha^2$). 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 $\theta(P) = \mathbb{P}(X_i = 1) \in [0, 1]$ for $P \in \mathcal{P}$.

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

$$\mathcal{M}_n := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}(|\hat{\theta}(Z_1, Z_2, \ldots, Z_n) - \theta(P)|) \geq c \frac{1}{\sqrt{n\alpha^2}}.$$

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

(c) 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 $\alpha \in \{2^{-k}, k = 1, 2, \ldots, 10\}$. Perform each experiment several times. Assuming that the proportion of users in the dataset is the true population proportion, how accurate is your estimator?

5. **(Fundamental Limits of Sign Identification in Sparse Linear Regression)** In sparse linear regression, we have n observations $Y_i = \langle X_i, \theta^* \rangle + \epsilon_i$, where $X_i \in \mathbb{R}^d$ are known (fixed) matrices and the vector θ^* has a small number $k \ll d$ of non-zero entries, and $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$. In this problem, we investigate the problem of *sign recovery*, that is, identifying the vector of signs $\mathsf{sign}(\theta_j^*), \forall j$, where $\mathsf{sign}(0) = 0$.
Assume we have the following process: fix a signal threshold $\theta_{\min} > 0$. First, a vector $S \in \{-1, 0, +1\}^d$ is chosen uniformly at random from the set of vectors $\mathcal{S}_k \equiv \{s \in \{-1, 0, +1\}^d : ||s||_1 = k\}, k \geq 2$. Then we define the vectors θ^s so that $\theta_j^s = \theta_{\min} s_j$, and conditional on $S = s$, we observe

$$Y = X\theta^s + \epsilon, \quad \epsilon \sim \mathsf{N}(0, \sigma^2 I_{n \times n}).$$

(Here $X \in \mathbb{R}^{n \times d}$ is a known fixed matrix.)

(a) Use Fano's inequality to show that for any estimator $\hat{S}$ of S, we have

$$\mathbb{P}(\hat{S} \neq S) \geq \frac{1}{2} \quad \text{unless} \quad n \geq \frac{\frac{d}{k} \ln \binom{d}{k}}{||n^{-1/2} X||_{\mathsf{Fr}}^2} \frac{\sigma^2}{\theta_{\min}^2}.$$

(b) Assume that $X \in \{-1, +1\}^{n \times d}$. Give a lower bound on how large n must be for sign recovery. Give a one line interpretation of the quantity $\frac{\theta_{\min}^2}{\sigma^2}$.