Problem Set 2: Minimax lower bounds

Stats 311/EE 377

Due: Thursday, 10/23/2014

Question 1 (Divergence between multivariate normal distributions): Let $P_1$ be $N(\theta_1, \Sigma)$ and $P_2$ be $N(\theta_2, \Sigma)$, where $\Sigma \succ 0$ is a positive definite matrix. What is $D_{\text{kl}}(P_1 \| P_2)$?

Question 2: 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 $X_i \overset{\text{i.i.d.}}{\sim} \text{Uni}(\theta, \theta + 1)$, 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
$$E[(X_{(1)} - \theta)^2] = \frac{2}{(n+1)(n+2)}.$$ (Hint: the fact that $E[Z] = \int_0^\infty P(Z \geq t)dt$ 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 $\mathcal{U} = \{\text{Uni}(\theta, \theta + 1) : \theta \in \mathbb{R}\}$ in squared error has lower bound $c/n^2$, where $c$ is a numerical constant.

Question 3: 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 $Z \sim Q(\cdot | X = x)$. We say that this $Z_i$ is $\alpha$-differentially private if for any subset $A \subset \mathcal{Z}$ and any pair $x, x' \in \mathcal{X}$,
$$\frac{Q(Z \in A | X = x)}{Q(Z \in A | X = x')} \leq \exp(\alpha).$$ (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 $\mu$) 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 $P_1, P_2$,
$$D_{\text{kl}}(M_1 \| M_2) + D_{\text{kl}}(M_2 \| M_1) \leq C(\exp(\alpha) - 1)^2 \|P_1 - P_2\|_{\text{TV}}^2.$$ (2)
(a) Show that for any $a, b > 0$

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

(b) As discussed in HW 1, when considering $D_{KL}(M_1 \| M_2)$, it is no loss of generality to assume that $Z = \{1, \ldots, k\}$ 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

$$|m_1(z) - m_2(z)| \leq c(e^a - 1) \inf_{x \in X} q(z \mid x) \|P_1 - P_2\|_{TV}.$$ 

(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 $\|P_1 - P_2\|_{TV} \to 0$, and choosing a Bernoulli distribution for $Q$ while taking $\alpha \to 0$. You do not need to prove this.

**Question 4:** In this question, we apply the results of Question 3 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 $\theta^* = E[X] = E[X_i] = P(X = 1)$. To avoid answer bias, each answer $X_i$ is perturbed by some channel $Q$, where $Q$ is $\alpha$-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 $(e^a - 1)^2 \leq 2\alpha^2$). In the questions, let $Q_\alpha$ denote the family of all $\alpha$-differentially private channels, and let $P$ denote the Bernoulli distributions with parameter $\theta(P) = P(X = 1) \in [0, 1]$ 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 $\theta^*$ in absolute value satisfies

$$\mathcal{M}_n(\theta(\mathcal{P}), \cdot, \cdot, \alpha) := \inf_{Q \in Q_\alpha} \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E\left[|\hat{\theta}(Z_1, \ldots, Z_n) - \theta(P)|\right] \geq \frac{1}{\sqrt{na^2}},$$

where $c > 0$ is a universal constant. Here the infimum is over channels $Q$ and estimators $\hat{\theta}$, 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 $\alpha$-differentially private and an estimator $\hat{\theta}$ such that $E[|\hat{\theta}(Z^n) - \theta|] \leq C/\sqrt{na^2}$, where $C > 0$ is a universal constant.

(c) Download the dataset at [http://web.stanford.edu/class/stats311/Data/drugs.txt](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 (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 5 (Sign identification in sparse linear regression): In sparse linear regression, we have \( n \) observations \( Y_i = \langle X_i, \theta^* \rangle + \varepsilon_i \), where \( X_i \in \mathbb{R}^d \) are known (fixed) matrices and the vector \( \theta^* \) has a small number \( k \ll d \) of non-zero indices, and \( \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \). In this problem, we investigate the problem of sign recovery, that is, identifying the vector of signs \( \text{sign}(\theta^*_j) \) for \( j = 1, \ldots, d \), where \( \text{sign}(0) = 0 \).

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 \( S_k := \{s \in \{-1, 0, 1\}^d : \|s\|_1 = k\} \). Then we define vectors \( \theta^*_s \) so that \( \theta^*_s = \theta_{\min}s_j \) and conditional on \( S = s \), we observe

\[ Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n}). \]

(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

\[ \Pr(\hat{S} \neq S) \geq \frac{1}{2} \quad \text{unless} \quad n \geq c \frac{d \log \binom{d}{k}}{\|n^{-1/2}X\|_F^2} \frac{\sigma^2}{\theta_{\min}^2}, \]

where \( c \) is a numerical constant. You may assume that \( k \geq 4 \) or \( \log \binom{d}{k} \geq 4 \log 2 \).

(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 sentence interpretation of \( \sigma^2/\theta_{\min}^2 \).