EE378C: Homework #3

Due on Wednesday, May 19, 2021

Please hand in your homework via Gradescope before 11:59 PM. Typing your solution using LaTeX is highly recommended.

1. The concept of orthogonal polynomials is useful not only in proving lower bounds, but also in constructing and analyzing unbiased estimators. Consider the same setting as Lecture 9: let $(P_\theta)_{\theta\in[\theta_0-\varepsilon,\theta_0+\varepsilon]}$ be a one-dimensional family of probability distributions with the following local expansion:

$$\frac{P_{\theta_0+u}(x)}{P_{\theta_0}(x)} = \sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{u^m}{m!}, \qquad \forall |u| \le \varepsilon, x \in \mathcal{X}.$$

In addition, we assume that the quantity $\sum_{x\in\mathcal{X}} P_{\theta_0+u}(x)P_{\theta_0+v}(x)/P_{\theta_0}(x)$ depends only on θ_0 and uv, for all $u, v \in [-\varepsilon, \varepsilon]$. Lecture 9 shows that $\{p_m(x;\theta_0)\}_{m\ge 0}$ are orthogonal in the sense that

$$\mathbb{E}_{X\sim P_{\theta_0}}[p_m(X;\theta_0)p_n(X;\theta_0)] = A_m(\theta_0)\cdot \mathbb{1}(m=n)$$

for some constants $\{A_m(\theta_0)\}_{m\ge 0}$.

(a) Show that for $X \sim P_{\theta_0+u}$, the estimator $p_m(X;\theta_0)$ is an unbiased estimator of u^m up to scaling. In other words, show that for all $u \in [-\varepsilon, \varepsilon]$,

$$\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)] = c_m u^m$$

for some constant c_m independent of u. Find the expression of c_m using $A_m(\theta_0)$.

(b) We can also say something about the second moment $\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)^2]$, with some additional properties of p_m. Suppose we could show that

$$p_m(x;\theta_0) = \sum_{\ell=0}^{m} b(m,\ell,\theta_0,u)\cdot p_\ell(x;\theta_0+u), \qquad \forall |u| \le \varepsilon, x \in \mathcal{X},$$

and we assume that both the local expansion and the orthogonality condition could be extended from θ_0 to any $\theta_0 + u$. Show that

$$\mathbb{E}_{X\sim P_{\theta_0+u}}[p_m(X;\theta_0)^2] = \sum_{\ell=0}^{m} b(m,\ell,\theta_0,u)^2\cdot A_\ell(\theta_0+u).$$

(c) Next we use a specific example to see how we could find the expression of $b(m,\ell,\theta_0,u)$. Suppose that $\mathcal{X} = \mathbb{N}, P_\theta = \mathsf{Poi}(\theta)$, so that for all $t \ge 0$, it holds that

$$\sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{\theta_0^m t^m}{m!} = \frac{P_{\theta_0(1+t)}(x)}{P_{\theta_0}(x)} = e^{-\theta_0 t}(1+t)^x = e^{ut}\cdot e^{-(\theta_0+u)t}(1+t)^x$$

$$= e^{ut} \cdot \frac{P_{(\theta_0+u)(1+t)}(x)}{P_{\theta_0+u}(x)} = e^{ut} \cdot \sum_{\ell=0}^{\infty} p_\ell(x; \theta_0+u) \frac{(\theta_0+u)^\ell t^\ell}{\ell!}.$$

Use this chain of equalities to argue that in the Poisson model, we have

$$b(m, \ell, \theta_0, u) = \binom{m}{\ell} \frac{(\theta_0+u)^\ell u^{m-\ell}}{\theta_0^m}.$$

2. This problem fills in the missing steps of the tree-based inequality in Lecture 10. Let $T = ([m], E)$ be an undirected graph with vertex set $[m]$ and edge set E, and be a tree in the sense that T is both connected and acyclic.

   (a) Show that for any real numbers $x_1, \cdots, x_m$, it holds that

   $$\sum_{i=1}^m x_i - \max_{i\in[m]} x_i \geq \sum_{(i,j)\in E} \min\{x_i, x_j\}.$$

   (b) Use the result in Part (a), show that for probability distributions $P_1, \cdots, P_m$,

   $$\min_\Psi \frac{1}{m} \sum_{i=1}^m P_i(\Psi \neq i) \geq \frac{1}{m} \sum_{(i,j)\in E} (1 - \|P_i - P_j\|_{\mathrm{TV}}),$$

   where the minimum is over all possible tests $\Psi : \mathcal{X} \to [m]$.

   (c) Evaluate the terms on both sides of (b) under $P_i = \mathcal{N}(i\Delta, 1)$ and a line tree with edge set $E = \{(1,2), (2,3), \cdots, (m-1, m)\}$, and show that they are equal.

3. In this problem, we aim to show that under the p-dimensional Gaussian location model $X \sim \mathcal{N}(\theta, I_p)$ with $\|\theta\|_0 \leq 1$, the minimax risk of denoising the 1-sparse vector in high dimensions satisfies

   $$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^p : \|\theta\|_0 \leq 1} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \geq (2 - o(1)) \log p, \qquad p \to \infty.$$

   In fact, the same upper bound (with constant $2 + o(1)$) could also be achieved.

   (a) We first establish an inequality which is similar to the generalized Fano's inequality, but involves the χ^2-divergence rather than the KL divergence in the definition of mutual information. Similar to Lecture 10, consider a generic prior π on Θ and loss function $L(\theta, a)$. Show that for every $\Delta > 0$,

   $$\inf_T \mathbb{E}_{\theta\sim\pi} \mathbb{P}_\theta[L(\theta, T(X)) > \Delta] \geq 1 - p_\Delta - \sqrt{p_\Delta(1-p_\Delta) I_{\chi^2}(\theta; X)},$$

   where the quantity $p_\Delta \triangleq \max_{a\in\mathcal{A}} \pi(\{\theta \in \Theta : L(\theta, a) \leq \Delta\})$ is defined in Lecture 10, and $I_{\chi^2}(\theta; X)$ is the χ^2-mutual information defined as

   $$I_{\chi^2}(\theta; X) \triangleq \inf_{Q_X} \mathbb{E}_{\theta\sim\pi}[\chi^2(P_{X|\theta}, Q_X)].$$

   Hint: consider the kernel $K : \Theta \times \mathcal{X} \to \{0, 1\}$ which sends (θ, x) to $\mathbb{1}(L(\theta, T(x)) > \Delta)$, apply a data-processing inequality specialized to χ^2-divergence. It is acceptable to prove a weaker result, provided that your result is sufficient to handle Part (c).

(b) Consider the uniform distribution $\pi = \mathsf{Unif}(\{\tau e_1, \cdots, \tau e_p\})$, where $e_1, \cdots, e_p$ are canonical vectors of $\mathbb{R}^p$, and $\tau > 0$ is a parameter to be determined. We apply the above method with $L(\theta, a) = \|\theta - a\|_2^2$ and $\Delta = (1-\delta)\tau^2$. Show that $p_\Delta \le 1/(p\delta)$. *Note: this shows that the best separation parameter here is $(1-o(1))\tau^2$, as opposed to $\tau^2/2$ in the original Fano's method.*

(c) Choosing $\tau = \sqrt{(2-\varepsilon)\log p}$, prove the claimed minimax lower bound using the above method and choosing $\varepsilon, \delta \to 0$ appropriately.

(d) Based on the minimax lower bound for 1-sparse vectors, argue that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \ge (2 - o(1)) s \log(p/s), \qquad p/s \to \infty.$$

4. This problem is a continuation of the learning theory example covered in Lecture 11. We have a function class $\mathcal{F}$ with VC dimension d, and n training data $(x_1, y_1), \cdots, (x_n, y_n)$ drawn from an unknown joint distribution P_{XY}, with $\mathcal{Y} = \{0, 1\}$. Define the following class $\mathcal{P}(\mathcal{F}, \varepsilon)$ of joint distributions where the best classifier has an error at most ε:

$$\mathcal{P}(\mathcal{F}, \varepsilon) = \left\{ P_{XY} : \inf_{f^\star \in \mathcal{F}} P_{XY}(Y \ne f^\star(X)) \le \varepsilon \right\}.$$

So $\varepsilon = 0$ corresponds to the well-specified case, and $\varepsilon = 1$ corresponds to the misspecified case. Define the minimax excess risk $R^\star(\mathcal{F}, \varepsilon)$ over $\mathcal{P}(\mathcal{F}, \varepsilon)$ as

$$R^\star(\mathcal{F}, \varepsilon) = \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}(\mathcal{F}, \varepsilon)} \mathbb{E}\left[P_{XY}(Y \ne \widehat{f}(X)) - \inf_{f^\star \in \mathcal{F}} P_{XY}(Y \ne f^\star(X)) \right].$$

Show that for all $\varepsilon \in [0, 1]$,

$$R^\star(\mathcal{F}, \varepsilon) = \Omega\left(\min\left\{ \sqrt{\frac{d}{n} \cdot \varepsilon} + \frac{d}{n}, 1 \right\} \right).$$

Note: the empirical risk minimization (ERM) approach attains the above lower bound possibly up to logarithmic factors. A possible hint is to try out the marginal distribution construction in the optimistic case and the conditional distribution construction in the pessimistic case, with appropriate parameters tailored for this problem.

5. In this problem we construct explicit packings in some examples, and therefore prove lower bounds on the packing number $M(A, d, \varepsilon)$. Recall that $M(A, d, \varepsilon)$ denotes the maximum number m of points $x_1, \cdots, x_m$ such that $d(x_i, x_j) \ge \varepsilon$ for every $i \ne j \in [d]$.

(a) Suppose that $A = \{\pm 1\}^n$ is the binary hypercube, and $d = d_{\mathrm{H}}$ is the Hamming distance. Show the following Gilbert–Varshamov bound: for $\varepsilon \in [0, 1/2]$,

$$M(\{\pm 1\}^n, d_{\mathrm{H}}, n\varepsilon) \ge 2^{n(1 - h_2(\varepsilon))},$$

where $h_2(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. *Hint: you may use the following result: for every $x \in \{\pm 1\}^n$ and $\varepsilon \in [0, 1/2]$,*

$$|\{y \in \{0,1\}^n : d_{\mathrm{H}}(x, y) \le n\varepsilon\}| \le 2^{n h_2(\varepsilon)}.$$

(b) Suppose that A is the set of all non-decreasing functions $f : [0, 1] \to [0, 1]$, and d is the L_2 norm between functions. Show that for $\varepsilon \in [0, 1]$, there exist universal constants $c_1, c_2 > 0$ such that

$$\log M(A, L_2([0, 1]), c_1\varepsilon) \geq \frac{c_2}{\varepsilon}.$$

(c) Suppose that A is the set of all convex functions $f : [0, 1] \to [0, 1]$, and d is the L_2 norm between functions. Show that for $\varepsilon \in [0, 1]$, there exist universal constants $c_1, c_2 > 0$ such that

$$\log M(A, L_2([0, 1]), c_1\varepsilon) \geq \frac{c_2}{\sqrt{\varepsilon}}.$$

Hint: for (b) and (c), you may try to break into several small intervals, find two possible function constructions in each interval, and concatenate them. The result in (a) might also be useful.