ORIE 7790 High Dimensional Probability and Statistics Lecture 18 - 04/14/2020

Lecture 18: Minimax Lower Bounds

Lecturer: Yudong Chen *Scribe: Liwei Jiang*

References:

- Wainwright book: Chap 15.1, 15.3 and 15.4
- Duchi note, `http://web.stanford.edu/class/stats311/lecture-notes.pdf`: Chap 7, 10
- "Introduction to Nonparametric Estimation" by Alexandre Tsybakov
- "Assouad, Fano, and Le Cam" by Bin Yu, `https://www.stat.berkeley.edu/~binyu/ps/LeCam.pdf`: A nice, short article on three basic methods for proving lower bounds

1 Minimax Framework

We consider the following setting:

- A family of distributions: $\{\mathbb{P}_\theta : \theta \in \Theta\}$
- Observe $X \sim \mathbb{P}_\theta$. We want to estimate θ.
- Estimator: $\hat{\theta}(\cdot)$. This is a measurable function of X.

Given metric $\rho(\Theta \times \Theta \to \mathbb{R}_+)$, define

- Wors-case risk for the estimator $\hat{\theta}$: $\sup_{\theta\in\Theta} \mathbb{E}\left[\rho(\hat{\theta}(X), \theta)\right]$
- Minimax risk: $\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}\left[\rho(\hat{\theta}(X), \theta)\right]$, where the infimum is over all estimators.

2 From Estimation to Testing

Consider a multiple test with M hypotheses: $\theta_1, \ldots, \theta_M \in \Theta$. The test procedure is $\psi(X) \in \{1, 2, \ldots, M\}$; that is, given the data X, we pick hypothesis $\theta_{\psi(X)}$.

Theorem 1. *If $\{\theta_1, \ldots, \theta_M\}$ is a 2δ-packing of Θ w.r.t. the metric ρ, then*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}\left[\rho(\hat{\theta}(X), \theta)\right] \geq \delta \cdot \inf_{\psi} \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}[\psi(X) \neq j |\, \mathbb{P}_{\theta_j}].$$

With this theorem, the problem of lower bounding the minimax risk reduces to bounding the average test error on the RHS above. To this send, we will draw techniques from information theory.

3 Some Information Theory

We'll give some definitions and related inequalities from information theory.

- Entropy: $H(Q) \triangleq -\int q(x) \log q(x) dx$. Here $q(\cdot)$ is the density of the distribution Q.

- Conditional Entropy: For $(X, Y) \sim Q_{X,Y}$,

$$H(X|Y) \triangleq \mathbb{E}_Y[H(Q_{X|Y})] = \mathbb{E}_Y[-\int q(x|Y) \log q(x|Y) dx].$$

- (KL-Divergence): $D(P||Q) \triangleq \int p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_P[\log \frac{p(X)}{q(X)}]$.
- Mutual Information: $(X, Y) \sim Q_{X,Y}$,

$$\begin{aligned} I(X, Y) &\triangleq D(Q_{X,Y}||Q_X \cdot Q_Y) \\ &= H(X) + H(Y) - H(X, Y) && \text{by definition} \\ &= H(Y) - H(Y|X). && \text{by chain rule below} \end{aligned}$$

Remark Note that entropy, KL and mutual information are all non-negative. The also satisfy the following properties:

- (Conditioning reduces entropy) $H(X|Y) \leq H(X)$.
- (Chain rule) $H(X, Y) = H(Y) + H(X|Y)$.
- (Chain rule) $H(X, Y|Z) = H(Y|Z) + H(X|Y, Z)$.

4 Fano's method

Assume that the index J for the true hypothesis is sampled uniformly from $\{1, 2, \dots, M\}$ and that $X|_{J=j} \sim \mathbb{P}_{\theta_j}$.

Theorem 2 (Fano's Inequality)**.** *For any text procedure ψ, we have*

$$\mathbb{P}[\psi(X) \neq J] \geq 1 - \frac{I(X; J) + \log 2}{\log M}.$$

Proof Let $q_e \triangleq \mathbb{P}[\psi(X) \neq J]$ be the error probably and $h(q_e) \triangleq -q_e \log(q_e) = (1 - q_e) \log(1 - q_e)$ be the (binary) entropy of Bernoulli(q_e). The proof contains two steps:

1. Show that $h(q_e) + q_e \log M \geq H(J|X)$.
2. Show that inequality above implies Fano's inequality.

Step 1: Define the indicator random variable $V \triangleq 1_{\{\psi(X) \neq J\}}$. We have $V \sim \text{Ber}(q_e)$, $H(V) = h(q_e)$. The key idea is to decompose the entropy $H(V, J|X)$ in two ways using the chain rule:

$$H(V, J|X) = \begin{cases} H(J|X) + H(V|J, X) \overset{(i)}{=} H(J|X), \\ H(V|X) + H(J|V, X) \overset{(ii)}{\leq} h(q_e) + H(J|V, X). \end{cases}$$

Step (i) holds because V is determined by J, X and hence $H(V|J, X) = 0$. Step 2 holds because conditioning reduces entropy.

Note that by definition of entropy, we have

$$H(J|V, X) = \mathbb{P}[V = 1] H(J|V = 1, X) + \mathbb{P}[V = 0] H(J|V = 0, X) \leq q_e \log M$$

The last inequality holds due to the following observations: (a) when $V = 0$, J is known given X, so $H(J|V = 0, X) = 0$; (b) $H(J|V = 1, X) \leq \log M$ since the uniform distribution maximizes entropy.

Combing pieces, we get the desired inequality:

$$H(J|X) \leq h(q_e) + q_e \log M$$

Step 2: Using the alternative expression for the mutual information, we have

$$H(J|X) = H(J) - I(X, J) = \log M - I(X, J).$$

Combining with Step 1, we get

$$q_e \geq 1 - \frac{I(X, J) + h(q_e)}{\log M} \geq 1 - \frac{I(X, J) + \log 2}{M},$$

thereby proving the Fano's inequality. □

Combining Theorem 1 and 2, we can develop the so called "local" Fano's method by using a particular upper bound on the mutual information $I(X, J)$. In particular, letting $Q_X := \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}_{\theta_i}$ be the marginal distribution of X, we have

$$\begin{aligned} I(X, J) &= D(Q_{X,J} || Q_X \cdot Q_J) \\ &= \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta_j} || Q_X) \\ &\overset{(i)}{\leq} \frac{1}{M^2}\sum_{i,j} D(\mathbb{P}_{\theta_j} \,||\, \mathbb{P}_{\theta_i}) \\ &\leq \max_{i,j} D(\mathbb{P}_{\theta_j} \,||\, \mathbb{P}_{\theta_i}), \end{aligned}$$

where step (i) follows from the convexity of KL and Jensen's Inequality. Therefore, we have the following corollary:

Corollary 1 (Local Fano's Method)**.** *Let $\{\theta_1, \ldots, \theta_M\}$ be 2δ-packing of Θ satisfying*

$$\max_{i,j} D(\mathbb{P}_{\theta_i} \,||\, \mathbb{P}_{\theta_j}) \leq g(\delta).$$

Then we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\left[\rho(\hat{\theta}(X), \theta)\right] \geq \delta\big(1 - \frac{g(\delta) + \log 2}{\log M}\big).$$

In the next lecture, we will apply Corollary 1 to concrete statistical problems.