36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 15: March 19

Lecturer: Siva Balakrishnan *Scribe: Tian Tong*

15.1 Recap of minimax lower bounds

The smallest worst-case risk among all estimators, known as the minimax risk, is defined as

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta(\mathbb{P})))],$$

where the estimator $\widehat{\theta}$ is based on n i.i.d. samples from $\mathbb{P}$; $\rho : \Omega \times \Omega \to [0, \infty)$ is a semi-metric; $\Phi : [0, \infty) \to [0, \infty)$ is an increasing function.

To develop a lower bound of the minimax risk, first reduce it to a testing problem. Suppose that $\{\theta^1, \cdots, \theta^M\}$ are 2δ separated, i.e., $\rho(\theta^i, \theta^j) \geq 2\delta, \forall i \neq j$. Sample the random integer J uniformly from $[M] = \{1, \cdots, M\}$, and then sample $Z \sim \mathbb{P}_{\theta^J}$. Let $\mathbb{Q}$ denote the joint distribution of the pair (Z, J). Let $\psi : \mathcal{Z} \to [M]$ be an M-ary testing function. The minimax risk is lower bounded by the error probability as

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J].$$

Next, there are two methods for the testing problem.

1. Le Cam's method applies to the binary testing problem, i.e., $M = 2$. The minimax risk for binary testing can be expressed explicitly in terms of the TV distance as

$$\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}(1 - \mathrm{TV}(\mathbb{P}_{\theta^1}, \mathbb{P}_{\theta^2})).$$

2. Fano's method applies to the M-ary testing problem. It follows from the Fano's inequality from information theory as

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

   Using the convexity of the KL divergence, the mutual information $I(Z; J)$ can be upper bounded by

$$I(Z; J) = \frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}, \bar{\mathbb{P}}) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}, \mathbb{P}_{\theta^k}).$$

15.2 Some divergence measures

To develop techniques for lower bounding the error probability, some background on some important divergence measures is required. Introduce the total variation (TV) distance, the KL divergence, and the Hellinger distance as

$$\begin{aligned}
\mathrm{TV}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int |p(x)-q(x)|dx;\\
\mathrm{KL}(\mathbb{P},\mathbb{Q}) &= \int p(x)\log\frac{p(x)}{q(x)}dx;\\
H^2(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int (\sqrt{p(x)}-\sqrt{q(x)})^2dx.
\end{aligned}$$

They are related by a sequence of inequalities as

$$H^2(\mathbb{P},\mathbb{Q}) \le \mathrm{TV}(\mathbb{P},\mathbb{Q}) \le \sqrt{2}H(\mathbb{P},\mathbb{Q}) \le \sqrt{\mathrm{KL}(\mathbb{P},\mathbb{Q})} \le \sqrt{\chi^2(\mathbb{P},\mathbb{Q})}.$$

Given n i.i.d. samples, it is natural to express the divergence between the product measures $\mathbb{P}^n$ and $\mathbb{Q}^n$ in terms of divergences between the individual pairs. The TV distance behaves badly and is unable to decompose. The KL divergence exhibits an attractive property as

$$\mathrm{KL}(\mathbb{P}^n,\mathbb{Q}^n) = \sum_{i=1}^n \mathrm{KL}(\mathbb{P}_i,\mathbb{Q}_i) = n\mathrm{KL}(\mathbb{P},\mathbb{Q}) \text{ for i.i.d. cases.}$$

The Hellinger distance has a similar property as

$$\begin{aligned}
H^2(\mathbb{P}^n,\mathbb{Q}^n) &= 1 - \underbrace{\int \sqrt{p(x^n)q(x^n)}dx^n}_{\text{affinity}} = 1-\prod_{i=1}^n \int \sqrt{p(x_i)q(x_i)}dx_i\\
&= 1-\prod_{i=1}^n (1-H^2(\mathbb{P}_i,\mathbb{Q}_i))\\
&= 1-(1-H^2(\mathbb{P},\mathbb{Q}))^n \le nH^2(\mathbb{P},\mathbb{Q}) \text{ for i.i.d. cases.}
\end{aligned}$$

15.3 Examples of Le Cam's method

The spirit of deriving minimax lower bounds is to design parameters with relatively large separation distance, while keeping the divergence measures relatively small.

15.3.1 Gaussian location family

Consider $X_1, \cdots, X_n \sim N(\theta, 1)$, with $\theta \in \mathbb{R}$. The metric is $\Phi \circ \rho(\hat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$.

Apply Le Cam's method with two parameters $\theta^1 = 0, \theta^2 = \delta$, which are δ separated. The two distributions are $\mathbb{P}_{\theta^1} = N(0,1), \mathbb{P}_{\theta^2} = N(\delta, 1)$. Calculate the KL divergence as $\mathrm{KL}(\mathbb{P}^n_{\theta^1}, \mathbb{P}^n_{\theta^2}) = n\delta^2/2$. Set δ such that $n\delta^2 = c$ for some small constant c.

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq (\frac{\delta}{2})^2 \frac{1}{2}(1 - \sqrt{\mathrm{KL}(\mathbb{P}^n_{\theta^1}, \mathbb{P}^n_{\theta^2})}) = \frac{\delta^2}{8}(1 - \sqrt{\frac{n\delta^2}{2}}) \gtrsim \frac{1}{n}.$$

15.3.2 Uniform location family

Consider $X_1, \cdots, X_n \sim \mathrm{Unif}[\theta, \theta + 1]$, with $\theta \in \mathbb{R}$. The metric is $\Phi \circ \rho(\hat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$.

Apply Le Cam's method with two parameters $\theta^1 = 0, \theta^2 = \delta$, which are δ separated. The two distributions are $\mathbb{P}_{\theta^1} = \mathrm{Unif}[0, 1], \mathbb{P}_{\theta^2} = \mathrm{Unif}[\delta, \delta + 1]$. Notice that the supports of two distributions mismatch, so the KL divergence is infinite and not applicable here. Instead, calculate the Hellinger distance as

$$H^2(\mathbb{P}_{\theta^1}, \mathbb{P}_{\theta^2}) = \frac{1}{2}\int_0^{\delta} (1-0)^2 dx + \frac{1}{2}\int_{\delta}^{1+\delta} (0-1)^2 dx = \delta,$$

and therefore $H^2(\mathbb{P}^n_{\theta^1}, \mathbb{P}^n_{\theta^2}) \leq n\delta$. Set $\delta = c/n$ for some small constant c.

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq (\frac{\delta}{2})^2 \frac{1}{2}(1 - \sqrt{2}H(\mathbb{P}^n_{\theta^1}, \mathbb{P}^n_{\theta^2})) = \frac{\delta^2}{8}(1 - \sqrt{2n\delta}) \gtrsim \frac{1}{n^2}.$$

15.3.3 Pointwise estimation of Lipschitz densities

Consider $Y_i = f(X_i) + \epsilon_i$, with $X_i \sim \mathrm{Unif}[0,1]$, $\epsilon_i \sim N(0,1)$, and

$$f \in \mathcal{F}_L = \{f : [0,1] \to \mathbb{R} : |f(x) - f(y)| \leq L|x - y|, \forall x, y \in [0,1], f(0) = 0\}$$

as a Lipschitz function. The goal is to estimate $\theta = f(0.5)$. The metric is $\Phi \circ \rho(\hat{\theta}, \theta) = |\widehat{\theta} - \theta|$.

Apply Le Cam's method with two functions $f_1(x) = 0, f_2(x) = L(h - |x - 0.5|)_+$, where $\theta^1 = 0, \theta^2 = Lh$ are Lh separated. The two distributions are

$$\begin{aligned} \mathbb{P}_{f_1} : X &\sim \mathrm{Unif}[0,1], \\ y|X &\sim N(0,1), \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}_{f_2} : X &\sim \mathrm{Unif}[0,1] \\ y|X &\sim N(f_2(x), 1). \end{aligned}$$

Calculate the KL divergence as

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}) &= \int p_{f_1}(x,y) \log \frac{p_{f_1}(x,y)}{p_{f_2}(x,y)} dxdy \\
&= \int p_{f_1}(x,y) \log \frac{p_{f_1}(y|x)}{p_{f_2}(y|x)} dxdy \\
&= \int p_{f_1}(x) \int p_{f_1}(y|x) \log \frac{p_{f_1}(y|x)}{p_{f_2}(y|x)} dy \cdot dx \\
&= \int p_{f_1}(x) \mathrm{KL}(N(0,1), N(f_2(x),1)) dx \\
&= \int_0^1 \frac{f_2^2(x)}{2} dx = 2 \int_0^h \frac{L^2 x^2}{2} dx = \frac{L^2 h^3}{3}.
\end{aligned}$$

Set h such that $nL^2h^3 = c$ for some constant c.

$$\mathfrak{M}(\theta(\mathcal{P}); \rho) \geq \frac{Lh}{2} \cdot \frac{1}{2}(1 - \sqrt{\mathrm{KL}(\mathbb{P}_{f_1}^n, \mathbb{P}_{f_2}^n)}) = \frac{Lh}{4}(1 - \sqrt{\frac{nL^2h^3}{3}}) \gtrsim (\frac{L}{n})^{1/3}.$$

To summarize our three examples:

1. In normal mean estimation, we see the standard parametric rate (in low-dimensions) of $1/\sqrt{n}$ (on the non-squared scale).
2. In uniform location estimation, we see a non-standard, faster than parametric rate of $1/n$ (on the non-squared scale).
3. For estimating Lipschitz functions at a point, we see the slower than parametric rate of $n^{-1/3}$. More generally, for estimating β-Holder functions at a point, in d dimensions, we will obtain the rate $n^{-\beta/(2\beta+d)}$, using essentially the same construction.

In each case, it is straightforward to construct corresponding upper bounds.