

## Lecture 30

Lecturer: Christina Lee Yu

Scribe: Ramchandran Muthukumar

**Remark 1** For more information on the topic please see Lecture notes for STAT311 by John Duchi at Stanford which was used as a reference in preparing these notes.

The goal of this lecture is to demonstrate how lower bounds for the minimax estimator can be obtained for two simple applications.

## 1 Recap

We had earlier looked at the Le-Cam's Inequality in the context of simple-vs-simple hypothesis testing and its relation to Total Variation Distance and the K-L Divergence. These results have been summarized here for reference.

**Theorem 1** *Le-Cam's Method - Reducing Estimation to Hypothesis Testing*

Suppose that  $X_1, \dots, X_n \sim \mathcal{P}_\theta$  for  $\theta \in \Theta$  are independent and identically distributed,  $\rho : \Theta \times \Theta \rightarrow \mathbb{R}_+$  is a semi-metric, and  $\Phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a nondecreasing function such that  $\Phi(0) = 0$ . Then for any  $\theta_0, \theta_1 \in \Theta$  such that  $\rho(\theta_0, \theta_1) > 2\delta$  then

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} [\Phi(\rho(\hat{\theta}(X), \theta))] \geq \Phi(\delta) \inf_{\Psi} \frac{1}{2} (\Pr(\Psi(X) \neq 0 \mid \mathcal{P}_{\theta_0}) + \Pr(\Psi(X) \neq 1 \mid \mathcal{P}_{\theta_1}))$$

where the infimum is taken over all functions  $\Psi : \mathcal{X} \rightarrow \{0, 1\}$ .

**Theorem 2** *Relating Hypothesis Test Error to Total Variation Distance*

Let  $\mathcal{X}$  be an arbitrary set. For any distributions  $P_0$  and  $P_1$  on  $\mathcal{X}$ ,

$$\inf_{\Psi} \{P_0(\Psi(X) \neq 0) + P_1(\Psi(X) \neq 1)\} = 1 - \|P_0 - P_1\|_{TV}$$

where the infimum is taken over all tests  $\Psi : \mathcal{X} \rightarrow \{0, 1\}$

**Theorem 3** *Pinkster's Inequality - Relating Total Variation Distance to K-L Divergence*

For any two distributions  $P_0$  and  $P_1$  on  $\mathcal{X}$ ,

$$\|P_0 - P_1\|_{TV} \leq \sqrt{\frac{D_{KL}(P_0 \| P_1)}{2}}$$

Combining Theorems ??, ?? and ??, we can rewrite the minimax lower bound in terms of the K-L Divergence which is easier to compute.

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} [\Phi(\rho(\hat{\theta}(X), \theta))] \geq \frac{1}{2} \Phi(\delta) \left( 1 - \sqrt{\frac{D_{KL}(P_0 \| P_1)}{2}} \right) \quad (1)$$

We also note the following property of K-L Divergence.

**Lemma 4** *K-L divergence on Product Distributions*

Let  $P_0$  and  $P_1$  denote distributions on  $\mathcal{X}$ , Let  $P_0^n$  and  $P_1^n$  denote their product distributions. Then,

$$D_{KL}(P_0^n \| P_1^n) = n D_{KL}(P_0 \| P_1)$$

## 2 Application 1 - Testing Normal Mean

Suppose we observe data  $X_1, X_2, \dots, X_n \sim \mathbf{N}(\mu, \sigma^2)$ . We wish to estimate  $\mu$  (when  $\sigma^2$  is known).

### 2.1 Lower bound for hypothesis testing

In the simple-vs-simple hypothesis testing let the null hypothesis be  $H_0 : \mu = \mu_0$  and let the alternate hypothesis be  $H_1 : \mu = \mu_1$ . We denote  $P_0 = \mathbf{N}(\mu_0, \sigma^2)$  and  $P_1 = \mathbf{N}(\mu_1, \sigma^2)$ . Let  $P_0^n$  denote the product distribution over  $n$  samples iid from  $P_0$ , and let  $P_1^n$  denote the product distribution over  $n$  samples iid from  $P_1$ .

Let us first compute a lower bound on the total probability of error of hypothesis testing, which we recall from Theorem 2 and 3 that

$$\begin{aligned} \inf_{\Psi} (\Pr(\Psi(X) \neq 0 \mid \mu = \mu_0) + \Pr(\Psi(X) \neq 1 \mid \mu = \mu_1)) &= 1 - \|P_0^n - P_1^n\|_{TV} \\ &\geq 1 - \sqrt{\frac{D_{KL}(P_0^n \| P_1^n)}{2}}. \end{aligned}$$

The bound with respect to KL-divergence is easier to bound for product distributions than directly computing the total variation distance between product distributions. In fact, we can show that

$$D_{KL}(P_0^n \| P_1^n) = n D_{KL}(P_0 \| P_1).$$

We can compute  $D_{KL}(P_0 \| P_1)$  as

$$\begin{aligned} D_{KL}(P_0 \| P_1) &= \int \mathcal{P}_{\mu_0}(x) \log \left( \frac{P_{\mu_0}(x)}{P_{\mu_1}(x)} \right) dx \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - \mu_0)^2}{2\sigma^2} \right) \left[ \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2} + \frac{2(\mu_0 - \mu_1)}{2\sigma^2} \right] dx \\ &= \frac{1}{2\sigma^2} (\mu_1 - \mu_0)^2 \end{aligned}$$

Therefore, the lower bound for the total probability of error for testing between two Gaussians of mean  $\mu_0$  and  $\mu_1$  is

$$1 - \frac{|\mu_1 - \mu_0|}{2\sigma} \sqrt{n}.$$

This implies that if  $n < \frac{\sigma^2}{(\mu_1 - \mu_0)^2}$ , then the total probability of error for hypothesis testing must be larger than  $\frac{1}{2}$ .

### 2.2 Upper bound for hypothesis testing

Without loss of generality, assume that  $\mu_0 < \mu_1$ . Consider the test which rejects the null hypothesis if  $\frac{1}{n} \cdot \sum_{i=1}^n X_i \geq \frac{1}{2}(\mu_1 - \mu_0)$ . We can use Hoeffding's inequality to bound the total probability of error, in particular the test can only makes an error in the case that

$$\left| \frac{1}{n} \cdot \sum_{i=1}^n X_i - \mu \right| \geq \frac{1}{2}(\mu_1 - \mu_0).$$

We can bound this event using Hoeffding's inequality,

$$\mathbb{P} \left( \left| \frac{1}{n} \cdot \sum_{i=1}^n X_i - \mu \right| \geq \frac{1}{2}(\mu_1 - \mu_0) \right) \leq 2 \exp \left( \frac{-n(\mu_1 - \mu_0)^2}{8\sigma^2} \right)$$

Therefore, if  $\Psi$  indicates the specific test defined above, then

$$(\Pr(\Psi(X) \neq 0 \mid \mu = \mu_0) + \Pr(\Psi(X) \neq 1 \mid \mu = \mu_1)) \leq 4 \exp\left(\frac{-n(\mu_1 - \mu_0)^2}{8\sigma^2}\right).$$

Therefore, if  $n > \frac{8 \ln(8)\sigma^2}{(\mu_1 - \mu_0)^2}$ , then the above simple test will achieve total probability of error less than  $\frac{1}{2}$ .

### 2.3 Lower bound for mean squared error in estimation

Next, suppose we would like to compute a lower bound on the minimax risk for the estimation task with respect to the squared loss (corresponding to mean squared error),

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} \left[ (\hat{\theta}(X) - \theta)^2 \right]$$

The semi-metric is  $\rho(\mu_0, \mu_1) = |\mu_0 - \mu_1|$  and the loss is  $\Phi(\rho) = \rho^2$ .

From equation ?? and Lemma ??, it follows that for any choice of  $\mu_0$  and  $\mu_1$  such that  $|\mu_0 - \mu_1| \geq 2\delta$ , the minimax estimator is lower bounded by

$$\begin{aligned} \inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} \left[ \Phi(\rho(\hat{\theta}(X), \theta)) \right] &\geq \frac{1}{2} \Phi(\delta) \left( 1 - \sqrt{\frac{n D_{KL}(P_0 \| P_1)}{2}} \right) \\ &= \frac{1}{2} \delta^2 \left( 1 - \frac{|\mu_1 - \mu_0|}{2\sigma} \sqrt{n} \right). \end{aligned}$$

We would like to choose  $\delta, \mu_0$ , and  $\mu_1$  to maximize the lower bound. For some pair of  $\mu_0$  and  $\mu_1$ , the maximum delta we can choose is  $(\mu_0 - \mu_1)/2$ , while still satisfying the constraint that  $\mu_0$  and  $\mu_1$  are at least  $2\delta$  distance away. By plugging this value of  $\delta$  in, it follows that

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} \left[ \Phi(\rho(\hat{\theta}(X), \theta)) \right] \geq \frac{|\mu_1 - \mu_0|^2}{8} \left( 1 - \frac{|\mu_1 - \mu_0|}{2\sigma} \sqrt{n} \right)$$

Finally we want to choose the value of  $\mu_0$  and  $\mu_1$  again to maximize the lower bound. If we differentiate the right hand side wrt  $|\mu_1 - \mu_0|$ , we get

$$\frac{|\mu_1 - \mu_0|}{4} - \frac{3|\mu_1 - \mu_0|^2}{16\sigma} \sqrt{n}.$$

The stationary point at  $|\mu_1 - \mu_0|$  is a maximum not a minimum, and we can show the lower bound is maximized at

$$|\mu_1 - \mu_0| = \frac{4\sigma}{3\sqrt{n}}.$$

Hence by plugging in this value for  $|\mu_1 - \mu_0|$ , we can show that

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} \left[ \Phi(\rho(\hat{\theta}(X), \theta)) \right] \geq \frac{1}{8} \left( \frac{4\sigma}{3\sqrt{n}} \right)^2 \left( 1 - \frac{2}{3} \right) = \frac{1}{8} \cdot \frac{16\sigma^2}{9n} \cdot \frac{1}{3} = \frac{2\sigma^2}{27n}$$

## 2.4 Upper bound for estimation

Consider the simple estimator  $\hat{\mu} := \frac{1}{n} \cdot \sum_{i=1}^n X_i$ . We can show an upper bound on the worst case mean-squared error using Hoeffding's Inequality.

$$\begin{aligned} \mathbb{E}[(\hat{\mu} - \mu)^2] &= \int_0^\infty \mathbb{P}((\hat{\mu} - \mu)^2 > t) dt \\ &\leq \int_0^\infty 2 \exp\left(\frac{-nt}{2\sigma^2}\right) dt \\ &= 2 \left(\frac{-2\sigma^2}{n}\right) \exp\left(\frac{-nt}{2\sigma^2}\right) \Big|_0^\infty \\ &= \frac{4\sigma^2}{n} \end{aligned}$$

Thus the sample mean estimator cannot do worse than  $\frac{4\sigma^2}{n}$  and no estimator can do better than  $\frac{2\sigma^2}{27n}$ .

## 3 Application 2 - Kernel Regression

In Kernel Regression we observe data  $Y_i = f(X_i) + \epsilon_i$  where  $X_i$  is a known feature matrix and the errors/noise  $\epsilon_i \sim \mathbf{N}(0, 1)$ .

We wish to estimate the function  $f$  under the assumption that it is Lipschitz continuous (with constant  $L$ ).

The **Nadaraya-Watson Estimator** for fixed  $x_0$  is

$$\hat{f}(x) = \frac{\sum_{i=1}^n K(x_i, x) Y_i}{\sum_{i=1}^n K(x_i, x)}$$

for some choice of kernel function  $K$ . We showed earlier that for some constant  $c$  and fixed  $x_0$ ,

$$\sup_{f \text{ is } L\text{-Lipschitz}} \mathbb{E}[(\hat{f}(x_0) - f(x_0))^2] \leq c \left(\frac{L\sigma^2}{n}\right)^{\frac{2}{3}}$$

We will now show a lower bound.

As our risk is the MSE, we choose the loss and the semi-metric according to,

$$\Phi(\rho) = \rho^2, \quad \rho(f_0, f_1) = |f_0(x_0) - f_1(x_0)| \quad (\text{distance at } x_0).$$

We choose the two functions to test on as  $f_0(x) = 0$  for all  $x$  and

$$f_1(x) = \begin{cases} L(\epsilon - |x - x_0|) & \forall |x - x_0| \leq \epsilon \\ 0 & \text{otherwise.} \end{cases}$$

Note that  $\rho(f_0, f_1) = L\epsilon$ . Corresponding to  $f_0$  and  $f_1$ , we have the two distributions

$$P_0 \left( \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \mid f = f_0 \right) \sim \mathbf{N} \left( \begin{bmatrix} f_0(X_1) \\ \vdots \\ f_0(X_n) \end{bmatrix}, \sigma^2 \mathbf{I} \right) = \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad \text{and} \quad P_1 \left( \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \mid f = f_1 \right) \sim \mathbf{N} \left( \begin{bmatrix} f_1(X_1) \\ \vdots \\ f_1(X_n) \end{bmatrix}, \sigma^2 \mathbf{I} \right)$$

So to obtain a lower bound we need to compute the KL divergence of these two multivariate gaussian distributions.

$$\begin{aligned}
D_{KL}(\mathbf{N}(\mu_0, \sigma^2 \mathbf{I}) \parallel \mathbf{N}(\mu_0, \sigma^2 \mathbf{I})) &= \int P_0(\vec{Y}) \log \left( \frac{P_0(\vec{Y})}{P_1(\vec{Y})} \right) d\vec{Y} \\
&= \int \prod_{i=0}^n P_{0i}(Y_i) \log \left( \frac{\prod_{i=0}^n P_{0i}(Y_i)}{\prod_{i=0}^n P_{1i}(Y_i)} \right) d\vec{Y} \\
&= \sum_{i=0}^n \int \log \left( \frac{P_{0i}(Y_i)}{P_{1i}(Y_i)} \right) P_{0i}(Y_i) dY_i \cdot \underbrace{\left( \prod_{j \neq i} P_{0j}(Y_j) dY_j \right)}_{\text{integrates to one}} \\
&= \sum_{i=0}^n \int \log \left( \frac{P_{0i}(Y_i)}{P_{1i}(Y_i)} \right) P_{0i}(Y_i) dY_i \quad \equiv \text{KL divergence of normal distributions} \\
&= \sum_{i=0}^n \frac{1}{2\sigma^2} (f_0(X_i) - f_1(X_i))^2 \\
&= \sum_{i=0}^n \mathbf{1}\{|X_i - X_0| < \epsilon\} \cdot \frac{L^2(\epsilon - |X_i - X - 0|)^2}{2\sigma^2} \\
&= \frac{L^2}{2\sigma^2} \cdot 2 \sum_{i=0}^{\epsilon n} \left( \frac{i}{n} \right)^2 \quad \text{under assumption that } X_i = \frac{i}{n} \\
&= \frac{L^2}{\sigma^2 n^2} \cdot \left[ \frac{\epsilon n(\epsilon n + 1)(2\epsilon n + 1)}{6} \right] \\
&\approx \frac{L^2}{\sigma^2 n^2} \cdot \frac{\epsilon^3 n^3}{3}.
\end{aligned}$$

Therefore the minimax lower bound is

$$\begin{aligned}
\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} [\Phi(\rho(\hat{\theta}(X), \theta))] &\geq \frac{1}{2} \Phi(\rho) \cdot \left( 1 - \sqrt{\frac{1}{2} D_{KL}(P_0 \parallel P_1)} \right) \\
&\approx \frac{1}{2} \left( \frac{L\epsilon}{2} \right)^2 \cdot \left( 1 - \sqrt{\frac{L^2 \epsilon^3 n}{6\sigma^2}} \right).
\end{aligned}$$

By differentiating the right hand side wrt  $\epsilon$  and equating to zero we can see that the lower bound is maximized at

$$\epsilon = \Theta \left( \left( \frac{\sigma}{L\sqrt{n}} \right)^{\frac{2}{3}} \right).$$

By plugging in this value of  $\epsilon$ , we show that the minimax lower bound is

$$\inf_{\hat{\theta}(X)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\theta} [\Phi(\rho(\hat{\theta}(X), \theta))] = \Omega \left( \left( \frac{L\sigma^2}{n} \right)^{\frac{2}{3}} \right),$$

matching our upper bound up to a constant.