CS861: Theoretical Foundations of Machine Learning Lecture 9 - 25/09/2023
University of Wisconsin–Madison, Fall 2023

Lectu[illegible]s testing and Le Cam's method

Lecturer: Kirthe[illegible] *Scribed by: Haoyue Bai, Ying Fu*

Disclaimer: *These n[illegible]ected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we begin by recapping **point estimate** and **minimax optimality** from the previous class. Then, we will introduce the concept of **hypothesis testing**, along with the theorem of **reduction from estimation to testing**. Finally, we will introduce the **Le Cam methods**, which are fundamental tools in establishing minimax lower bounds.

1 From Estimation to Testing

A standard first step in proving minimax bounds is to "reduce" the estimation problem to a testing problem. Then, we need to show that the estimation risk can be lower bounded by the probability of error in testing problems, which we can develop tools for. We first define the hypothesis test problems we will use.

Definition 1 (Hypothesis Test). *Let Q be a class of distributions, and let $Q_1, Q_2, .., Q_N$ be a partition of Q. Let S be a dataset drawn from some $P \in Q$. A (multiple) hypothesis test Ψ is a function of the data which maps to $\{1, ..., N\} \triangleq [N]$. If $\Psi(S) = j$, the test has decided that $P \in Q_j$.*

- *In this class, $Q = \{P_1, ..., P_N\}$, $Q_i = \{P_j\}$.*
- *$\mathbb{P}_{S\sim P_j}(\Psi(S) \neq j)$ is the probability of error (when $S \sim P_i$).*

With this setup, we obtain the classical reduction from estimation to testing.

Theorem 1 (Reduction from Estimation to Testing). *Let $\{P_1, ..., P_N\} \subseteq P$, and let $\delta = \min_{j\neq k} \rho(\theta(P_j), \theta(P_n))$, Then*

$$R_n^* = \inf_{\widehat{\theta}} \sup_{P} \mathbb{E}_S\left[\Phi \cdot \rho(\theta(P), \widehat{\theta}(S))\right] \geq \Phi\left(\frac{\delta}{2}\right) \inf_{\Psi} \max_{j\in[N]} \mathbb{P}_{S\sim P_j}\left(\Psi(S) \neq j\right).$$

Proof For brevity, $\theta_j = \theta(P_j)$, $\mathbb{P}_j(\cdot) = \mathbb{P}_{S\sim P_j}(\cdot)$, $\mathbb{E}_j[\cdot] = \mathbb{E}_{S\sim P_j}[\cdot]$.
First,

$$R_n^* = \inf_{\widehat{\theta}} \sup_{p\in P} \mathbb{E}_S\left[\ell(\theta, \widehat{\theta})\right] \geq \inf_{\widehat{\theta}} \max_{j\in[N]} \mathbb{E}_j\left[\ell(\theta_j, \widehat{\theta})\right].$$

By Markov's inequality,

$$\mathbb{E}_j\left[\ell(\theta_j, \widehat{\theta})\right] \geq t\mathbb{P}_j\left[\ell(\theta_j, \widehat{\theta}) > t\right].$$

Set $t = \Phi\left(\frac{\delta}{2}\right)$,

$$\begin{aligned} R_n^* &\geq \inf_{\widehat{\theta}} \max_{j\in[N]} \Phi\left(\frac{\delta}{2}\right) \mathbb{P}_j\left(\Phi \circ \rho\left(\theta_j, \widehat{\theta}\right) > \Phi\left(\frac{\delta}{2}\right)\right) \\ &= \Phi\left(\frac{\delta}{2}\right) \inf_{\widehat{\theta}} \max_{j\in[N]} \mathbb{P}_j\left(\rho\left(\theta_j, \widehat{\theta}\right) > \frac{\delta}{2}\right) \qquad \text{(Since } \Phi(\cdot) \text{ is a nondecreasing function)} \end{aligned}$$

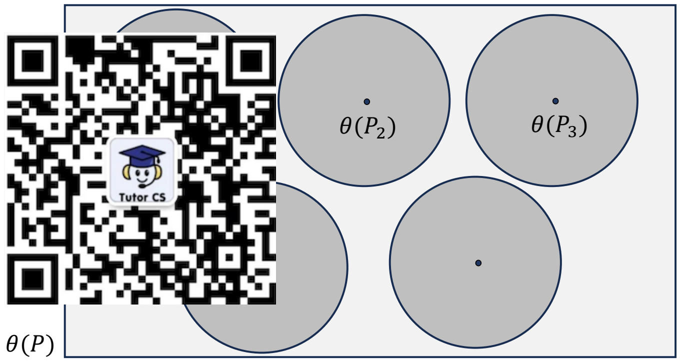

Figure 1: Illustrative figure for Theorem 1. The radius of each circle is $\delta/2$. If N is too large (δ is small), $\Psi(\delta/2)$ will be small. But if N is small (δ is large), $\mathbb{P}_{S\sim P_j}(\Psi(S)\neq j)$ will be small as it may be harder to distinguish between the alternatives.

Given an estimator $\widehat{\theta}$, we define the following test,

$$\Psi_{\widehat{\theta}}(S) = \arg\min_j \rho\left(\widehat{\theta}(S), \theta_j\right)$$

Given the data was generated by P_j, but $\Psi_{\widehat{\theta}} = k \neq j$, then,

$$\begin{aligned}
\delta &\le \rho\left(\theta_j, \theta_k\right) \quad \text{(By definition of } \delta\text{)} \\
&\le \rho\left(\theta_j, \widehat{\theta}\right) + \rho\left(\widehat{\theta}, \theta_k\right) \quad \text{(By triangle inequality)} \\
&\le \rho\left(\theta_j, \widehat{\theta}\right) + \rho\left(\theta_j, \widehat{\theta}\right) \\
&= 2\rho\left(\theta_j, \widehat{\theta}\right). \\
\therefore\ \Psi_{\widehat{\theta}} \neq j &\Rightarrow \rho\left(\theta_j, \widehat{\theta}\right) \ge \frac{\delta}{2}.
\end{aligned}$$

Therefore, $\mathbb{P}_j\left(\rho\left(\widehat{\theta}, \theta_j\right) \ge \frac{\delta}{2}\right) \ge \mathbb{P}_j\left(\Psi_{\widehat{\theta}} \neq j\right)$, and we have,

$$\begin{aligned}
R_n^* &\ge \Phi\left(\frac{\delta}{2}\right) \inf_{\Psi_{\widehat{\theta}}} \max_{j\in[N]} P_j\left(\Psi_{\widehat{\theta}} \neq j\right) \\
&\ge \Phi\left(\frac{\delta}{2}\right) \inf_{\Psi} \max_{j\in[N]} P_j\left(\Psi \neq j\right).
\end{aligned}$$

□

2 Distances/divergences between distributions

Consider two probability distributions, P and Q. Let $p(x)$ and $q(x)$ be their probability density functions. We usually have the following distance and divergence measurements between two probability distributions. They are key ingredients in formulating lower bounds on the performance of inference procedures.

1. KL divergence:

$$\mathrm{KL}(P,Q) = \int \log\left(\frac{dP(x)}{dQ(x)}\right) dP(x) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx\,.$$

2. Total Variation (TV) distance:

$$\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|.$$

3. L_1 distance:

$$||P - Q||_1 = \int |p(x) - q(x)|\, dx\,.$$

4. Helliger distance $H(P,Q)$:

$$\begin{aligned} \mathrm{H}^2(P,Q) &= \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx \\ &= 2 - 2\int \sqrt{p(x)q(x)}\, dx\,. \end{aligned}$$

Also, define affinity:

$$\begin{aligned} ||P \wedge Q|| &= \int \min\left(p(x), q(x)\right) dx \\ &= \int \left(p(x) \wedge q(x)\right) dx \end{aligned}$$

2.1 Inequalities between divergences and product distributions

Here we present a few inequalities and their consequences when applied to product distributions, which will be quite useful for proving our lower bounds. These inequalities will relate to three divergences, i.e. total variation distance, Kullback-Leibler divergence, and Hellinger distance.

1. Since KL-divergence and Hellinger distance both are easier to manipulate on product distributions than is total variation. Consider the product distribution $P^n = \underbrace{P \times \cdots \times P}_{\text{n times}}$ and $Q^n = \underbrace{Q \times \cdots \times Q}_{\text{n times}}$.

   Then,

$$\mathrm{KL}(P^n, Q^n) = n \times \mathrm{KL}(P,Q)$$
$$\mathrm{H}^2(P^n, Q^n) = 2 - 2\Big(1 - \frac{1}{2}H^2(P,Q)\Big)^n$$

2. $\mathrm{TV}(P,Q) = \frac{1}{2}||P - Q||_1 = 1 - ||P \wedge Q||$.
3. $\mathrm{H}^2(P,Q) \le ||P - Q||_1 = 2\mathrm{TV}(P,Q)$.
4. Pinsker's inequality:

$$TV(P,Q) \le \sqrt{\frac{1}{2}KL(P,Q)}.$$

5. $\|P \wedge Q\| \ge \frac{1}{2}\exp(-\mathrm{KL}(P,Q)$.

   We will prove statement 5 below. You will prove the remaining statements in your homework.

Proof

$$
\begin{aligned}
2\|P \wedge Q\| &= 2\int \text{m[illegible]} \\
&\geq 2\int \text{m[illegible]} \min(p(x), q(x))\, dx\Big)^2 \\
&= \int \text{mi[illegible]} \min(p(x), q(x))\, dx\Big) \\
&= \Big(\int \text{m[illegible]} \max(p(x), q(x))\, dx\Big) \\
&\geq \left(\int \sqrt{\min(p(x), q(x)) \max(p(x), q(x))}\, dx\right)^2 \quad \text{(By Cauchy–Schwarz inequality)} \\
&= \left(\int \sqrt{p(x)q(x)}\, dx\right)^2 \\
&= \exp\left(2\log\left(\int \sqrt{p(x)q(x)}\right) dx\right) \\
&= \exp\left(2\log\left(\int p(x)\sqrt{\frac{q(x)}{p(x)}}\right) dx\right) \\
&\geq \exp\left(2\int p(x)\log\left(\sqrt{\frac{q(x)}{p(x)}}\right) dx\right) \quad \left(\text{By Jensen's inequality: } \log\left(\mathbb{E}\left[\sqrt{\frac{q(x)}{p(x)}}\right]\right) \geq \mathbb{E}\left[\log\left(\sqrt{\frac{q(x)}{p(x)}}\right)\right]\right) \\
&= \exp\left(-\int p(x)\log\left(\frac{p(x)}{q(x)}\right) dx\right) \\
&= \exp(-\mathrm{KL}(P, Q))
\end{aligned}
$$

Where the third equality to fourth equality follows by,

$$\int \min(p(x), q(x))\, dx + \int \max(p(x), q(x))\, dx = \int p(x)\, dx + \int q(x)\, dx = 2$$

□

3 Le Cam's Method

Consider this scenario: nature chooses one of a possible set of (say) $k+1$ words, indexed by probability distributions $P_0, P_1, \cdots P_k$ and conditional on nature's choice of the word —the distribution $P^* \in \{P_0, \cdots P_k\}$ chosen—we observe data S drawn from P^*. Intuitively, it will be difficult to decide which distribution P_i is the true P^* if all the distributions are similar—the distance/divergence between the P_i is small and easy if the distance/divergence between the distribution P_i is large.

The simplest case is when there are only two possible distributions, P_0 and P_1, and our goal is to make a decision on whether P_0 and P_1 are the distribution generating data we observe. Suppose that nature chooses one of the distributions P_0 or P_1 at random, and let $V = \{0, 1\}$ index the choice. Conditional on $V = v$, we then observe samples S drawn from P_v, then, for any test $\Psi : \mathcal{S} \Rightarrow \{0, 1\}$, the probability of error is then

$$P(\Psi(S) \neq V) = \frac{1}{2}P_0(\Psi \neq 0) + \frac{1}{2}P_1(\Psi \neq 1)$$

Now, we introduce the Neyman-Pearson Test, and then we will show that it can minimize the sum of errors.

3.1 Neyman-Pearson Test

Given a binary hypothe[illegible]ernatives P_0 and P_1 with densities p_0 and p_1, let S denote an i.i.d dataset. Then, the [illegible]s the form:

$$[illegible] \begin{cases} 0 & \text{if } p_0(S) \geq p_1(S) \\ 1 & \text{if } p_0(S) < p_1(S) \end{cases}$$

Lemma 1. *For any ot[illegible]-Pearson test minimizes the sum of errors. That is,* $\forall \Psi$,

$$[illegible] 1) \geq P_0(\Psi_{\text{NP}} \neq 0) + P_1(\Psi_{\text{NP}} \neq 1)$$

where $P_0(\Psi \neq 0)$ *is actually the* $\mathbb{P}_{S\sim P_0}(\Psi \neq 0)$*, for short.*

Proof

$$\begin{aligned}
&P_0(\Psi \neq 0) + P_1(\Psi \neq 1) \\
&= P_0(\Psi = 1) + P_1(\Psi = 0) \\
&= \int_{\Psi=1} p_0(x)\,dx + \int_{\Psi=0} p_1(x)\,dx \\
&= \int_{\Psi=1,\Psi_{\text{NP}}=1} p_0(x)\,dx + \int_{\Psi=1,\Psi_{\text{NP}}=0} p_0(x)\,dx + \int_{\Psi=0,\Psi_{\text{NP}}=0} p_1(x)\,dx + \int_{\Psi=0,\Psi_{\text{NP}}=1} p_1(x)\,dx \\
&\geq \int_{\Psi=1,\Psi_{\text{NP}}=1} p_0(x)\,dx + \int_{\Psi=1,\Psi_{\text{NP}}=0} p_1(x)\,dx + \int_{\Psi=0,\Psi_{\text{NP}}=0} p_1(x)\,dx + \int_{\Psi=0,\Psi_{\text{NP}}=1} p_0(x)\,dx \quad \text{(by Definition of NP Test)} \\
&= \int_{\Psi=1} p_0(x)\,dx + \int_{\Psi=0} p_1(x)\,dx \\
&= P_0(\Psi_{\text{NP}} = 1) + P_1(\Psi_{\text{NP}} = 0) \\
&= P_0(\Psi_{\text{NP}} \neq 0) + P_1(\Psi_{\text{NP}} \neq 1)
\end{aligned}$$

□

Next, we show the connection between hypothesis testing and total variation distance and later use this to yield lower bounds on minimax error by Le Cam's Method.

Corollary 1. *For any hypothesis test* Ψ*, we have,*

$$P_0(\Psi \neq 0) + P_1(\Psi \neq 1) \geq \|P_0 \wedge P_1\| \;= 1 - \text{TV}(P_0, P_1) \;\geq \frac{1}{2}\exp(-\text{KL}(P_0, P_1))$$

From this Corollary, we can see that the smaller the KL divergence or TV distance between P_0 and P_1, i.e., the more similar P_0 and P_1, the larger the testing error, which also verifies the intuition of our introduced scenario.

Proof

$$\begin{aligned}
P_0(\Psi \neq 0) + P_1(\Psi \neq 1) &= \int_{\Psi_{\text{NP}}=1} p_0(x)\,dx + \int_{\Psi_{\text{NP}}=0} p_1(x)\,dx \\
&= \int_{P_0 \leq P_1} p_0(x)\,dx + \int_{P_1 < P_0} p_1(x)\,dx \quad \text{(by Definition of NP test)} \\
&= \int \min\left(p_0(x), p_1(x)\right) dx \\
&= \|P_0 \wedge P_1\|
\end{aligned}$$

Other parts of inequalities come from Section 2. $\square$

Putting them together, am's Method can yield lower bounds for minimax estimation.

Theorem 2 (Le Cam's *P, let $\delta = \rho\big(\theta(P_0), \theta(P_1)\big)$ and let S be an i.i.d dataset of n points, then,*

$$\frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\|P_0^n \wedge P_1^n\|$$

Proof

$$\begin{aligned} R_n^* &\geq \Phi\left(\frac{\delta}{2}\right) \inf_{\Psi} \max_{j \in \{0,1\}} P_j^n(\Psi \neq j) \quad \text{(By Theorem 1)} \\ &\geq \Phi\left(\frac{\delta}{2}\right) \inf_{\Psi}\left[\frac{1}{2}\left(P_0^n(\Psi \neq 0) + P_1^n(\Psi \neq 1)\right)\right] \quad \text{(As max is larger than the average.)} \\ &\geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\|P_0^n \wedge P_1^n\| \quad \text{(By Corollary 1)} \end{aligned}$$

$\square$