**CS861: Theoretical Foundations of Machine Learning** Lecture 9 - 09/23/2024
University of Wisconsin–Madison, Fall 2024

# Lecture 09: Le Cam's Method

*Lecturer: Kirthevasan Kandasamy* *Scribed by: Jingyun Jia, Xinta Yang*



In this lecture, we begin by recapping the minimax lower bounds from Reduction to Testing Theorem and the KL divergence. We then introduce Le Cam's method to obtain a lower bound by considering binary hypothesis testing specifically. Finally, we obtain a specific form of lower bound as a consequence of Le Cam's method and then see its applications through some examples of mean estimation and regression.

## 1 Properties of Divergences (cont'd)

We begin by proving statement 5 from last lecture below.

**Proposition 1.** *The relation* $\|P \wedge Q\| \geq \frac{1}{2}\exp(-\mathrm{KL}(P,Q))$ *holds.*

**Proof**

$$\begin{aligned}
2\|P \wedge Q\| &= 2\int \min(p(x), q(x))\, dx \\
&\geq 2\int \min(p(x), q(x))\, dx - \left(\int \min(p(x), q(x))\, dx\right)^2 \\
&= \int \min(p(x), q(x))\, dx \left(2 - \int \min(p(x), q(x))\, dx\right) \\
&= \left(\int \min(p(x), q(x))\, dx\right)\left(\int \max(p(x), q(x))\, dx\right) \qquad (1)
\end{aligned}$$

because $\min(p(x), q(x)) + \max(p(x), q(x)) = p(x) + q(x)$ and $\int p(x)dx = \int q(x)dx = 1$. With Cauchy-Schwartz inequality, i.e. $\int |fg|dx \leq \sqrt{\int f^2 dx} \cdot \sqrt{\int g^2 dx}$,

$$(1) \geq \left(\int \sqrt{\min(p(x), q(x))\max(p(x), q(x))}\, dx\right)^2 \qquad (2)$$

Continue, with $\min(p,q) \cdot \max(p,q) = pq$,

$$\begin{aligned}
(2) &= \left(\int \sqrt{p(x)q(x)}\, dx\right)^2 \\
&= \exp\left(2\log\left(\int p(x)\sqrt{\frac{q(x)}{p(x)}}\right) dx\right) \qquad \text{(property of exp and log)} \\
&\geq \exp\left(2\int p(x)\log\left(\sqrt{\frac{q(x)}{p(x)}}\right) dx\right) \qquad (3)
\end{aligned}$$

where inequality (3) follows from Jensen's inequality

$$\log\left(\mathbb{E}\left[\sqrt{\frac{q(x)}{p(x)}}\right]\right) \geq \mathbb{E}\left[\log\left(\sqrt{\frac{q(x)}{p(x)}}\right)\right]$$

Finally, by property of log again,

$$\begin{aligned}(3) &= \exp\left(-\int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx\right)\\ &= \exp(-\mathrm{KL}(P,Q))\end{aligned}$$

□

## 2 Le Cam's Method

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

**Definition 1** (Neyman-Pearson Test). *Given a binary hypothesis test between two alternatives $P_0$ and $P_1$ with densities $p_0$ and $p_1$, let $S$ denote an i.i.d dataset. Then, the Neyman-Pearson test is the form:*

$$\psi_{\mathrm{NP}}(S) = \begin{cases} 0 & \text{if } p_0(S) \geq p_1(S) \\ 1 & \text{if } p_0(S) < p_1(S) \end{cases}$$

**Lemma 1.** *For any other test $\psi$, the Neyman-Pearson test minimizes the sum of errors. That is, $\forall\psi$,*

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) \geq P_0(\psi_{\mathrm{NP}} \neq 0) + P_1(\psi_{\mathrm{NP}} \neq 1)$$

*where $P_0(\psi \neq 0) = \mathbb{P}_{S\sim P_0}(\psi \neq 0)$.*

**Proof**

$$\begin{aligned}
&P_0(\psi \neq 0) + P_1(\psi \neq 1)\\
&= P_0(\psi = 1) + P_1(\psi = 0)\\
&= \int_{\psi=1} p_0(x)\,dx + \int_{\psi=0} p_1(x)\,dx\\
&= \int_{\psi=1,\psi_{\mathrm{NP}}=1} p_0(x)\,dx + \int_{\psi=1,\psi_{\mathrm{NP}}=0} p_0(x)\,dx + \int_{\psi=0,\psi_{\mathrm{NP}}=0} p_1(x)\,dx + \int_{\psi=0,\psi_{\mathrm{NP}}=1} p_1(x)\,dx\\
&\geq \int_{\psi=1,\psi_{\mathrm{NP}}=1} p_0(x)\,dx + \int_{\psi=1,\psi_{\mathrm{NP}}=0} p_1(x)\,dx + \int_{\psi=0,\psi_{\mathrm{NP}}=0} p_1(x)\,dx + \int_{\psi=0,\psi_{\mathrm{NP}}=1} p_0(x)\,dx \qquad (4)\\
&= \int_{\psi=1} p_0(x)\,dx + \int_{\psi=0} p_1(x)\,dx = P_0(\psi_{\mathrm{NP}} = 1) + P_1(\psi_{\mathrm{NP}} = 0) = P_0(\psi_{\mathrm{NP}} \neq 0) + P_1(\psi_{\mathrm{NP}} \neq 1)
\end{aligned}$$

where the inequality (4) is by how Neyman-Pearson lemma is setup. □

One important note, for $S = (X_1, ..., X_n)$, densities $p_0, p_1$ are the joint distribution of those $n$ random variables, i.e. $p_0^{(n)}(x_1, ..., x_n) = p_0(x_1)\cdots p_0(x_n)$, where $p_0$ is the density of a single $X$ because $X_i$'s are i.i.d.

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** (Bretagnolle-Huber inequality)**.** *For any hypothesis test $\psi$, we have,*

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

**Proof**

$$\begin{aligned}
P_0(\psi \neq 0) + P_1(\psi \neq 1) &\geq \int_{\psi_{\mathrm{NP}}=1} p_0(x)\,dx + \int_{\psi_{\mathrm{NP}}=0} p_1(x)\,dx \\
&\geq \int_{p_0 \leq p_1} p_0(x)\,dx + \int_{p_1 < p_0} p_1(x)\,dx \quad \text{(Definition of NP test)} \\
&= \int \min\left(p_0(x), p_1(x)\right) dx \\
&= \|P_0 \wedge P_1\| \\
&\geq \frac{1}{2}\exp(-\mathrm{KL}(P_0, P_1))
\end{aligned}$$

where the first inequality is due to the Neyman-Pearson lemma above, and the last inequality is due to the relation between TV distance and KL divergence. □

Same as in the Neyman-Pearson test, $P_0, P_1$ are *joint* distributions if $S$ contains more than one sample point. 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. For binary hypothesis testing, we can simply combine "max≥avg" with the BH inequality and get the nice result below:

$$\begin{aligned}
\inf_{\psi} \sup_{j \in \{0,1\}} P_j(\psi(S) \neq j) &\geq \inf_{\psi} \left(\frac{1}{2}P_0(\psi(S) \neq 0) + \frac{1}{2}P_1(\psi(S) \neq 1)\right) && (5) \\
&\geq \frac{1}{2}\|P_0 \wedge P_1\| && (6)
\end{aligned}$$

Combining all the results above, we can now show Le Cam's method for estimation problems.

**Theorem 1** (Le Cam's Method for Estimation Problems)**.** *Let $P_0, P_1 \in P$, let $\delta = \rho\big(\theta(P_0), \theta(P_1)\big)$ and $S$ be drawn i.i.d. from the distribution in $P$. Then,*

$$\begin{aligned}
R_n^* &= \inf_{\widehat{\theta}} \sup_{P} \mathbb{E}_S \left[\Phi \circ \rho(\theta(P), \widehat{\theta}(S))\right] \\
&\geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\|P_0^n \wedge P_1^n\|
\end{aligned}$$

*(here we directly write $P_0^n, P_1^n$ to distinguish from $P_0, P_1$ because $S$ contains $n$ points, whereas before we omit the superscript $n$ and emphasize they are joint distributions since we don't know the sample size.)*

**Proof** We use the Reduction to Test Theorem (previous lecture) and inequality (6) to prove the theorem.

$$R_n^* \geq \Phi\left(\frac{\delta}{2}\right) \cdot \inf_{\psi} \max_{j \in \{0,1\}} P_j^n(\psi \neq j) \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\|P_0^n \wedge P_1^n\|$$

□

**Corollary 2** (Lower bound risk by a constant)**.** *Let $S$ has sample size $n$, i.i.d. from some $P \in \mathcal{P}$. Pick $P_0, P_1 \in \mathcal{P}$. If both the following hold,*

$$\rho(\theta(P_0), \theta(P_1)) \geq \delta \tag{7}$$

$$KL(P_0^n, P_1^n) \leq \frac{1}{n} \log 2 \tag{8}$$

*then $R_n^* \geq \frac{1}{8} \cdot \Phi\left(\frac{\delta}{2}\right)$.*

**Proof** Combine the properties of various divergences (listed at the beginning of this document) and the two assumptions here,

$$||P_0^n \wedge P_1^n|| \geq \frac{1}{2} e^{-KL(P_0^n, P_1^n)} = \frac{1}{2} e^{-nKL(P_0, P_1)} \geq \frac{1}{4}$$

Invoke the theorem 1 above, we obtain the desired result. □

**Remark** The first assumption (5) says two distributions are distinguishable, which is the essential part for the Reduction to Test theorem; while the second assumption (6) says such difference is not too much, and we call such phenomenon *statistically indistinguishable*.

# 3 Examples for Le Cam's Method

To be more precise, the main tool here to lower bound $R_n^*$ is Corollary 2 above.

## 3.1 Family of normal distributions

Take $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, where $\sigma^2$ is known. $S = \{x_1, ..., x_n\}$ are $n$ i.i.d. samples. We focus on the estimator $\theta(P) = \mu$, and use $\Phi(t) = t^2$, $\rho(\theta_1, \theta_2) = |\theta_1 - \theta_2|$ for the distances. Take two distributions from the family, $P_0 = \mathcal{N}(0, \sigma^2)$, $P_1 = \mathcal{N}(\delta, \sigma^2)$, then they have $\rho(\theta(P_0), \theta(P_1)) = \delta$, giving the first assumption of Corollary 2. For the second assumption, because

$$KL(P_0, P_1) = \frac{\delta^2}{2\sigma^2} \tag{9}$$

we pick

$$\delta = \sigma \sqrt{\frac{2 \log 2}{n}} \tag{10}$$

Now, invoke Corollary 2, we obtain

$$R_n^* \geq \frac{1}{8} \left(\frac{\delta}{2}\right)^2 = \frac{\log 2}{16} \cdot \frac{\sigma^2}{n}$$

Finally, because sample mean has risk $\sigma^2/n$, we conclude it's minimax rate.

## 3.2 Distribution with finite support

Take $\mathcal{P} = \{P : \text{supp}(P) \subset [0, 1]\}$ and use the same $\theta, \Phi, \rho$ as in the previous example. Take two distributions from $\mathcal{P}$, $P_0 = \text{Ber}(1/2 + \delta)$ and $P_1 = \text{Ber}(1/2)$. From the ending example of previous lecture,

$$KL(P_0, P_1) \leq \frac{(1/2 + \delta - 1/2)^2}{1/2(1 - 1/2)} = 4\delta^2$$

For the second assumption of Corollary 2, pick

$$\delta = \frac{1}{2} \sqrt{\frac{\log 2}{n}}$$

Therefore, we conclude

$$R_n^* \geq \frac{\log 2}{128} \cdot \frac{1}{n}$$

Finally, because sample mean achieves the risk $1/4n$, it is the minimax rate.

### 3.3 A simplified regression problem

Let $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \overset{i.i.d.}{\sim} \text{Unif}([0,1])$ and $y_i$ is drawn from a distribution with mean $f(x_i)$ and variance bounded by $\sigma^2$. We also assume some regularity of the underlying $f$. Formally, the distribution family we work with is

$$\mathcal{P} = \{P_{x,y} : P_x = \text{Unif}([0,1]), \mathbb{E}[Y|X = x] = f(x), f \text{ L-Lipschitz and bounded in } [0,1], \text{Var}[Y|X = x] \leq \sigma^2\}$$

We estimate $\theta(P) = f(1/2)$, i.e. we only care about the middle point of $f$ instead of the whole picture of $f(x)$. Finally, we use the same $\Phi, \rho$ as before.

Now time to solve the lower bound. Pick two underlying functions $f_0, f_1$ so the distributions

$$P_0 : Y|X = x \sim \mathcal{N}(f_0(x), \sigma^2)$$

$$P_1 : Y|X = x \sim \mathcal{N}(f_1(x), \sigma^2)$$

Remember in $\mathcal{P}$, we don't force the conditional distribution to be normal, but picking normal distribution certainly satisfies the condition of $\mathcal{P}$. Next, the first assumption of Corollary 2 requires

$$|f_0(1/2) - f_1(1/2)| \geq \delta$$

and $f$'s are required to be L-Lipschitz, so we define $f_0 \equiv 0$; and $f_1 = 0$ outside $[1/2 - \delta, 1/2 + \delta]$, $f_1(1/2) = \delta$, and linear elsewhere (diagram for $f_1$ below.) Rest is left for next lecture.

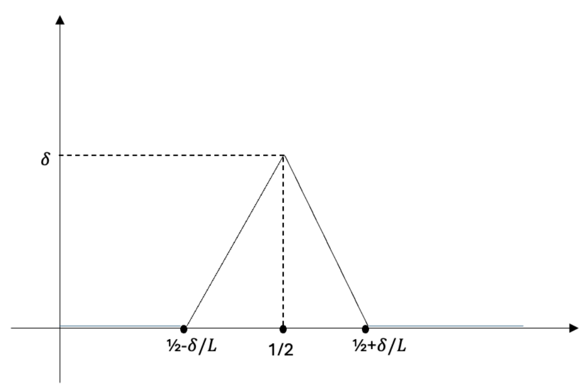


**Figure 1:** Definition of $f_1$, where $f_1 = 0$ outside $[1/2 - \delta, 1/2 + \delta]$.


## Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Haoyue Ba, Ying Fu and Deep Patel.