CS861: Theoretical Foundations of Machine Learning Lecture 8 - 09/20/2024

University of Wisconsin–Madison, Fall 2024

Lecture 08: Minimax Theory and Hypothesis Testing

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

In this lecture, we begin by recapping the minimax lower bounds for point estimation from the previous class. Then we will introduce the minimax theory in the general estimation problem and the concept of hypothesis testing, along with the theorem of the reduction from estimation to testing. Finally, we will briefly discuss different distances and divergences between distributions.

1 Minimax Theory

We now extend the point estimation to general estimation problems. An estimation problem has the following components:

1. A known family of distributions $\mathcal{P}$
2. A dataset S of n i.i.d points drawn from the unknown distribution $P \in \mathcal{P}$
3. A function(parameter) $\theta : \mathcal{P} \to \Theta$. We wish to estimate $\theta(P)$ from S, e.g. the population mean$\theta(P) = \mathbb{E}_{X\sim P}[X]$ and $\theta(P) =$ the density of P.
4. An estimator $\hat{\theta} = \widehat{\theta}(S) \in \Theta$
5. A loss function ℓ, $\ell = \Phi \circ \rho$, satisfies the following conditions:
   - $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing.
   - $\rho : \Theta \times \Theta \to \mathbb{R}_+$ is a pseudo-metric. That is, it satisfies the following properties for all $\theta_1, \theta_2, \theta_3 \in \Theta$,
     (i) $\rho(\theta_1, \theta_1) = 0$,
     (ii) $\rho(\theta_1, \theta_2) = \rho(\theta_2, \theta_1)$,
     (iii) $\rho(\theta_1, \theta_2) \le \rho(\theta_1, \theta_3) + \rho(\theta_3, \theta_2)$.

   When we estimate $\theta(P)$ with $\widehat{\theta}$, the loss is $\ell(\theta(P), \widehat{\theta}) = \Phi(\rho(\theta(P), \widehat{\theta}))$.
6. The risk of an estimator $\widehat{\theta}$ is,

$$R(P, \widehat{\theta}) = \mathbb{E}_{S\sim P^n}\left[\Phi \circ \rho(\theta(P), \widehat{\theta}(S))\right] = \mathbb{E}_{S\sim P^n}\left[\Phi \circ \rho(\theta, \hat{\theta})\right].$$

Note that, as before we have overloaded notation so that θ denotes the parameter $\theta \in \Theta$ and the function $\theta : \mathcal{P} \to \Theta$. Similarly, $\widehat{\theta}$ denotes the estimate $\widehat{\theta} \in \Theta$ and the estimator, which maps the data to Θ.

Definition 1. *We can now define the minimax risk R^* as follows,*

$$R^* = \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_S R(P, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_S\left[\Phi \circ \rho\left(\theta(P), \widehat{\theta}(S)\right)\right].$$

Example 1. Normal mean estimation Let $S = \{X_1, .., X_n\}$ drawn i.i.d. from $N(\mu, \sigma^2)$, where σ^2 is known. Here, $\mathcal{P} = \{N(\mu, \sigma^2); \mu \in \mathbb{R}\}$. We wish to estimate $\theta(P) = \mathbb{E}_{X\sim P}[X]$, so $\Theta = \mathbb{R}$. If we use the squared loss $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, then $\rho = |\theta_1 - \theta_2|$ and $\Phi(t) = t^2$.

Example 2. (Regression) Here, $\mathcal{P}$ is the set of all distributions with support on $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is the input space. = The parameter space $\Theta = \{h : \mathcal{X} \to \mathbb{R}\}$, is the class of functions mapping $\mathcal{X}$ to $\mathcal{R}$. We wish to estimate the regression function, which is given by

$$\theta(P)(\cdot) = \mathbb{E}[Y|X = \cdot]$$

If we use the L_2 loss, $\ell(\theta_1, \theta_2) = \int_{\mathcal{X}} (\theta_1(x) - \theta_2(x))^2 \, dx$, then, we have

$$\rho(\theta_1, \theta_2) = \sqrt{\int_{\mathcal{X}} (\theta_1(x) - \theta_2(x))^2 \, dx} \triangleq \|\theta_1 - \theta_2\|_2, \quad \text{and} \quad \Phi(t) = t^2.$$

2 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 2 (Hypothesis Test)**.** *Let $\mathcal{Q}$ be a class of distributions, and let $\mathcal{Q}_1, \mathcal{Q}_2, .., \mathcal{Q}_N$ be a partition of $\mathcal{Q}$. Let S be a dataset drawn from some distribution $P \in \mathcal{Q}$. 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 $P \in \mathcal{Q}_j$.*

In this class, we will focus on cases where $\mathcal{Q} = \{P_1, ..., P_N\}$, $\mathcal{Q}_i = \{P_j\}$. $\mathbb{P}_{S\sim P_j}(\Psi(S) \neq j)$ is the probability of error (when $S \sim P_i$).

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

Theorem 3 (Reduction from Estimation to Testing)**.** *Let $\{P_1, ..., P_N\} \subseteq \mathcal{P}$, and let $\delta = \min_{j\neq k} \rho(\theta(P_j), \theta(P_k))$, Then*

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

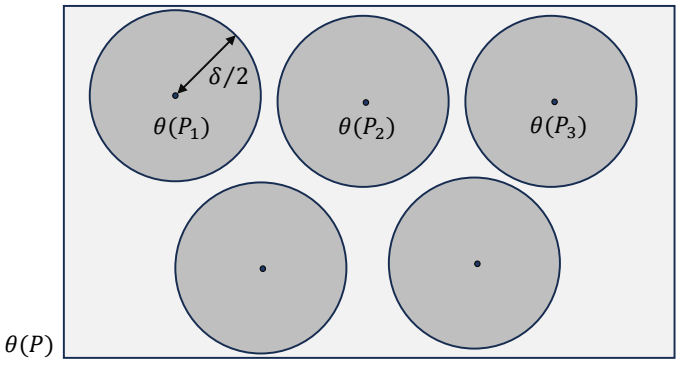

Figure 1: Illustrative figure for Theorem 3. The radius of each circle is $\delta/2$. If N is too large (δ is small), $\Psi(\delta/2)$ will be small. But if N is small (δ is large), $\mathbb{P}_{S\sim P_j}(\Psi(S) \neq j)$ will be small as it may be harder to distinguish between the alternatives.

Proof First observe

$$R^* = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_S[\Phi \circ \rho(\theta(P), \widehat{\theta}(S))] \tag{1}$$

$$\geq \inf_{\widehat{\theta}} \max_{j=1,\ldots,N} \mathbb{E}_j[\Phi \circ \rho(\theta_j, \widehat{\theta}(S))] \tag{2}$$

because instead of taking sup over all possible distributions P, we only take sup over the selected subset $\{P_1, ..., P_N\}$, which gives a smaller supremum (notation details: $\theta_j = \theta(P_j), P_j(\cdot) = P_{S\sim P_j}(\cdot), \mathbb{E}_j[\cdot] = \mathbb{E}_{S\sim P_j}[\cdot]$). Now, Markov's inequality gives

$$\mathbb{E}_j[\Phi \circ \rho(\theta_j, \widehat{\theta}(S))] \geq t \cdot P_j(\Phi \circ \rho(\theta_j, \widehat{\theta}(S)) \geq t)$$

Substitute into (2) with $t \triangleq \Phi(\delta/2)$, we get

$$R^* \geq \inf_{\widehat{\theta}} \max_{j=1,\ldots,N} \Phi\left(\frac{\delta}{2}\right) \cdot P_j\left(\Phi \circ \rho(\theta_j, \widehat{\theta}) \geq \Phi\left(\frac{\delta}{2}\right)\right) \tag{3}$$

$$= \inf_{\widehat{\theta}} \max_{j=1,\ldots,N} \Phi\left(\frac{\delta}{2}\right) \cdot P_j\left(\rho(\theta_j, \widehat{\theta}) \geq \frac{\delta}{2}\right) \tag{4}$$

where (4) follows by Φ being non-decreasing.

Next, we continue by pulling in hypothesis test perspective. Given any estimator $\widehat{\theta}$, define a corresponding hypothesis test $\Psi_{\widehat{\theta}}(S) = \arg\min_{j=1,\ldots,N} \rho(\theta_j, \widehat{\theta}(S))$. Claim event $\{\Psi_{\widehat{\theta}}(S) \neq j\}$ implies event $\{\rho(\theta_j, \widehat{\theta}(S)) \geq \delta/2\}$ (i.e. the first set is a subset of the second). Suppose not, then we would have

$$\rho(\theta_j, \widehat{\theta}(S)) < \delta/2$$

while

$$\Psi_{\widehat{\theta}}(S) = k \neq j$$

which then give

$$\rho(\widehat{\theta}(S), \theta_k) \leq \rho(\widehat{\theta}(S), \theta_j) < \delta/2$$

by the definition of Ψ. By triangle inequality,

$$\rho(\theta_j, \theta_k) \leq \rho(\theta_j, \widehat{\theta}(S)) + \rho(\widehat{\theta}(S), \theta_k) < \delta$$

But this cannot happen because our choice of θ's insist $\rho(\theta_j, \theta_k) \geq \delta$ for any $k \neq j$. Thus our proof by contradiction works and we get

$$P_j(\Psi_{\widehat{\theta}}(S) \neq j) \leq P_j\left(\rho(\theta_j, \widehat{\theta}(S)) \geq \frac{\delta}{2}\right) \tag{5}$$

Substitute (5) back into (4) yields

$$R^* \geq \inf_{\widehat{\theta}} \max_{j=1,\ldots,N} \Psi\left(\frac{\delta}{2}\right) \cdot P_j(\Psi_{\widehat{\theta}}(S) \neq j) \tag{6}$$

$$= \inf_{\substack{\Psi_{\widehat{\theta}} \\ \text{for all } \widehat{\theta}}} \max_{j=1,\ldots,N} \Psi\left(\frac{\delta}{2}\right) \cdot P_j(\Psi_{\widehat{\theta}}(S) \neq j) \tag{7}$$

$$\geq \Psi\left(\frac{\delta}{2}\right) \cdot \inf_{\text{all } \Psi} \max_{j=1,\ldots,N} P_j(\Psi(S) \neq j) \tag{8}$$

where (7) is because the quantity only involves $\Psi_{\widehat{\theta}}$, so it does not matter if I put $\widehat{\theta}$ or $\Psi_{\widehat{\theta}}$ for the infimum. And (8) is because $\{\Psi_{\widehat{\theta}} : \text{ all } \widehat{\theta}\} \subset \{\text{all possible hypothesis tests}\}$, and infimum over a larger set gets smaller. This concludes the proof. □

Going forward, we will apply this theorem to develop the LeCam (via binary hypothesis testing) and Fano (via multiple hypothesis testing) methods.

3 Digression: Distances / Divergence between Distributions

Let P, Q be two probability distributions with density functions $p(x), q(x)$ respectively. The following are some important distances / divergence. Here I use $\int$ to mean $\int_{\mathbb{R}}$, $\int_{\mathbb{R}^n}$ or $\int_{\Omega}$ if there is no confusion.

1. **KL-divergence (Kullback-Leibler)**

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

2. **Total Variation (TV) distance**

$$TV(P, Q) = \sup_A |P(A) - Q(A)|$$

where $A \subset \Omega$ is any event set.

3. L_1 **distance**

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

4. **Hellinger distance**

$$H^2(P, Q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx = 2 - 2\int \sqrt{p(x)q(x)} dx$$

Finally, we also define the following **affinity measure** between two distributions P, Q:

$$||P \wedge Q|| = \int \min(p(x), q(x)) dx$$

. Notice affinity is always in between 0 and 1, and if p, q are close, think about the extreme case where $p = q$, then affinity measure will be close to 1; on the other hand, if p, q looks very different, think about the case when p is high, q is close to 0 (and the other way around), then affinity is close to 0.

Here are some useful facts about the quantities above:

1. For product distributions, i.e. think about n i.i.d. sample points,

$$KL(P^n, Q^n) = n \cdot KL(P, Q)$$

because under P^n the density function is $p(x_1)p(x_2)\cdots p(x_n)$ by i.i.d. and then use property of log;

$$H^2(P^n, Q^n) = 2 - 2\left(1 - \frac{1}{2}H^2(P, Q)\right)^n$$

same idea here.

2.

$$TV(P, Q) = \frac{1}{2}||P - Q||_1 = 1 - ||P \wedge Q||$$

3.

$$H^2(P, Q) \leq ||P - Q||_1 = 2TV(P, Q)$$

4. (Pinsker's inequality)

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

5.

$$||P \wedge Q|| \geq \frac{1}{2} \cdot e^{-KL(P,Q)}$$

Proof for (1)-(4) in the list above is for HW2, and (5) is left for next lecture. To end the lecture, we provide couple examples.

1. $P = \mathcal{N}(\mu_1, \sigma^2)$ and $Q = \mathcal{N}(\mu_2, \sigma^2)$, then

$$\begin{aligned}
KL(P,Q) &= \int \log \frac{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}} \cdot p(x)dx \\
&= \frac{1}{2\sigma^2} \int \left[(x-\mu_2)^2 - (x-\mu_1)^2\right] \cdot p(x)dx \\
&= \frac{1}{2\sigma^2} \left[\int (x-\mu_2)^2 p(x)dx - \mathrm{Var}_{X\sim P}(X)\right] \\
&= \frac{1}{2\sigma^2} \left[\int \left[(x-\mu_1)^2 + (\mu_1-\mu_2)^2 + 2(x-\mu_1)(\mu_1-\mu_2)\right] p(x)dx - \mathrm{Var}_{X\sim P}(X)\right] \\
&= \frac{1}{2\sigma^2} \left[\mathrm{Var}_{X\sim P}(X) + (\mu_1-\mu_2)^2 + 2(\mu_1-\mu_2)\int (x-\mu_1)p(x)dx - \mathrm{Var}_{X\sim P}(X)\right] \\
&= \frac{1}{2\sigma^2}(\mu_1-\mu_2)^2
\end{aligned}$$

   where the idea is to repeatedly use $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[(X-\mu_1)^2] = \int (x-\mu_1)^2 p(x)dx$ for example in $X \sim P$; and $\mathbb{E}[X] = \mu_1$ for $X \sim P$ to make the integral become 0 in the second to last equality.

2. $P = \mathrm{Ber}(p)$ and $Q = \mathrm{Ber}(q)$, where "Ber" stands for Bernoulli distribution. Then, use probability mass function (p.m.f.) for discrete random variable here and recall $p.m.f.(x) = p^x \cdot (1-p)^{1-x}$ for $X \sim \mathrm{Ber}(p)$, we get

$$\begin{aligned}
KL(P,Q) &= \sum_{x=0,1} \log \frac{p^x(1-p)^{1-x}}{q^x(1-q)^{1-x}} \cdot p^x(1-p)^{1-x} \\
&= (1-p)[\log(1-p) - \log(1-q)] + p(\log p - \log q)
\end{aligned}$$

   But such quantity is not too easy to work with. Use Pinsker's inequality above, we have $KL(P,Q) \geq 2TV^2(P,Q)$, and $TV(P,Q) = \frac{1}{2}\sum_x |p(x) - q(x)| = |p-q|$ for discrete P, Q (thus use summation instead of integral on the right hand side), so

$$KL(P,Q) \geq 2(p-q)^2$$

   Next, going back to the precise $KL(P,Q)$ above, write in the form of

$$\log \frac{1-p}{1-q} + p \cdot \log \frac{p(1-q)}{q(1-p)}$$

   now use the inequality $\log x \leq x - 1$, we get

$$KL(P,Q) \leq \frac{(p-q)^2}{q(1-q)(1-p)}$$

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Xindi Lin, Zhihao Zhao, Haoyue Bai and Ying Fu.