CS861: Theoretical Foundations of Machine Learning Lecture 10 - 09/25/2024

University of Wisconsin–Madison, Fall 2024

Lecture 10: Le Cam's Method Examples, Information Theory Review

Lecturer: Kirthevasan Kandasamy *Scribed by: Thomas Zeng, Daisuke Yamada*

We have explored the idea of *minimax optimality* and tried to obtain lower bounds for *minimax risk* by reducing estimation problems to hypothesis testing. Specifically in the previous lecture, we defined the *Le Cam's method* to reduce point estimation to binary hypothesis testing and proceeded to establish the lower bound. In this lecture, we continue with the examples of Le Cam's method to illustrate its consequences. We then start reviewing materials from information theory, which will be necessary in the coming lectures.

1 A Simplified Regression Problem

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \overset{\text{i.i.d}}{\sim} \text{Unif}([0,1])$ and y_i is drawn independently from a distribution with mean $f(x_i)$ and variance bounded by σ^2. We assume that $f : [0,1] \to [0,1]$, is L-Lipschitz. That is,

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

Our goal is to estimate $\theta(P) := f(1/2)$ (i.e., rather than the entire f on [0,1]). Hence, this is a point estimation problem.

1.1 Lower Bound

We first obtain the lower bound. To this end, we construct two L-Lipschitz functions f_0 and f_1 such that

$$P_0 : Y|X = x \sim \mathcal{N}(f_0(x), \sigma^2)$$
$$P_1 : Y|X = x \sim \mathcal{N}(f_1(x), \sigma^2).$$

Recall that we want the gap between two parameters to be large, i.e. $\delta = |f_0(1/2) - f_1(1/2)|$, while ensuring that the two distributions are hard to distinguish, i.e. small $\text{KL}(P_0, P_1)$. We choose the following f_0 and f_1:

$$f_0(x) := 0 \ \ \forall\, x \in [0,1]$$
$$f_1(x) := \begin{cases} L(x - (1/2 - \delta/L)), & \text{if } x \in [1/2 - \delta/L, 1/2) \\ L(1/2 + \delta/L - x), & \text{if } x \in [1/2, 1/2 + \delta/L) \\ 0, & \text{else} \end{cases}$$

The function f_1 is in Figure 1.

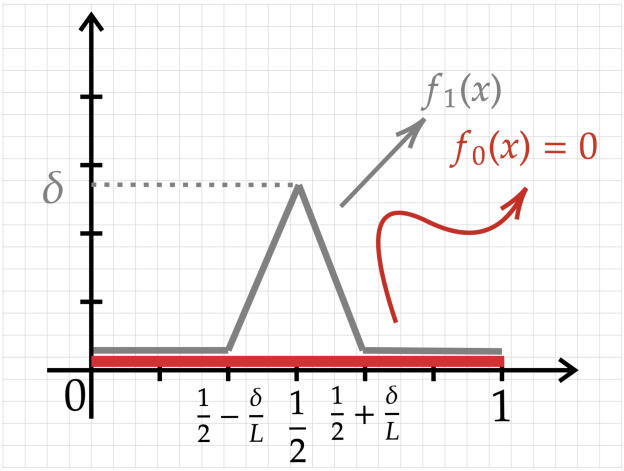

Figure 1: Here, we'd need $\delta/L \leq 1/2$ for f_1 to be well-defined.

For obtaining the lower bound as derived in Corrollary 2 from previous lecture, recall that we need to choose δ such that $KL(P_0, P_1) \leq \frac{\log 2}{n}$. So, we first compute $KL(P_0, P_1)$:

$$\begin{aligned}
\mathrm{KL}(P_0, P_1) &= \int_0^1 \int_{-\infty}^{\infty} P_0(x,y) \log\Big(\frac{P_0(x,y)}{P_1(x,y)}\Big) dy dx \\
&= \int_0^1 \int_{-\infty}^{\infty} P_0(y|x) \underbrace{P_0(x)}_{=1} \log\Big(\frac{P_0(y|x)\overbrace{P_0(x)}^{=1}}{P_1(y|x)\underbrace{P_1(x)}_{=1}}\Big) dy dx \\
&= \int_0^1 \int_{-\infty}^{\infty} P_0(y|x) \log\Big(\frac{P_0(y|x)}{P_1(y|x)}\Big) dy dx \\
&= \int_0^1 \mathrm{KL}(\mathcal{N}(0,\sigma^2), \mathcal{N}(f(x),\sigma^2)) dx \\
&= \int_0^1 \frac{1}{2\sigma^2}(0 - f(x))^2 dx \qquad \Big(\because \mathrm{KL}\left(\mathcal{N}(\mu_1,\sigma^2), \mathcal{N}(\mu_2,\sigma^2)\right) = \frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2\Big) \\
&= \frac{1}{2\sigma^2}\Big(\int_{1/2-\delta/L}^{1/2} L^2(x - 1/2 + \delta/L)^2 dx + \int_{1/2}^{1/2+\delta/L} L^2(1/2 + \delta/L - x)^2 dx\Big) \\
&= \frac{\delta^3}{3\sigma^2 L}.
\end{aligned}$$

Hence, we can choose

$$\delta := \frac{(3\sigma^2 \cdot L \cdot \log 2)^{1/3}}{n^{1/3}}$$

to ensure the inequality. Corrollary 2 now implies that

$$R_n^* \geq \frac{1}{8}\Phi\Big(\frac{\delta}{2}\Big) = \frac{1}{8} \cdot \frac{\delta^2}{4} = C \cdot \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} \tag{1}$$

where $C = \frac{1}{32}(3\log 2)^{1/3}$. This gives us the lower bound.

Remark 1.1. *Notice that for our construction, since we need $\delta/L \leq 1/2$ (see Figure 1), the lower bound in Equation (1) applies only if $n \geq 3\sigma^2(\log 2)^2/L^2$.*

Remark 1.2. *Observe that we obtained a weaker bound of $1/n^{2/3}$, compared to when we estimated the normal mean in which case we had a $\frac{1}{n}$ rate. This is because we only observe data samples around the point $X = 1/2$.*

1.2 Upper Bound

Let us now derive the upper bound. While our focus is on lower bounds, the techniques can be useful in general regression problems later on. The idea is to estimate $\theta = f(1/2)$ by taking the average of points that are close to $x = 1/2$. Towards this, we define:

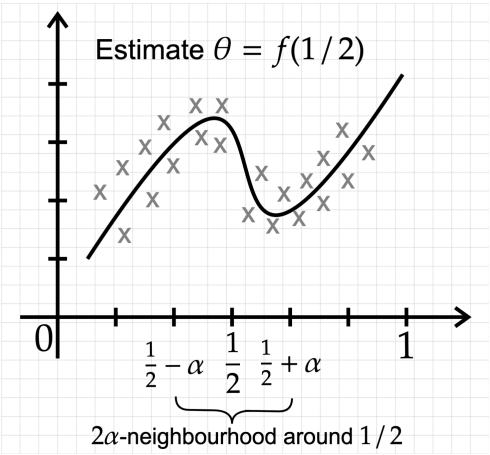

Figure 2: Idea: Estimate $\theta = f(1/2)$ by taking the average of points that are close to $x = 1/2$

$$N(S) := \sum_{i=1}^{n} \mathbb{1}\{x_i \in (1/2 - h, 1/2 + h)\},$$

and

$$\widehat{\theta}(S) := \begin{cases} 1/2, & \text{if } N = 0 \\ \frac{1}{N} \sum_{i=1}^{n} y_i \mathbb{1}\{x_i \in (1/2 - h, 1/2 + h)\}, & \text{if } N > 0 \end{cases}$$

We wish to compute the risk of this estimator. Define now a *good event*, $G := \{N(S) \geq nh\}$. Observe that $N(S) \sim \text{Binom}(n, 2h)$, and thus $\mathbb{E}[N(S)] = 2n\alpha$. Intuitively, G happens whenever at least half of the points fall near $\theta(1/2)$. It then follows by Hoeffding's inequality that

$$\begin{aligned} \mathbb{P}(G^c) &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}\{x_i \in (1/2 - h, 1/2 + h)\} \leq nha \right) \\ &\leq \exp\left(-2nh^2\right), \end{aligned}$$

which in turn leads by the tower property to the following:

$$\mathbb{E}[(\widehat{\theta}(S) - \theta)^2] = \mathbb{E}[(\widehat{\theta}(S) - \theta)^2 | G] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}[(\widehat{\theta}(S) - \theta)^2 | G^c]}_{\leq \theta^2 \leq 1} \underbrace{\mathbb{P}(G^c)}_{\leq \exp(-2nh^2)} . \tag{2}$$

Given this, we want to upper bound the term $\mathbb{E}[(\hat{\theta}(S)-\theta)^2|G]$ as follows. For notation, we let $A_i = \mathbb{1}\{x_i \in (1/2-h, 1/2+h)\}$).

$$\begin{aligned}(\widehat{\theta}(S)-\theta)^2 &= \Big(\frac{1}{N}\sum_{i=1}^{n} A_i y_i - \theta\Big)^2 \\ &= \Big(\underbrace{\frac{1}{N}\sum_{i=1}^{n} A_i(y_i - f(x_i))}_{v} + \underbrace{\frac{1}{N}\sum_{i=1}^{n} A_i(f(x_i)-\theta)}_{b}\Big)^2.\end{aligned}$$

Here, we can think of v, b as the variance and bias, respectively. Then, we have

$$\therefore \mathbb{E}[(\widehat{\theta}(S)-\theta)^2|G] = \mathbb{E}[(v+b)^2|G] = \mathbb{E}[b^2|G] + \mathbb{E}[v^2|G] + 2\mathbb{E}[bv|G]. \tag{3}$$

We now compute each of the three conditional expectations.

$$\begin{aligned}\mathbb{E}[v^2|G] &= \mathbb{E}\Big[\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^{N} A_i(y_i - f(x_i))\Big)^2 |G, X_1, \ldots, X_n\Big]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{1}{N^2}\sum_{i=1}^{N} A_i(y_i - f(x_i))^2 |G, X_1, \ldots, X_n\Big]\Big] \\ &\le \mathbb{E}\Big[\frac{1}{N^2}\cdot N\sigma^2 |G\Big] \\ &= \mathbb{E}\left[\frac{1}{N}\cdot\sigma^2|G\right] \\ &\le \frac{\sigma^2}{nh},\end{aligned}$$

where the first inequality follows from the bounded variance and the second inequality follows from the definition of good event G. Now, for any $x_i \in (1/2-h, 1/2+h)$, we have by assumption

$$|f(x_i) - f(1/2)| \le L|x_i - 1/2| \le Lh,$$

which gives

$$\begin{aligned}|b| &\le \left|\frac{1}{N}\sum_{i=1}^{N} A_i(f(x_i)-\theta)\right| \\ &\le \frac{1}{N}\sum_{i=1}^{N} A_i\,|f(x_i)-\theta| \\ &\le \frac{1}{N}Lh\sum_{i=1}^{N} A_i = Lh.\end{aligned}$$

Hence, we obtain

$$\mathbb{E}[b^2|G] \le L^2h^2.$$

Finally, we have

$$\begin{aligned}\mathbb{E}[bv|G] &= \mathbb{E}\Big[\Big[\Big(\frac{1}{n}\sum_{i=1}^{n} A_i(y_i - f(x_i))\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} A_i(f(x_i)-\theta)\Big)|G, X_1, \ldots, X_n\Big]\Big] \\ &= 0 \;\; (\because y_i - f(x_i)|X_i = x_i),\end{aligned}$$

since the cross terms vanish as they have zero mean and are independent. Combining these, we get

$$\mathbb{E}[(\hat{\theta}(S)-\theta)^2] \leq \frac{\sigma^2}{nh} + L^2h^2 + e^{-2nh^2}. \tag{4}$$

Choosing $h = \sigma^{2/3}/(L^{2/3}n^{1/3})$ gives

$$\Rightarrow \mathbb{E}[(\hat{\theta}(S)-\theta)^2] \leq \frac{2\sigma^{4/3}L^{2/3}}{n^{2/3}} + \exp\left(-\frac{2\sigma^{4/3}}{L^{4/3}}n^{1/3}\right). \tag{5}$$

Remark 1.3. *Observe that the first term above is precisely the lower bound we obtained earlier. But this requires the specific choice of h as above and the knowledge of both σ and L, which may or may not be known. If one need to decide h without knowing σ and L, they can for instance set $h = \frac{1}{N^{1/3}}$. This will lead to the following loose upper bound that is tight in n but not σ or L:*

$$\mathbb{E}[(\widehat{\theta}(S)-\theta)^2] \leq e^{-2n^{1/3}} + \frac{1}{n^{2/3}}(L^2+\sigma^2). \tag{6}$$

2 Insufficiency of Le Cam's method

The previous section shows us how Le Cam's method can be applied to point estimations via binary hypothesis testing. However, when we do high-dimensional parameter estimation, we would want to distinguish between multiple hypotheses for better estimation bounds. The following is an example of such problems: mean estimation for d-dimensional Gaussians.

Consider the family $\mathcal{P} = \left\{\mathcal{N}\left(\mu, \sigma^2 I\right) \mid \mu \in R^d\right\}$, with known σ^2. Suppose we have $S \sim P \in \mathcal{P}$, and

$$\Phi \circ \rho(\theta_1, \theta_2) := \|\theta_1 - \theta_2\|^2.$$

We wish to estimate $\theta(P) = \mathbb{E}_{x\sim P}[x]$. We start with the upper bound.

Consider the sample mean $\widehat{\theta}(S) = \frac{1}{n}\sum_{i=1}^n x_i$. Then, we have

$$\begin{aligned} R(\widehat{\theta}, P) &= \mathbb{E}_{S\sim P}\left[\left(\widehat{\theta}(S) - \theta(P)\right)^2\right] \\ &= \sum_{j=1}^d \mathbb{E}_S\left[\left(\frac{1}{n}\sum_{i=1}^n x_{ij} - \theta_j\right)^2\right] \\ &\leq \frac{\sigma^2 d}{n}, \end{aligned}$$

since we know for 1-d Gaussian we have σ^2/n. Notice that the above upper bound becomes increasingly loose as d increases. Now, we examine the lower bound via Le Cam's method.

Let $P_0 = \mathcal{N}(0, \sigma^2 I)$ and $P_1 = \mathcal{N}(\delta v, \sigma^2 I)$, where $v \in \mathbb{R}^d$ s.t. $\|v\|_2 = 1$. As before, we need to choose δ such that $KL(P_0, P_1) \leq \frac{\log 2}{n}$. Since P_0 and P_1 are Gaussian, we have $KL(P_0, P_1) = \delta^2/2\sigma^2$. Thus, choosing $\delta = \sqrt{2\log 2/n}$ suffices. By Corrollary 2, we have

$$R_n^* = \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_S\left[\Phi \circ \rho\left(\theta(P), \widehat{\theta}(S)\right)\right] \geq \underbrace{\frac{\log 2}{16} \cdot \frac{\sigma^2}{n}}_{\text{No } d \text{ factor here}}$$

3 Review of Information Theory

Now, we pivot and review some materials from information theory, which will be useful in the coming lectures.

3.1 Entropy

Definition 1 (Entropy of a random variable)**.** *Let X be an RV with distribution P. Then, the entropy of X is given by*

$$H(X) = \mathbb{E}_X[-\log p(X)].$$

For discrete X, we have

$$H(X) = -\sum_{x\in\chi} p(x)\log(p(x)),$$

while for continuous X,

$$H(X) = -\int_{\chi} p(x)\log(p(x))dx.$$

Intuitively, entropy measures the spread of the distribution, or in other words, it measures the amount of information or uncertainty about the possible outcomes contained in the variable.

For example, we have

- $X \sim \text{Bern}(p)$: $H(X) = -p\log p - (1-p)\log(1-p)$
- $X \sim \mathcal{N}(\mu, \sigma^2)$: $H(X) = \frac{1}{2}\log(2\pi e\sigma^2)$

Remark 3.1. *For any discrete RV X, we have*

$$0 \underset{(a)}{\leq} H(X) \underset{(b)}{\leq} \log|\mathcal{X}|$$

(a) : use $\log 1/p(x) \geq 0$ since $p(x) \leq 1$

(b) : by Jensen's inequality.

Definition 2 (Joint/Conditional Entropy)**.** *We define the joint and conditional entries as follows:*

$$H(X,Y) := \mathbb{E}_{x,y\sim P_{XY}}[-\log p(x,y)]$$

$$H(X|Y) := \mathbb{E}_{x,y\sim P_{XY}}[-\log p(x|y)].$$

The conditional entropy captures how much information is left in X after knowing $Y = y$. That is,

$$H(X|Y=y) = -\sum_{x\in\chi} p(x|y)\log(p(x|y)).$$

Taking expectation over the values y that Y can take, we see

$$\begin{aligned}
\mathbb{E}_Y[H(X|Y)] &= \sum_y p(y)H(X|Y=y) \\
&= \sum_{x,y} -p(x,y)\log p(x|y) \\
&= \mathbb{E}_{x,y} - \log p(x|y) \\
&= H(X|Y).
\end{aligned}$$