CS861: Theoretical Foundations of Machine Learning Lecture 10 - 27/09/2023

University of Wi[illegible] Fall 2023

Lec[illegible]'s Method (Some Examples)

Lecturer: Kirthe[illegible] *Scribed by: Ying Fu, Deep Patel*

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

Previously, we defined the notion of **Minimax Optimality**, following which we obtained lower bounds for this **Minimax Risk** by '**reducing**' the problem of **estimation** to that of **hypothesis testing**. Upon this 'reduction,' we then looked at **Le Cam's** method to obtain *a* **lower bound** by considering binary hypothesis testing specifically. In this lecture, we begin by obtaining 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.

Corollary 1 (From Le Cam's Method)**.** *Let $P_0, P_1 \in \mathcal{P}$ and $\delta = \rho(\theta(P_0), \theta(P_1))$. If $\mathrm{KL}(P_0, P_1) \le \frac{1}{n}\log 2$. Suppose the dataset S has n i.i.d samples. Then,*

$$R_n^* \ge \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) \tag{1}$$

Proof From our last lecture, we know that,

$$\begin{aligned}
R_n^* &\ge \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\|P_0^n \wedge P_1^n\| \\
&\ge \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)\frac{1}{2}\exp\left(-\mathrm{KL}\left(P_0^n, P_1^n\right)\right) \quad \left(\text{By the inequality}: \|P_0 \wedge P_1\| \ge \frac{1}{2}\exp\left(-\mathrm{KL}(P_0, P_1)\right)\right) \\
&= \frac{1}{4}\Phi\left(\frac{\delta}{2}\right)\exp\left(-n\mathrm{KL}\left(P_0, P_1\right)\right) \quad \left(\text{ Since } \mathrm{KL}(P_0^n, P_1^n) = n\mathrm{KL}(P_0, P_1)\right) \\
&\ge \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) \quad (\text{By assumption})
\end{aligned}$$

□

Remark 0.1. *Corollary 1 formalizes the intuition that if statistically speaking, we can find two distributions that are close to each other (i.e. low KL-divergence) but their parameters, $\theta(P_0)$ and $\theta(P_1)$ are far apart (i.e. high δ), then the minimax optimal error is also high. This is because the lower bound, as seen above in the Corollary, will be high owing to difficulty in distinguishing between the two distributions via the binary hypothesis test.*

1 Estimating Mean of Normal Distribution (with known variance)

Suppose we have $S = \{X_1, \cdots X_n\}$ with n i.i.d samples drawn from $P \in \mathcal{P}$, where $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$ with σ^2 is known. We also have $\Phi \circ \rho(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$. Define $P_0 = N\left(0, \sigma^2\right)$ and $P_1 = \mathcal{N}\left(\delta, \sigma^2\right)$ such that $|\theta(P_0) - \theta(P_1)| = \delta$. Then the KL divergence between P_0 and P_1 is,

$$\mathrm{KL}(P_0 \qquad \mathrm{L}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2\Big)$$

In order to satisfy the re 1 – which is $\mathrm{KL}(P_0, P_1) \leq \frac{\log 2}{n}$ – we will choose $\delta = \sigma\sqrt{\frac{2\log 2}{n}}$. Then, by Corollary 1, w

$$\begin{aligned} & \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) \\ &= \frac{1}{8}\left(\frac{\sigma}{2}\sqrt{\frac{2\log 2}{n}}\right)^2 \\ &= \frac{\log 2}{16}\frac{\sigma^2}{n} \end{aligned}$$

Previously, we have seen that the sample mean estimator achieves the σ^2/n rate.

2 Estimating Mean of Bernoulli Distribution

Before we proceed, we will state some facts about Bernoulli distributions that we will be using for deriving the lower bounds:

- If $P \sim \mathrm{Bern}(p)$ and $Q \sim \mathrm{Bern}(q)$, then we have

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

  The proof of relations in Equation (2) above is left for exercise. The first inequality(tagged as ①) is by Pinsker's inequality, and the second inequality(tagged as ②) will use the fact that $\log(x) \leq x - 1$.

- Therefore, if q is bounded away from 0 and 1, then,

$$\mathrm{KL}(P, Q) \in \Theta((p-q)^2)$$

  This is also a general statement about sub-Gaussian distributions (i.e., $\mathrm{KL}(P, Q) \in \Theta((\mu_P - \mu_Q)^2)$).

With this setup, let's begin our example of mean estimation for Bernoulli distributions. Let $S = \{X_1, \cdots X_n\}$ with n i.i.d samples drawn from $P \in \mathcal{P}$, where $\mathcal{P} = \{\mathrm{Bern}(\mu \mid \mu \in [0, 1]\}$. Let $P_1 = \mathrm{Bern}\left(\frac{1}{2}\right)$ and $P_2 = \mathrm{Bern}\left(\frac{1}{2} + \delta\right)$, then we have $|\theta(P_0) - \theta(P_1)| = \delta$. Then, by the facts introduced above, we have

$$\mathrm{KL}(P_0, P_1) \leq \frac{(p_0 - p_1)^2}{p_1(1 - p_1)} = \frac{\delta^2}{1/4}$$

In order to satisfy the requirement of Corollary 1 of $\mathrm{KL}(P_0, P_1) \leq \frac{\log 2}{n}$, we will choose $\delta = \frac{1}{2}\sqrt{\frac{\log 2}{n}}$. With this chosen δ, by Corollary 1, we get

$$\begin{aligned} R_n^* &\geq \frac{1}{8}\Phi\left(\frac{\delta}{2}\right) \\ &= \frac{1}{8}\left(\sqrt{\frac{\log 2}{n}} \cdot \frac{1}{4}\right)^2 \\ &= \frac{\log 2}{128}\frac{1}{n} \end{aligned}$$

Previously, we have seen that the sample mean estimator achieves $\frac{1}{n}$ rate.

3 A Simplified Regression Problem

Let $S = \{X_1, \ldots, X_n\}$ $\in \mathcal{P}$ where $X_i \overset{\text{i.i.d}}{\sim} \text{unif}([0,1])$ and $Y_i \overset{\text{i.i.d}}{\sim}$ σ-subGaussian distribution with mean sume that the function, $f : [0,1] \to [0,1]$, is $L-$Lipschitz. Thus, by our assumptic

$$\mathcal{P} = \{P_{XY} \mid \underbrace{P_X = \text{unif}(}_{\text{marginal c}} \cdot] \text{ is } L-\text{Lipschitz with range } [0,1],\ Y|X \text{ is } \sigma-\text{subgaussian}\}$$

We will begin with a si blem, where we are interested in estimating the value of the regression function at x $\mathbb{E}[Y|X = 1/2]$ and the criterion for measuring 'good'-ness is the standard squared error loss: $\Phi \circ \rho(\theta_1, \theta_2) \triangleq (\theta_1 - \theta_2)^2$.

3.1 Lower Bound

Let us first obtain the lower bound for the minimax risk. For this, we need to choose a P_0 and P_1. To keep things simple, we will choose them to be the following:

$$P_0 : P_0(y|x) = \mathcal{N}(f_0(x), \sigma^2),\ \underbrace{P_0(x) = \text{unif}([0,1])}_{\text{by assumption}}$$

$$P_1 : P_1(y|x) = \mathcal{N}(f_1(x), \sigma^2),\ \underbrace{P_1(x) = \text{unif}([0,1])}_{\text{by assumption}}$$

Recall that we we want the gap between two parameters to be large, i.e. large $\delta = |f_0(1/2) - f_1(1/2)|$, while ensuring that the two distributions are hard to distinguish, i.e. small $\text{KL}(P_0, P_1)$. Given that we require the regression functions, f_0 and f_1, to be L-Lipschitz, for the simplistic setting of $f_0 \equiv 0$, the regression function f_1 would look something like the one shown below in Figure 1.

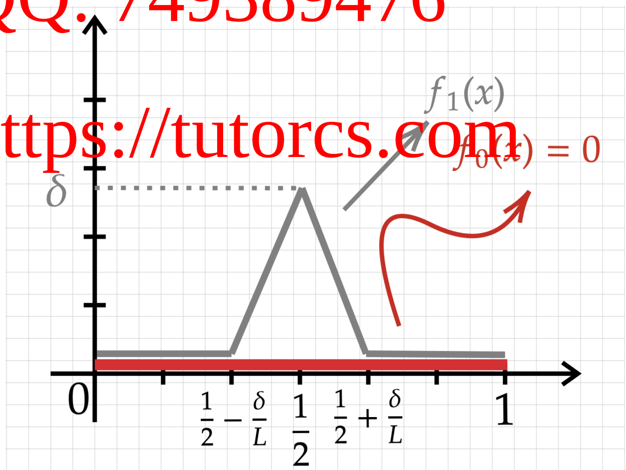

Figure 1: Here, we'd need $\delta/L \leq 1/2$ for f_1 to be well-defined.

We can mathematically define these functions, f_0 and f_1, as follows:

$$f_0(x) \triangleq 0 \ \ \forall\, x \in [0,1]$$

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

For obtaining the lower bound as derived in Corollary 1, recall that we need to choose δ such that $KL(P_0, P_1) \leq \frac{\log 2}{n}$. To do this, let us first compute $KL(P_0, P_1)$:

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

So, we can choose $\delta = \frac{(3\sigma^2 \cdot L \log 2)^{1/3}}{n^{1/3}}$ to ensure $KL(P_0, P_1) \leq \frac{\log 2}{n}$. Application of Corollary 1 now tells us that

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

where $C = \frac{1}{32}(3 \log 2)^{1/3}$.

Remark 3.1. *One minor detail here is that, for our construction, since we need $\delta/L \leq 1/2$ (see Figure 1), the lower bound in Equation (3) applies only if $n \geq \frac{3\sigma^2(\log 2)^2}{L^2}$.*

Remark 3.2. *Previously, when we were estimating the mean of a Gaussian distribution, we were getting a $\frac{1}{n}$ 'rate' in the lower bound as opposed to $\frac{1}{n^{2/3}}$ that we have obtained here for regression in Equation (3). This is because we observe data samples around the point $X = \frac{1}{2}$. We will note later that we will get similar, weaker rates even when we look at more realistic regression settings wherein we are not doing point estimation like we are here in our toy setting.*

3.2 Upper Bound

Having derived the lower bound, let us derive the upper bound now. While our focus is on lower bounds, this exercise will be useful when we derive upper bounds for general regression problems later on. For this, we will rely on the idea of estimating $\theta = f(1/2)$ by calculating the average of points in a neighborhood around the point $\frac{1}{2}$. Towards this, we'll now define a few quantities as follows:

$$\begin{aligned}
N(S) &\triangleq \sum_{i=1}^{n} \mathbb{I}\{X_i \in (1/2 - \alpha, 1/2 + \alpha)\} \\
\hat{\theta}(S) &\triangleq \begin{cases} 0, & \text{if } N(S) = 0 \\ \frac{1}{N(S)} \sum_{i=1}^{n} \mathbb{I}\{X_i \in (1/2 - \alpha, 1/2 + \alpha)\} y_i, & \text{if } N(S) > 0 \end{cases}
\end{aligned}$$

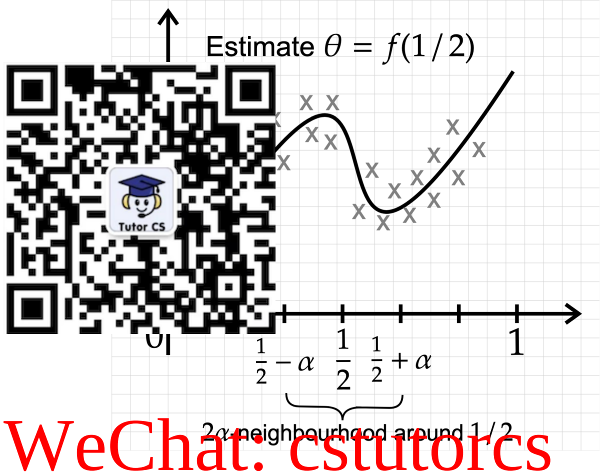

Figure 2: Idea: Estimate $\theta = f(1/2)$ by calculating the average of points in a neighborhood around the point 1/2

Having defined the estimator $\hat{\theta}(S)$, we now wish to compute the risk of this estimator. For this, we will define a 'Good Event': $G \triangleq \{N(S) \geq n\alpha\}$.

Remark 3.3 (Rationale for defining G the way we have)**.** *If we observe points sampled uniformly between 0 and 1, the probability of each of them falling in one of the "2α intervals" is 2α. Thus, we see that* $N(S) \sim \text{Binomial}(n, 2\alpha)$*. Since,* $\mathbb{E}[N(S)] = 2n\alpha$*, the 'Good Event' is saying that 'at least half of these points should fall near* $\theta = \frac{1}{2}$*'.*

Given the way we have defined G, it immediately leads us to the following inequality:

$$\mathbb{P}(G^c) = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{I}\{X_i \in (1/2 - \alpha, 1/2 + \alpha)\} \leq n\alpha\right]$$
$$\leq \exp(-2n\alpha^2) \quad (\because \text{By Hoeffding's inequality})$$

Thus, by the tower property of expectation operator, we can write the following:

$$\mathbb{E}[(\hat{\theta}(S) - \theta)^2] = \mathbb{E}[(\hat{\theta}(S) - \theta)^2|G] \underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}[(\hat{\theta}(S) - \theta)^2|G^c]}_{\leq \theta^2 \leq 1} \underbrace{\mathbb{P}(G^c)}_{\leq \exp(-2n\alpha^2)} \tag{4}$$

To upper bound $\mathbb{E}[(\hat{\theta}(S) - \theta)^2|G]$, we will expand $(\hat{\theta}(S) - \theta)^2$ as follows:

$$(\hat{\theta}(S) - \theta)^2 = \left(\frac{1}{N(S)} \sum_{i=1}^{n} A_i Y_i - \theta\right)^2 = \left(\frac{1}{N(S)} \sum_{i=1}^{n} A_i (Y_i - \theta)\right)^2$$
$$(\text{where } A_i = \mathbb{I}\{X_i \in (1/2 - \alpha, 1/2 + \alpha)\})$$
$$= \left(\underbrace{\frac{1}{N(S)} \sum_{i=1}^{n} A_i (Y_i - f(X_i))}_{v} + \underbrace{\frac{1}{N(S)} \sum_{i=1}^{n} A_i (f(X_i) - \theta)}_{b}\right)^2$$

Here, we can think of the quantities v, b as the variance and bias respectively. Therefore we can write the following

$$\therefore \mathbb{E}[(\hat{\theta}(S) - \theta)^2|G] = \mathbb{E}[(v + b)^2|G] = \mathbb{E}[b^2|G] + \mathbb{E}[v^2|G] + 2\mathbb{E}[bv|G] \tag{5}$$

Calculating and bounding each of the conditional expectations separately yields the following:

$\mathbb{E}[v$ [illegible] $_i(Y_i - f(X_i))\Big)^2|G, X_1, \dots, X_n\Big]\Big]$

[illegible] $A_i(Y_i - f(X_i))^2|G, X_1, \dots, X_n\Big]\Big]$

[illegible] $)\sigma^2|G\Big](\because Y_i - f(x_i) \text{ is } \sigma - \text{subGaussian})$

$$= \mathbb{E}[\frac{1}{N(S)} \cdot \sigma^2|G] \tag{6}$$

$$\leq \frac{\sigma^2}{n\alpha} (\because N(S) \geq n\alpha \text{ under the good event } G) \tag{7}$$

Similarly, we obtain the following

$$\mathbb{E}[b^2|G] = \mathbb{E}\Big[\Big(\frac{1}{N(S)} \sum_{i=1}^{N(S)} A_i \underbrace{f(x_i) - \theta}_{f(x_i) - f(1/2) \leq L\alpha}\Big)^2|G\Big]$$

$$\leq L^2\alpha^2 \tag{8}$$

and

$$\mathbb{E}[bv|G] = \mathbb{E}\Big[\Big[\Big(\frac{1}{n}\sum_{i=1}^{n} A_i(Y_i - f(X_i))\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} A_i(f(X_i) - \theta)\Big)|G, X_1, \dots, X_n\Big]\Big]$$

$$= 0 \ (\because Y_i - f(X_i)|X_i = x_i \text{ have zero mean and are independent}) \tag{9}$$

Combining Equations (4) – (9) together, we get

$$\mathbb{E}[(\hat{\theta}(S) - \theta)^2] \leq \frac{\sigma^2}{n\alpha} + L^2\alpha^2 + e^{-2n\alpha^2} \tag{10}$$

For a tighter upper bound, optimizing the R.H.S. above yields

$$\alpha^* = \frac{\sigma^{2/3}}{L^{2/3}n^{1/3}} \tag{11}$$

$$\Rightarrow \mathbb{E}[(\hat{\theta}(S) - \theta)^2] \leq \frac{2\sigma^{4/3}L^{2/3}}{n^{2/3}} + e^{-\frac{2\sigma^{4/3}}{L^{4/3}}n^{1/3}} \tag{12}$$

Remark 3.4. *Note that the first term in R.H.S. of Equation (12) above is precisely the lower bound we had got earlier. But this requires that we choose α^* as per Equation (11), which requires knowledge of both σ and L, which may or may not be known.*

However, if one wanted to choose α without the knowledge of σ and L, say, $\alpha = \frac{1}{N^{1/3}}$ instead of $\alpha = \alpha^*$, then we get the following loose upper bound for Equation (10) which is tight with respect to n but not σ or L:

$$\mathbb{E}[(\hat{\theta}(S) - \theta)^2] \leq e^{-2n^{1/3}} + \frac{1}{n^{2/3}}(L^2 + \sigma^2) \tag{13}$$