CS861: Theoretical Foundations of Machine Learning Lecture 11 - 29/09/2023
University of Wisconsin–Madison, Fall 2023

ew of Information Theory

Lecturer: Kirthe *Scribed by: Deep Patel, Keran Chen*

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

We have been looking at the notion of Minimiax Optimality for a few lectures wherein we introduced the notion of Minimax Risk and "reduced" the problem of estimation to that of hypothesis testing for obtaining lower bounds for the Minimax risk. Specifically, we "reduced" to binary hypothesis testing and derived the lower bound using Le Cam's method. We looked at some mean estimation and toy settings for regression problem for application of Le Cam's method. In this lecture, we will do a brief review of Information Theory and set the stage for Fano's method that will be more appropriate for the lower bounds we are interested in obtaining in this course.

1 Insufficiency of Le Cam's method

As we consider only binary hypothesis testing, Le Cam's method is usually sufficient only for point estimation. However, when we are doing high-dimensional parameter estimation, it would make sense to be able to distinguish between multiple hypotheses for better estimation bounds. To illustrate this, consider the following (imperfect) example of mean estimation for d-dimensional Gaussian distributions: Consider the family $\mathcal{P} = \left\{\mathcal{N}\left(\mu, \sigma^2 I\right) \mid \mu \in R^d\right\}$, where σ^2 is known. We are given a set $S = \{X_1, X_2, \ldots, X_n\} \overset{i.i.d.}{\sim} P \in \mathcal{P}$. Let $\Phi \circ \rho(\theta_1, \theta_2) \triangleq \|\theta_1 - \theta_2\|^2$. Consider $\hat{\theta}(S) = \frac{1}{n}\sum_{i=1}^n X_i$ as our estimator for the mean.

The upper bound for this mean estimator can be obtained as follows:

$$\begin{aligned} R(\hat{\theta}, P) &= \mathbb{E}_{S\sim P}\left[\left(\hat{\theta}(S) - \theta(P)\right)^2\right] \\ &= \sum_{j=1}^d \mathbb{E}_S\left[\left(\frac{1}{n}\sum_{i=1}^n X_{ij} - \theta_j\right)^2\right] \\ &\le \frac{\sigma^2 d}{n} \quad (\because \text{ for each 1-D Gaussian, upper bound is } \frac{\sigma^2}{n}) \end{aligned}$$

Note that the upper bound shown above becomes increasingly loose as the dimensionality, d, grows larger. Using Le Cam's method, we can obtain the lower bound as follows: 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. $\|v\|_2 = 1$). We want to apply Corollary 1 from Lecture 10 to obtain the lower bound. But for this, we need to choose δ such that $KL(P_0, P_1) \le \frac{\log 2}{n}$. Since P_0 and P_1 are Gaussian, we have $KL(P_0, P_1) = \frac{\delta^2}{2\sigma^2}$. Thus, we choose $\delta = \sqrt{\frac{\log 2}{n}}$. Whence, by Corollary 1 of Lecture 10, we have

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

2 Review of Information Theory

2.1 Entropy

Definition 1 (Entropy

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

Discrete random va $p(x)\log(p(x))$

Continuous random va $)\log(p(x))dx$

For example, if $X \sim \text{Bern}(p) \Rightarrow H(X) = -p\log p - (1-p)\log(1-p)$. Similarly, if $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow H(X) = \frac{1}{2}\log 2\pi e\sigma^2$

Remark 2.1 (Interpretation of Entropy for discrete random variables)**.** *Entropy measures the spread of the distribution. Another interpretation of entropy is that it measures the amount of information or uncertainty about the possible outcomes contained in a variable.*

Remark 2.2. *Some properties of Entropy (for Discrete R.V.):*

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

(a) : This uses the fact that $\log\frac{1}{p(x)} \geq 0$ *since* $p(x) \leq 1$

(b) : Refer to Lemma 6 towards the end of this lecture notes.

Definition 2 (Conditional Entropy)**.** *First define the entropy of X conditioned on knowing $Y = y$ as follows,*

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

The conditional entropy is the expectation of $H(X|Y=y)$ over Y. We have,

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

More generally, we can write

$$H(X|Y) = -E_{X,Y}[\log(p(X|Y))]$$

Definition 3 (Joint Entropy of two random variables)**.**

$$H(X,Y) = -E_{X,Y}[\log(p(X,Y))]$$

Lemma 1 (Chain Rule for Entropy)**.**

$$H(X_1, ...X_n) = \sum_{i=1}^{n} H(X_i|X_1...X_{i-1}) \tag{1}$$

$$H(X_1, ...X_n|Y) = \sum_{i=1}^{n} H(X_i|X_1...X_{i-1}, Y) \tag{2}$$

Proof *We will prove the first statement when $n = 2$. First note that we can write,*

$$p(x_1, x_2 \quad \log(p(x_1, x_2)) = -\log(p(x_1)) - \log(p(x_2|x_1))$$

Take expectation with r

$$) = H(X_1) + H(X_2|X_1)$$

Then use induction:

$$\begin{aligned} H(X_1, , X_n \quad & H(X_n|X_1, , X_{n-1}) \\ &= H(X_1, \dots, X_{n-2}) + H(X_{n-1}|X_1, \dots, X_{n-2}) + H(X_n|X_1, \dots, X_{n-1}) \\ &= \dots\dots \\ &= \sum_{i=1}^{n} H(X_i|X_1 \dots X_{i-1}) \end{aligned}$$

The second statement can be proved in a a similar fashion. □

Definition 4 (KL divergence (relative entropy) of distibution P and Q).

$$KL\left(P, Q\right) = E_{X\sim P}\left[\log\left(\frac{p\left(X\right)}{q\left(X\right)}\right)\right]$$

Lemma 2 ($KL(P, Q) \geq 0$ with equality iff $P = Q$).

$$KL\left(P, Q\right) = E_{X\sim P}\left[-\log\left(\frac{q\left(X\right)}{p\left(X\right)}\right)\right] \geq -\log\left(E_{X\sim P}\left[\frac{q\left(X\right)}{p\left(X\right)}\right]\right) = 0$$

The equality condition follows from the tightness condition for Jensen's inequality.

2.2 Mutual Information

Definition 5. *Mutual Information is the KL divergence between the joint distribution P_{XY} and product of marginals $P_X \times P_Y$*

$$I(X;Y) = KL(P_{XY}, P_X \times P_Y) = E_{P_{XY}}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right]$$

Some properties of Mutual Information:

- (Non-Negativity) $I(X;Y) \geq 0$ with equality IFF $X \perp\!\!\!\perp Y$
- (Symmetry) $I(X;Y) = I(Y;X)$

Lemma 3. *Mutual Information can be expressed in terms of entropy, conditional entropy, and joint entropy as follows:*

1.) $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

2.) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

3.) $I(X;Y) = H(X)$

Proof

1.)

$$
\begin{aligned}
&{}_{X,Y}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right]\\
&{}_{X,Y}\left[\log \frac{P(X)P(Y|X)}{P(X)P(Y)}\right]\\
&\mathbb{E}_Y[\log P(Y)] + \mathbb{E}_{X,Y}[\log P(Y|X)]\\
&Y) - H(Y|X)
\end{aligned}
$$

Similarly, one can $) - H(X|Y)$.

2.) By Chain Rule (Lemma 1), we have

$$
\begin{aligned}
H(X,Y) &= H(Y) + H(X|Y)\\
&= H(Y) + H(X) - I(X;Y)
\end{aligned}
$$

3.)

$$
\begin{aligned}
I(X;X) &= H(X) - H(X|X)\\
&= H(X) - 0
\end{aligned}
$$

□

Definition 6 (Conditional Mutual Information).

$$
\begin{aligned}
I(X;Y|Z) &\triangleq H(X|Z) - H(X|Y,Z)\\
&= \mathbb{E}_{X,Y,Z}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right]
\end{aligned}
$$

Lemma 4 (Chain Rule for Mutual Information).

$$
I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1})
$$

Proof We will see the proof for the case of $n = 2$ but exactly the same proof strategy works for any n.

$$
\begin{aligned}
I(X_1,X_2;Y) &\overset{\text{(by def.)}}{=} H(X_1,X_2) - H(X_1,X_2|Y)\\
&= H(X_1) + H(X_2|X_1) - (H(X_1|Y) + H(X_2|X_1,Y))\\
&\text{(Applying Lemma 1 for entropy and conditional entropy)}\\
&= I(X_1;Y) + I(X_2;Y|X_1)
\end{aligned}
$$

□

Lemma 5 (Conditioning reduces entropy). $H(X|Y) \leq H(X)$

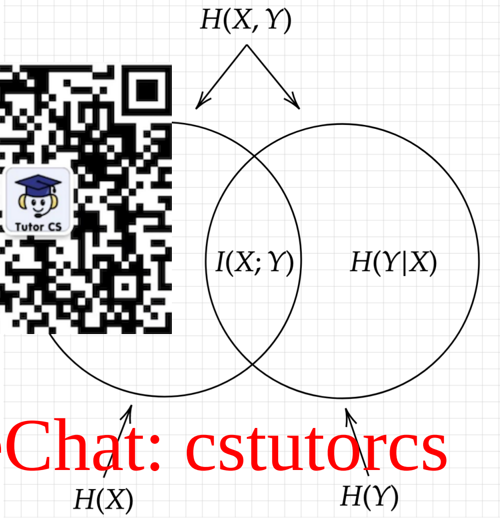

Figure 1: Summary of relationship between entropy and mutual information. Source: Chapter 2, Elements of Information Theory by Thomas M. Cover & Joy A. Thomas

Proof

$$I(X;Y) = H(X) - H(X|Y) \geq 0 \text{ (Mutual information is non-negative)}$$
$$\Rightarrow H(X|Y) \leq H(X)$$

□

Lemma 6 (Uniform distribution represents the maximum uncertainty)**.** *$0 \leq H(X) \leq \log|\mathcal{X}|$ for discrete random variable X. with equality IFF $X \overset{\text{unif}}{\sim} [|\mathcal{X}|]$.*

Proof Let P, U be the distribution of X and uniform random variable over $[|\mathcal{X}|]$. Let p, u be the corresponding PMFs. Then, we have

$$0 \leq KL(P,U) = \mathbb{E}_X\left[\log\frac{p(X)}{u(X)}\right] = -H(X) - \mathbb{E}_X\left[\log\frac{1}{|\mathcal{X}|}\right]$$
$$\Rightarrow H(X) \leq -\log\frac{1}{|\mathcal{X}|} = \log|\mathcal{X}|$$

$0 \leq H(X)$ is because $log(p(x)) \leq 0$ □

Lemma 7 (Data Processing Inequality)**.** *Say X, Y, Z are random variables such that $X \perp\!\!\!\perp Z|Y$. Then,*

$$I(X;Y) \geq I(X;Z)$$

Proof

$$I(X;Y,Z) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} \text{ (By applying Lemma 4)}$$
$$I(X;Y,Z) = I(X;Y) + \underbrace{I(X;Z|Y)}_{=0\ (\because X \perp\!\!\!\perp Z|Y)} \text{ (By applying Lemma 4)}$$
$$\therefore I(X;Y) \geq I(X;Z)$$

□

Remark 2.3. *Some r* *ata Processing Inequality (DPI) that we have proved above (Lemma 7):*

1.) *We can think of* *rkov Chain:* $X \longrightarrow Y \longrightarrow Z$.

2.) $I(X;Z) \leq I(X;Y$ *ns no more information about* X *than* Y*".*

3.) ***The road ahead***

- *In hypothesis testing, we will assume a prior on the alternatives* $\{P_1, \ldots, P_N\} \subseteq \mathcal{P}$
- $X \in [N]$ *forms a selection from* $\{P_1, \ldots, P_N\}$*. Then, the data,* Y*, is generated from the chosen* P_X*. Now we have a test,* Z*, which tries to estimate* X*.*
- *With the help of Fano's inequality, we will obtain a lower bound for the probability that* Z *fails to estimate* X*. We will use the data processing inequality to prove Fano's inequality.*

$$\mathbb{P}[Z \neq X] \geq \underbrace{\ldots\ldots}_{\text{lower bound}}$$