EE 6180: Advanced Topics in Artificial Intelligence	December 3, 2020
Lecture 15	
Lecturer: Abhishek Sinha	*Scribe: Sourav Sahoo*

1 Minimax Theory

Assume, there is a family of distributions $\mathcal{P}_\Theta$ parameterized by $\theta \in \Theta$. There exists n i.i.d. random variables, $X_1, \ldots, X_n$ from some distribution p_{θ^*}. We need to provide an estimate $T(X_1^n)$ for θ^*.

Definition 1 (Minimax Risk)**.** *Suppose $T(X_1^n)$ be an estimator of θ and $L(\cdot, \cdot)$ be a loss function. Then minimax risk is defined as:*

$$\mathfrak{M}_n = \inf_T \sup_{\theta^*} L(T(X_1^n), \theta^*) \tag{1}$$

This concept is used to find the information theoretic lower bounds, i.e, $\mathfrak{M}_n \geq B \implies$ irrespective of the chosen estimator, there exists an adversarial setting where one incurs a loss B. Now, we discuss a basic problem setting for the minimax theory.

1.1 Binary Hypothesis Testing

Definition 2 (Total Variation Distance)**.** *The total variation distance between two probability distributions P and Q is defined as:*

$$TV(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|\,.$$

Definition 3 (Pinsker's Inequality)**.** *Let the total variation distance and Kullback-Leibler distance between two probability distributions P and Q be denoted as $TV(P, Q)$ and $D(P||Q)$:*

$$TV(P, Q) \leq \sqrt{\frac{D(P||Q)}{2}}$$

where the inequality holds upto a constant logarithmic factor.

Let $\Theta = \{\theta_1, \theta_2\}$ and $\widehat{\theta}$ be the estimator. Here, the chosen loss function $L(\widehat{\theta}, \theta^*) = P_\theta(\widehat{\theta} \neq \theta)$. Now, we derive the fundamental lower bound of the minimax risk:

$$\begin{aligned}
\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta(\widehat{\theta} \neq \theta) &\geq \inf_{\widehat{\theta}} \left[\frac{1}{2} P_{\theta_1}(\widehat{\theta} \neq \theta_1) + \frac{1}{2} P_{\theta_2}(\widehat{\theta} \neq \theta_2)\right] && \text{maximum} \geq \text{average} \\
&= \inf_{\widehat{\theta}} \left[\frac{1}{2}(1 - P_{\theta_1}(\widehat{\theta} \neq \theta_2)) + \frac{1}{2} P_{\theta_2}(\widehat{\theta} \neq \theta_2)\right] \\
&= \frac{1}{2} - \frac{1}{2} \sup_{\widehat{\theta}} \left[P_{\theta_1}(\widehat{\theta} \neq \theta_2) - P_{\theta_2}(\widehat{\theta} \neq \theta_2)\right] \\
&\geq \frac{1}{2} - \frac{1}{2} TV(P_{\theta_1}, P_{\theta_2})
\end{aligned}$$

Example 4. *Let $P_1 = \mathcal{N}(-\mu, 1)$ and $P_2 = \mathcal{N}(\mu, 1)$. We have n i.i.d observations $X_1^n \sim P_\theta, \theta \in \{1, 2\}$. What is the relation between μ and n so that the error of estimation can be made arbitrarily small ?*

We have:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \{1,2\}} P_\theta(\widehat{\theta} \neq \theta) \geq \frac{1}{2} - \frac{1}{2} TV(P_{\theta_1}^n, P_{\theta_2}^n)$$

$$\geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{D(P^n_{\theta_1}||P^n_{\theta_2})}{2}}$$

The KL divergence for $\mathcal{N}(\theta_1, \Sigma)$ and $\mathcal{N}(\theta_2, \Sigma)$ is given as:

$$D(\mathcal{N}(\theta_1, \Sigma)||\mathcal{N}(\theta_2, \Sigma)) = \frac{1}{2}(\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2) \tag{2}$$

The product distributions are $P^n_{\theta_1} = \mathcal{N}(\mu\mathbf{1}, I_n)$ and $P^n_{\theta_2} = \mathcal{N}(-\mu\mathbf{1}, I_n) \implies D(P^n_{\theta_1}||P^n_{\theta_2}) = 2n\mu^2$. So,

$$\begin{aligned} P_e = \inf_{\widehat{\theta}} \sup_{\theta \in \{1,2\}} P_\theta(\widehat{\theta} \neq \theta) &\geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{D(P^n_{\theta_1}||P^n_{\theta_2})}{2}} \\ &= \frac{1}{2} - \frac{1}{2}\mu\sqrt{n} \end{aligned}$$

If $\mu\sqrt{n} \leq C \implies n \leq \mathcal{O}\left(\frac{1}{\mu^2}\right)$ for a constant $C > 0$, we have $P_e \geq \frac{1}{2} - \frac{1}{2}C$, i.e, we have a finite error. Furthermore, if $n \to \infty$, then $P_e \to 0$.

1.2 Multiple Hypothesis Testing

Now, we consider the case of multiple hypothesis testing. Before that, we discuss two famous results from information theory.

1.2.1 Data Processing Inequality and Fano's Inequality

Suppose we have a kernel $P_{Y|X}$ that takes an input X and gives out Y. Suppose it is given two input distributions P_X and Q_X and it generates P_Y and Q_Y respectively.

Theorem 5 (Data Processing Inequality)**.** $D(P_X||Q_X) \geq D(P_Y||Q_Y)$

Proof. Let P_{XY} and Q_{XY} be defined as:

$$Q_{XY} = Q_X P_{Y|X}, P_{XY} = P_X P_{Y|X} \tag{3}$$

Consider:

$$\begin{aligned} D(P_X||P_Y) &= \mathbb{E}\left[\log \frac{P_X}{Q_X}\right] \\ &\overset{(3)}{=} \mathbb{E}\left[\log \frac{P_{XY}}{Q_{XY}}\right] \\ &= \mathbb{E}\left[\log \frac{P_Y P_{X|Y}}{Q_Y Q_{X|Y}}\right] \\ &= D(P_Y||Q_Y) + D(P_{X|Y}||Q_{X|Y}) \\ &\geq D(P_Y||Q_Y) \end{aligned}$$

■

Consider the special case when the first distribution is joint distribution $P_\theta P_{X|\theta}$, the second distribution is the product distribution (joint distribution, assuming independence) $P_\theta P_X$, and the kernel is $\mathbf{1}(T(X) \neq \theta)$, i.e., probability that the estimator of X is θ. If the distributions are independent, the probability of error is simply $1 - \frac{1}{M}$ where M is the number of choices of θ (hypotheses). In case, they are not independent, let the probability of error be p_e. So, from data processing inequality, we have:

$$
\begin{aligned}
D(P_\theta P_{X|\theta} || P_X P_\theta) &\geq D\left((p_e, 1-p_e) || \left(\frac{1}{M}, 1 - \frac{1}{M}\right)\right) \\
\implies I(X, \theta) &\geq D\left((p_e, 1-p_e) || \left(\frac{1}{M}, 1 - \frac{1}{M}\right)\right) && \text{By definition of mutual information} \\
&= p_e \log \frac{p_e}{1 - 1/M} + (1-p_e) \log \frac{1-p_e}{1/M} \\
&= -h(p_e) + \log M - p_e \log(M-1) \\
&\geq -h(p_e) + \log M - p_e \log M \\
\implies p_e &\geq 1 - \frac{I(X, \theta) + h(p_e)}{\log M}
\end{aligned}
$$

where $h(p)$ is the binary entropy function. As $h(p_e) \leq 1$,

$$p_e \geq 1 - \frac{1 + I(X, \theta)}{\log M} \tag{4}$$

This is the celebrated Fano's inequality.