Course: E2 209 Date **January 8th, 2019**

Lecture 1

Instructor: Himanshu Tyagi *Scribe: Avi Mohan*

Part 1 of the course, as discussed during the lecture, will be concerned with distribution learning and testing. We will begin by describing some of the problems that will be addressed in this part of the course.

1 Motivating Examples

Example 1 (learning the bias of a coin)**.** Given a coin that may or may not be fair, how many tosses are required to

- test if the coin has bias 0.5 (fair) or $0.5 + \epsilon$?
- estimate the bias of the coin to within an accuracy of ϵ?

The two questions above are fundamentally different, with the former being a hypothesis testing problem and the latter a problem of estimation. However, it turns out that the number of tosses is of the order of $\frac{1}{\epsilon^2}$ in both cases.

Example 2 (uniformity testing)**.** Given IID[1] samples from some probability mass function (pmf) P on the set $[k] := \{1, 2, \ldots, k\}$, how many samples are required to test if P is *uniform* over $[k]$, i.e., $P(i) = \frac{1}{k}$, $\forall i \in [k]$, or ϵ-away from the uniform distribution (with respect to some measure of distance between probability distributions)? We will see that the number of samples here is of the order of $\frac{\sqrt{k}}{\epsilon^2}$. A point of note here is that the value of k itself is assumed to be known to us.

Example 3 (independence testing)**.** Given samples of a pair of random variables, how many samples does it take to determine if they are statistically independent? More formally, what is the smallest value of $n \in \mathbb{N}$, such that, given IID samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \sim P_{XY}$, one can determine of $P_{XY} = P_X \times P_Y$ or if P_{XY} is ϵ-away from all product measures?

Example 4 (Gaussian mixture learning)**.** We are given samples from a Gaussian mixture model, i.e., $\mathbf{X}_j \sim \sum_{i=1}^{m} w_i \mathcal{N}(\boldsymbol{\mu_i}, \mathbf{K_i}), 1 \leq j \leq n$, where $w_i \geq 0$ $\forall i \in [m]$, and $\sum_{i=1}^{m} w_i = 1$. What is the smallest value of n such that $((w_i, \boldsymbol{\mu_i}, \mathbf{K_i}), i \in [m])$ can be estimated to within some desired accuracy?

2 Formulating the problem: a decision-theoretic framework

- Goal: By observing a sample $X \sim P_\theta$, $\theta \in \Theta$, (or multiple IID samples) output an estimate $\hat{\theta}$. The family of distributions $(P_\theta)_{\theta \in \Theta}$ is assumed to be defined over some measure space $(\mathcal{X}, \mathcal{F})$ and the estimator is some measurable function $\hat{\theta} : \mathcal{X} \to \Theta$, with[2] $x \mapsto \hat{\theta}(x)$. Let the set of all such estimators be denoted by $\mathcal{E}(\mathcal{X}, \Theta)$. Note that while this only seems to cover the set of *deterministic* estimators, but it can be expanded using conditional expectation to include randomized estimators as well.
- Performance evaluation: How does one choose one estimator over another?
  - To quantify the performance of estimators, we associate with each a (measurable) *Risk* or *Loss* function $r : \Theta \times \Theta \to \mathbb{R}_+$, with $(\theta, \hat{\theta}) \mapsto r(\theta, \hat{\theta})$.

[1] IID stands for "independent and identically distributed."
[2] In more general formulations, it is, in fact, not necessary that the estimate lie within Θ itself.

  - Since the observation X is assumed to be drawn from a distribution, r is clearly a random variable and we define another quantity that we call *Average Risk* of a given estimator as $r_\theta(\hat{\theta}) := \mathbb{E}_\theta r(\theta, \hat{\theta})$, and $\mathbf{r}(\hat{\theta}) = \left(r_\theta(\hat{\theta})\right)_{\theta \in \Theta}$.
  - The *Risk region* of a given family of distributions is defined as $\mathcal{R}\left(\mathcal{P}_\Theta\right) = \bar{\mathsf{co}}\left(\left\{\mathbf{r}(\hat{\theta}), \hat{\theta} \in \mathcal{E}\left(\mathcal{X}, \Theta\right)\right\}\right)$, where, given a set $A \subset \mathbb{R}^d$, $\bar{\mathsf{co}}(A)$ is its convex closure.

- An estimator θ (also sometimes called a "policy") is said to be *Inadmissible* if there exists another estimator, say, $\hat{\alpha} \in \mathcal{E}$ that is uniformly better than $\hat{\theta}$, i.e., $r_\theta(\hat{\alpha}) \leq r_\theta(\hat{\theta}),\ \forall \theta \in \Theta$. An estimator that is not inadmissible is called (shockingly) *Admissible*.
  - Heuristically speaking, admissible policies are those that are the best for at least one θ.
  - However, inadmissible policies are not entirely useless. If finding an admissible policy becomes intractable, one might have to be content with an "approximately optimal" inadmissible policy (obtained through some iterative optimization, for example).

  Since different estimators might be optimal for different values of the parameter θ, it is helpful to have a single cost against which estimators can be compared. Towards this end, one can define two different metrics called the Bayes cost and the Minimax cost as follows.

- Bayesian cost: Define a prior measure π on Θ and let $\mathcal{P}(\Theta)$ be the set of all such measures. The *Bayes Risk* associated with this prior for a fixed policy $\hat{\theta}$ is defined as $R_\pi(\hat{\theta}) := \mathbb{E}_{\theta \sim \pi} r_\theta(\hat{\theta})$, and

$$R_\pi^* := \inf_{\hat{\theta} \in \Theta} R_\pi(\hat{\theta}) \tag{1}$$

  is the smallest risk for the given prior. A policy/estimator $\hat{\theta}$ that attains R_π^*, i.e., $R_\pi(\hat{\theta}) = R_\pi^*$ is said to be *Bayes Optimal* for that prior.

  Theorem 2.1 (Lucien le Cam)**.** Under some regularity conditions on $r, \mathcal{E}(\mathcal{X}, \Theta)$ and Θ, an estimator is admissible iff it is Bayes.

  Remark 2.1. The above theorem shows that Bayes optimal policies completely characterize the boundary of $\mathcal{R}\left(\mathcal{P}_\Theta\right)$. In practice, for a given problem, prior belief about the parameter and where within Θ it might lie can be encoded using π.

- The Minimax cost R^* is defined as

$$\begin{aligned} R^* &:= \inf_{\hat{\theta} \in \mathcal{E}} \sup_{\theta \in \Theta} r_\theta(\hat{\theta}) && (2) \\ &= \inf_{\hat{\theta} \in \mathcal{E}} \sup_{\pi \in \mathcal{P}_\Theta} R_\pi(\hat{\theta}) && (3) \\ &\geq \sup_{\pi \in \mathcal{P}_\Theta} \inf_{\hat{\theta} \in \mathcal{E}} R_\pi(\hat{\theta}) && (4) \end{aligned}$$

  Remark 2.2. Often (under the regularity conditions assumed in Thm. 2.1, for instance), the inequality in (4) becomes an equality, i.e., $R^* = \sup_{\pi \in \mathcal{P}_\Theta} \inf_{\hat{\theta} \in \mathcal{E}} R_\pi(\hat{\theta})$

  Remark 2.3. Suppose $R^* = \min_{\hat{\theta}} \max_\pi R_\pi(\hat{\theta})$. Then, denoting by π^* the least favorable prior, a good strategy for attaining R^* is to use a policy for this least favorable prior.

3 Distances between distributions

The difficulty of estimating θ is related to how "close" the distributions in the family $(\mathcal{P}_\theta)_\Theta$ are to each other. This closeness between distributions will be characterized using various measures of distance, such as the following.

Total variation distance: Given two distributions P and Q over $(\mathcal{X}, \mathcal{F})$, the total variation (TV) distance between them is defined as

$$\begin{aligned} d(P,Q) &:= \sup_{A\in\mathcal{F}} P(A) - Q(A) \\ &\overset{*a}{=} \sup_{A\in\mathcal{F}} 1 - P(A^C) - (1 - Q(A^C)) = \sup_{A\in\mathcal{F}} Q(A^C) - P(A^C) \\ &= \sup_{A\in\mathcal{F}} Q(A) - P(A), \end{aligned} \tag{5}$$

where in $*a$ $A^C = \mathcal{X} \setminus A$ the complement of set A.

Remark 3.1. If P and Q have densities f and g with respect to some measure μ, we have

$$d(P,Q) = \frac{1}{2}\int_{\mathcal{X}} \Big| f(x) - g(x) \Big| d\mu(x) \tag{6}$$

Remark 3.2. In fact, $d(\cdot,\cdot)$ is provably a true distance which means that $d(P,Q) = 0$ iff they agree on every measurable set. Furthermore, from (6), it is clear that TV distance always lies in $[0,1]$, since

$$\begin{aligned} d(P,Q) &= \frac{1}{2}\int_{\mathcal{X}} \Big| f(x) - g(x) \Big| d\mu(x) \qquad (7) \\ &\leq \frac{1}{2}\int_{\mathcal{X}} \mid f(x)\mid + \mid g(x) \mid d\mu(x) = \frac{1}{2}\left(\int_{\mathcal{X}} f(x)d\mu(x) + \int_{\mathcal{X}} g(x)d\mu(x)\right) \\ &= 1. \qquad (8) \end{aligned}$$

$d(P,Q) = 1$ if P and Q have disjoint supports.

3.1 Probability of error

As a quick application of TV distance, consider the problem of distinguishing between two measures P_1 and P_2 over $(\mathcal{X}, \mathcal{F})$ using a single sample X, and consider all tests of the form

$$\hat{\theta}_A(X) = 1\mathbb{I}_{\{X\in A\}} + 2\mathbb{I}_{\{X\in A^c\}}, \tag{9}$$

for some $A \in \mathcal{F}$. Assume a uniform prior over $\{P_1, P_2\}$. The probability of error (Bayes risk) in this case is given by

$$\begin{aligned} P_e^* &= \inf_{A\in\mathcal{F}} \frac{1}{2}\left(P_1\left(A^c\right) + P_2\left(A\right)\right) \qquad (10) \\ &= \inf_{A\in\mathcal{F}} \frac{1}{2}\left(1 - P_1\left(A\right) + P_2\left(A\right)\right) \\ &= \frac{1}{2}\left(1 - \sup_{A\in\mathcal{F}} P_1\left(A\right) - P_2\left(A\right)\right) \qquad (11) \\ &= \frac{1}{2}\left(1 - d(P_1,P_2)\right). \qquad (12) \end{aligned}$$

As can be seen from the above equation, smaller the distance between P_1 and P_2, larger the probability of error. We had mentioned that the difficulty of estimating θ is related to how close the measures are within Θ and this example illustrated this in a rather concrete manner.

Preview of Lecture 2: In the next lecture, we will continue our study of the TV distance and also study other notions of distance between probability measures. Returning to Sec. 3.1, we saw how TV distance influenced error probability in (12). But that was with a single sample. Suppose $n \geq 2$ samples are available to us. Can we make use of them in some way as to reduce P_e^*? Specifically, denoting by P_1^n and P_2^n the n-fold product measures of P_1 and P_2 (since samples are IID), a natural question to ask is how $d(P_1,P_2)$ and $d(P_1^n,P_2^n)$ are related. If TV distance increases with n, that's good news because it means that there might be exist means to reduce P_e^* in (12).