ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya — Lecture #2
Scribe: JA — 30th August, 2016

Please send errors to acharya@cornell.edu

We did a brief recap of the previous lecture. We then outlined the three things we will discuss today:

- Basics of minimax theory
- Learning discrete distributions
- Lower bound for learning Bernoulli distributions

1 Basic Minimax Theory

Minimax theory provides an elegant framework to study the performance of various estimators. We study it in the context of the learning distributions, but it can be defined for any statistical estimation problem.

Let $\mathcal{P}$ be a known collection of distributions. We want to learn an unknown distribution from $\mathcal{P}$ given samples X_1^n from it. Consider the following steps:

1. We pick an estimator that takes a set of n samples and outputs a distribution $Q_{X_1^n}$.
2. An adversary (with the knowledge of our estimator) picks a distribution $P \in \mathcal{P}$.
3. We observe n independent samples $X_1, \ldots, X_n$ from P.
4. We output $Q_{X_1^n}$ as our estimator.
5. We incur a loss $L(P, Q_{X_1^n})$ for some loss function L.

Given a loss function, we would like the loss to be small. In the last lecture we learnt Bernoulli distributions under total variation distance.

There are two related performance measures, the expected loss, and the sample complexity.

Goal 1: Design an estimator that has the least *expected loss* when we carry out the steps above. Let us call it L^*. It turns out that this expected value will be equal to:

$$L^* = \min_{Q} \max_{P \in \mathcal{P}} \mathbb{E}_{X_1^n \sim P} \left[L(P, Q_{X_1^n}) \right]. \tag{1}$$

Do not read the math for a second. All we want to do is to find the least number of samples, such that for any possible underlying underlying distribution P, the probability that the loss we incur exceeds ε is at most δ.

Please stare at this definition, and notice how it relates to the steps described. Also, notice the order of min, and max (we choose estimator, then adversary chooses an instance). This expected

loss is widely used in statistics, information theory etc. In the cs theory literature somehow these problems have been studied in the context of sample complexity.

Goal 2: Given an error parameter $\varepsilon > 0$, and an error probability δ (typically a constant, say 1/4). We want to find the following quantity:

$$n^*(\mathcal{P}, \varepsilon, \delta, L) \stackrel{\text{def}}{=} \min_n : \exists\, Q_{X_1^n} : \forall P \ \in \mathcal{P} \ \Pr\left(L(P, Q_{X_1^n}) > \varepsilon\right) < \delta \tag{2}$$

For most of the lectures, we will bother with δ being a fixed constant, say 1/4, and the loss is the total variation distance. Let

$$n^*(\mathcal{P}, \varepsilon) = n^*(\mathcal{P}, \varepsilon, 1/4, d_{TV}). \tag{3}$$

This is the least number of samples sufficient to estimate any distribution in $\mathcal{P}$ to accuracy ε with probability at least 3/4.

We will typically be interested in understanding $n^*(\mathcal{P}, \varepsilon)$ in terms of some parameterization of $\mathcal{P}$. Some example classes are:

1. Δ_k: This is the collection of all discrete distributions over a set $\mathcal{X}$ with $|\mathcal{X}| = k$. We would like to know $n^*(\mathcal{P}, \varepsilon)$ as a function of k, and ε.

   Note that Δ_2 is the collection of all Bernoulli distributions.

2. $\mathcal{M}_k$: These are the collection of all monotone (decreasing) distributions over say $\{1, 2, \ldots, k\}$. Such distributions satisfy $P(x) \geq P(y)$ for $x \leq y$.

3. One dimensional Gaussian distributions

4. High dimensional Gaussians with some structure

5. Mixtures of simple distributions

In the last lecture, we proved a bound on the number of samples required to learn a Bernoulli distribution. Using $p(1-p) \leq 1/4$ for any $p \in \mathbb{R}$, we have

$$n^*(\Delta_2, \varepsilon) \leq \frac{1}{\varepsilon^2}. \tag{4}$$

2 Learning Δ_k

Suppose we observe $X_1^n \stackrel{\text{def}}{=} X_1, \ldots, X_n$ from a distribution P over $\mathcal{X}$. Let

$$N_x \stackrel{\text{def}}{=} \{\#\text{ times symbol } x \text{ appears in } X_1^n\}.$$

For example, if $X_1^n = H\ T\ T\ H\ T$, $N_H = 2$, $N_T = 3$. Under independent sampling from P, where the probability of x is $P(x)$, N_x is a Binomial $Bin(n, P(x))$ distribution (Show this).

Define the empirical estimator P_{ML} as:

$$P_{ML}(x) = \frac{N_x}{n},$$

the distribution that assigns the empirical probability to each symbol. It is also the distribution that assigns the highest probability to the sequence X_1^n under independent sampling (requires a short proof).

Question: Show that of all distributions in Δ_k, the distribution P_{ML} assigns the highest probability to X_1^n, namely

$$P_{ML} = \arg\max_{P \in \mathcal{P}} \prod_{i=1}^{n} P(X_i) = \prod_{x \in \mathcal{X}} P(x)^{N_x}.$$

We will understand the complexity of the ML estimator under the total variation distance. Let

$$L_{ML} \stackrel{\text{def}}{=} |P - P_{ML}|_1 = \sum_{x \in \mathcal{X}} |P(x) - \frac{N_x}{n}|.$$

be the random variable denoting the L_1 distance between the ML distribution and the underlying distribution P. Bounding the sum of absolute values can be painful. A very useful routine is to upper bound these as squares, which can be easier to handle. The Cauchy-Schwarz inequality can be very helpful in this:

Lemma 1 (Cauchy Schwarz Inequality (CSI))**.** *Let $a_1, \ldots, a_n$, and $b_1, \ldots, b_n$ be real numbers. Then,*

$$\left(\sum a_i b_i\right)^2 \leq \left(\sum a_i^2\right) \cdot \left(\sum b_i^2\right)$$

Using CSI with $b_i = 1$, for any X_1^n,

$$L_{ML}^2 = \left(\sum_{x \in \mathcal{X}} |P(x) - \frac{N_x}{n}|\right)^2 \leq |\mathcal{X}| \cdot \left(\sum_{x \in \mathcal{X}} \left(P(x) - \frac{N_x}{n}\right)^2\right).$$

Taking expectations of both sides,

$$\mathbb{E}[L_{ML}^2] \leq |\mathcal{X}| \cdot \frac{1}{n^2} \mathbb{E}[\sum_{x \in \mathcal{X}} (N_x - nP(x))^2] = \frac{|\mathcal{X}|}{n^2} \cdot \left(\sum_{x \in \mathcal{X}} nP(x)(1 - P(x))\right) \leq \frac{|\mathcal{X}|}{n^2} \sum_{x \in \mathcal{X}} nP(x) = \frac{|\mathcal{X}|}{n}.$$

Recall that $L_{ML} = 2 \cdot d_{TV}(P, P_{ML})$. Applying Markov's Inequality for the random variable L_{ML}^2,

$$\begin{aligned} \Pr\left(d_{TV}(P, P_{ML}) \geq \varepsilon\right) &= \Pr\left(L_{ML} \geq 2\varepsilon\right) \\ &= \Pr\left(L_{ML}^2 \geq 4\varepsilon\right) \\ &\leq \frac{\mathbb{E}[L_{ML}^2}{4\varepsilon^2} \\ &\leq \frac{|\mathcal{X}|}{4n\varepsilon^2}. \end{aligned} \tag{5}$$

Therefore, when $n \geq |\mathcal{X}|/\varepsilon^2 = k/\varepsilon^2$, with probability at least $3/4$, P_{ML} is at a total variation at most ε from P. Therefore,

Theorem 2.

$$n^*(\Delta_k, \varepsilon) \leq \frac{k}{\varepsilon^2}.$$

3 Lower bound for Bernoulli Estimation

In the last lecture we saw that

$$n^*(\Delta_2, \varepsilon) \leq \frac{1}{\varepsilon^2},$$

which is achieved by the ML estimator.

We would like to prove a lower bound on the performance of *any estimator*. We will do it via a reduction to the following hypothesis testing problem. We will set it similar to the first class (Bayes error).

The distribution D is uniformly chosen to be either $P = Bern(\frac{1}{2})$, or $Q = Bern(\frac{1}{2}+2\varepsilon)$. We are given n independent samples from D, and we have to decide between P and Q. Giving n samples from D is equivalent to giving one sample from the distribution P^n defined as:

$$P^n(X_1^n) \stackrel{\text{def}}{=} P(X_1) \cdots P(X_n).$$

Therefore, the error of the best classifier (from previous lecture) is

$$\frac{1}{2} - \frac{1}{4} \cdot |P^n - Q^n|_1.$$

Let n_T be the minimum number of samples such that the testing error is at most $1/4$. We relate the testing and learning complexity with the following claim.

Claim 3. $n_T \leq n^*(\Delta_2, \varepsilon)$.

Proof. When we run the learning algorithm on D, the output is ε close to D with probability at least $3/4$. Since P and Q are TV 2ε, we simply output the P or Q which is closer to P_{ML}. The testing error is at most the error probability of learning, which is at most $1/4$. □

Let $n \geq n_T$. Then,

$$\frac{1}{4} \geq \frac{1}{2} - \frac{1}{4} \cdot |P^n - Q^n|_1$$

implying that

$$|P^n - Q^n|_1 \geq 1.$$

Now, by Pinsker's Inequality

$$1 \leq |P^n - Q^n|_1^2 \leq 2 \cdot D(Q^n||P^n). \tag{6}$$

Recall from last lecture that $D(Q^n||P^n) = n \cdot D(Q||P)$. Moreover, using $\log(1+x) \leq x$,

$$D(Q||P) = (\frac{1}{2} + 2\varepsilon) \log \frac{(\frac{1}{2} + 2\varepsilon)}{\frac{1}{2}} + (\frac{1}{2} - 2\varepsilon) \log \frac{(\frac{1}{2} - 2\varepsilon)}{\frac{1}{2}} \tag{7}$$

$$\leq (\frac{1}{2} + 2\varepsilon) \cdot (4\varepsilon) + (\frac{1}{2} - 2\varepsilon) \cdot (-4\varepsilon) \tag{8}$$

$$= 16 \cdot \varepsilon^2. \tag{9}$$

Therefore,

$$1 \leq 32n\varepsilon^2.$$

Combining these results we obtain

$$\frac{1}{32\varepsilon^2} \leq n^*(\Delta_2, \varepsilon) \leq \frac{1}{\varepsilon^2}. \tag{10}$$

In the order notation, this means,

$$n^*(\Delta_2, \varepsilon) = \Theta\left(\frac{1}{\varepsilon^2}\right).$$

Question: Improve the constants for $n^*(\Delta_2, \varepsilon)$ by improving any side of (10).

In the next lecture, we will cover some basics of Information theory, and techniques for proving lower bounds. In particular, we aim to prove a lower bound of the form $c \cdot k/\varepsilon^2$ on $n^*(\Delta_k, \varepsilon)$.