ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya — Lecture #3
Scribe: Sourbh Bhadane — 30th August, 2017

1 Recap

In the previous lecture, we saw

1. **Empirical estimator for TV-Estimation problem**

$$\mathbb{E}\left[l_1\left(p, \widehat{p_n}\right)\right] \leq \sqrt{\frac{k}{n}}$$

2. **Poisson Sampling**: If a problem can be solved in $O\left(n\right)$ samples, it can be solved in $O\left(\mathrm{Poi}\left(n\right)\right)$ samples.

2 Hypothesis Testing

Hypothesis testing is a statistical inference technique wherein, based on observations from a phenomenon, a choice is made between two or more hypotheses which are concerned with the origins of the observed phenomenon. It finds applications in diverse areas; for example, testing whether a lottery is fair or not. A real-life example of such a lottery is the Polish Multilotek where the initial machine was found to be biased against the number 50-59.

2.1 Problem Formulation

Let $\mathbb{P}$ and $\mathbb{Q}$ be two collections of probability distributions. In most cases, we will assume that $\mathbb{P} \cap \mathbb{Q} = \emptyset$. Let $p \in \mathbb{P} \cup \mathbb{Q}$ be an unknown probability distribution and let $X_1 \cdots X_n$ be sampled from p. Note that these may or may not be independently sampled. The goal of hypothesis testing is as follows :

Goal : Decide if $p \in \mathbb{P}$ or $p \in \mathbb{Q}$ with probability atleast $1 - \delta = 0.9$.

We are primarily interested in the setting where $\mathbb{P}$ and $\mathbb{Q}$ are defined over a large domain compared to the number of samples.

2.2 Test statistic

In order to capture the differences between $\mathbb{P}$ and $\mathbb{Q}$, a useful approach is to design a test statistic, $T : X_1, \cdots, X_n \to \mathbb{R}$. For a fixed threshold t, compare $T\left(X_1, \cdots, X_n\right)$ with t and make a decision. We will henceforth abuse notation and refer to $T\left(X_1, \cdots, X_n\right)$ as T.

Example 1. Fix $t = 0$ and let T be a test statistic designed such that we decide

$$p \in \mathbb{P} \text{ if } T < 0$$
$$p \in \mathbb{Q} \text{ if } T \geq 0$$

The goal of hypothesis testing is achieved if

$$\Pr\left(T \geq 0 \mid p \in \mathbb{P}\right) \leq 0.1, \Pr\left(T < 0 \mid p \in \mathbb{Q}\right) \leq 0.1$$

The following example illustrates hypothesis testing between two Bernoulli distributions.

Example 2.

$$\mathbb{P} = \left\{\text{Bern } \left(\frac{1}{2} - \epsilon\right)\right\}, \mathbb{Q} = \left\{\text{Bern } \left(\frac{1}{2} + \epsilon\right)\right\}$$

Let $p \in \mathbb{P} \cup \mathbb{Q}$. Given $X_1, \cdots, X_n \sim p$, we choose $T = N_1$ where N_x is defined to be the number of times the symbol x appears in $X_1, \cdots, X^n$. Therefore,

$$\mathbb{E}\left[T \mid p \in \mathbb{P}\right] = n\left(\frac{1}{2} - \epsilon\right)$$
$$\mathbb{E}\left[T \mid p \in \mathbb{Q}\right] = n\left(\frac{1}{2} + \epsilon\right)$$

We now find an upper bound on $\Pr\left(T \geq \frac{n}{2} \mid p \in \mathbb{P}\right)$ by applying the Chebyshev's inequality (see Appendix).

$$\text{Var}\left(T \mid p \in \mathbb{P}\right) = n\left(\frac{1}{2} - \epsilon\right)\left(1 - \left(\frac{1}{2} - \epsilon\right)\right) = n\left(\frac{1}{4} - \epsilon^2\right) \leq \frac{n}{4}$$

For $a = \sqrt{10}$, by Chebyshev's inequality, we have

$$\Pr\left(\mid T - E\left(T \mid p \in \mathbb{P}\right) \mid \geq \frac{\sqrt{10n}}{2}\right) \leq 0.1$$

When $n\epsilon > \frac{\sqrt{10n}}{2}$,

$$\Pr\left(T > \frac{n}{2} \mid p \in \mathbb{P}\right) < \Pr\left(\mid T - E\left(T \mid p \in \mathbb{P}\right) \mid \geq \frac{\sqrt{10n}}{2}\right) \leq 0.1$$

We can therefore conclude that $n > \dfrac{5}{2\epsilon^2} = O\left(\dfrac{1}{\epsilon^2}\right)$ samples are sufficient for hypothesis testing between two Bernoulli distributions.

3 Uniformity Testing

The uniformity testing problem is equivalent to a hypothesis testing problem when $\mathbb{P} = \{p : l_1(p, u) > \epsilon\}, \mathbb{Q} = u$, where u is the uniform distribution and p is any discrete distribution over $[k]$. Our objective is to find the minimum number of samples required to meet the hypothesis testing goal.

A naive approach to achieve the hypothesis testing goal is to take a lot of samples and learn the distribution. More formally, the plug-in estimator $\widehat{p}$ can be learned such that we decide

$$p \in \mathbb{P} \text{ if } l_1\left(\widehat{p}, u\right) \geq \frac{\epsilon}{2}$$
$$p \in \mathbb{Q} \text{ if } l_1\left(\widehat{p}, u\right) < \frac{\epsilon}{2}$$

It can be shown that the above approach requires atleast $\Omega\left(\frac{k}{\epsilon^2}\right)$ samples. We now prove that for the setting in which we are interested i.e. large domain size compared to the number of samples, it is possible to do better in terms of number of samples required.
Remark : As the number of samples $n \to \infty$, the naive approach can indeed be shown to be the best we can do. This is a general theme that is applicable to multiple problems and will be revisited later in the course.

Theorem 3. *$O\left(\frac{\sqrt{k}}{\epsilon^2}\right)$ samples are necessary and sufficient to solve the uniformity testing problem*

This lecture contained the proof of the sufficiency part alone. We state the following lemma that will be used in the proof.

Lemma 4. *Let $X \sim Poi(\lambda)$, $\mu \in \mathbb{R}$. Define Z to be $(X-\mu)^2 - X$ If*

$$\mathbb{E}[Z] = (\lambda - \mu)^2$$

then

$$Var(Z) = 2\lambda^2 + 4\lambda(\lambda-\mu)^2$$

Before presenting the proof of Theorem (3), we first present some intuition behind the approach. Observe that if we can design a test statistic T such that

$$\mathbb{E}[T] = l_1(p,u)$$

the following test suffices

$$p \in \mathbb{P} \text{ if } T \geq \frac{\epsilon}{2}$$
$$p \in \mathbb{Q} \text{ if } T < \frac{\epsilon}{2}$$

It turns out that designing the above test statistic is harder. Therefore, we relate the test statistic to the l_2 distance instead of the l_1 distance.

Proof. Take Poi(n) independent samples from p. For $x \in [k]$, let N_x denote the number of times symbol x occurs in the data. Define T as follows,

$$T = \sum_{x \in [k]} \left(N_x - \frac{n}{k}\right)^2 - N_x$$

We can prove that $\mathbb{E}[T] = n^2 (l_2(p,u))^2$. See Appendix for the proof of a generalized version of T. By Cauchy-Schwarz inequality, we have

$$k\left(l_2(p,u)^2\right) \geq l_1(p,u)^2$$

Therefore,

$$\mathbb{E}[T \mid l_1(p,u) > \epsilon] \geq \frac{n^2\epsilon^2}{k}$$

As a result, we choose t to be $\frac{n^2\epsilon^2}{2k}$. The upper bound on the minimum number of samples is obtained by comparing $\mathrm{Var}\left(T \mid p = u\right)$ and $\mathrm{Var}\left(T \mid l_1\left(p, u\right) > \epsilon\right)$ with $t = \frac{n^2\epsilon^2}{2k}$.
By applying Lemma (4) to N_x, we have

$$\mathrm{Var}\left(\left(N_x - \frac{n}{k}\right)^2 - N_x\right) = 2n^2 p\left(x\right)^2 + 4np\left(x\right)\left(np\left(x\right) - \frac{n}{k}\right)^2$$

Since N_x's are independent under Poisson sampling, we have

$$\mathrm{Var}\left(T \mid p = u\right) = 2\sum_{x\in[k]} n^2\left(u\left(x\right)^2\right) = \frac{2n^2}{k}$$

We now use the same reasoning that we used in Example 2 to obtain the upper bound on n using Chebyshev's inequality. Therefore, when

$$\frac{n^2\epsilon^2}{2k} > \sqrt{10}\left(\sqrt{\frac{2n^2}{k}}\right)$$
$$n > 4\sqrt{5}\left(\frac{\sqrt{k}}{\epsilon^2}\right)$$
$$\implies \mathrm{Pr}\left(T > \frac{n^2\epsilon^2}{2k} \mid p = u\right) < 0.1$$

Note that we still have to find a condition on n such that $\mathrm{Pr}\left(T < \frac{n^2\epsilon^2}{2k} \mid l_1\left(p, u\right) > \epsilon\right) < 0.1$ by computing $\mathrm{Var}\left(T \mid l_1\left(p, u\right) > \epsilon\right)$ and applying Chebyshev's inequality. This will be done in the next class.

□

4 Appendix

We now state the Chebyshev's inequality and some lemmas with proofs that were not included in the main scribe.

4.1 Chebyshev's Inequality

Proposition 5. *Let X be a random variable with finite mean μ and finite non-zero variance σ^2. Then for any real number $a > 0$,*

$$\mathrm{Pr}\left(\mid X - \mu \mid \geq a\sigma\right) \leq \frac{1}{a^2}$$

4.2 Property of Poisson Sampling

Lemma. *Generate* $Poi(n)$ *samples independently from* p*. Let* N_x *denote the number of times symbol* $x \in [k]$ *appears in the data. For any discrete distribution* q *over* $[k]$*,*

$$\mathbb{E}\left[\sum_{x\in[k]} \left(N_x - nq\left(x\right)\right)^2 - N_x\right] = n^2 \sum_{x\in[k]} \left(p\left(x\right) - q\left(x\right)\right)^2$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\sum_{x\in[k]} \left(N_x - nq\left(x\right)\right)^2 - N_x\right] &= \sum_{x\in[k]} \mathbb{E}\left[N_x\left(N_x - 1\right)\right] + n^2\left(q\left(x\right)\right)^2 - 2\mathbb{E}\left[N_x\right]q\left(x\right)\\ &= \sum_{x\in[k]} \left(np\left(x\right)\right)^2 + \left(nq\left(x\right)\right)^2 - 2n^2 p\left(x\right)q\left(x\right)\\ &= n^2 \sum_{x\in[k]} \left(p\left(x\right) - q\left(x\right)\right)^2\end{aligned}$$

□