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

Instructor: Jayadev Acharya — Lecture #2
Scribe: Huanyu Zhang — 28th August, 2017

1 Recap

$|\mathcal{X}| = k$, ε is an accuracy parameter, and δ is an error parameter.

2 Learning discrete distributions

TV-Estimation Problem: Given $X_1, X_2, ..., X_n$ independent samples drawn from an unknown distribution p over $[k]$, we need to output $\hat{p}$ s.t. with probability at least $1-\delta$, $d_{TV}(p, \hat{p}) < \varepsilon$. Here we assume $\delta = 0.1$ (for now).

Suppose we observe $X_1^n \stackrel{\text{def}}{=} X_1, X_2, ..., X_n$ from a distribution p over $\mathcal{X}$. Let

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

We define the empirical estimator $\hat{p_n}(x) = \frac{N_x}{n}$.

Theorem 1. *The empirical estimator satisfies*

$$\mathbb{E}_{X_1^n}\left[\ell_1(p, \hat{p_n})\right] \le \sqrt{\frac{k}{n}}$$

Lemma 2 (Cauchy-Schwarz Inequality). *let $a_1, ..., a_m, b_1, ..., b_m \in \mathbb{R}$, we have*

$$(\sum_{i=1}^m a_i \cdot b_i)^2 \le (\sum_{i=1}^m a_i^2) \cdot (\sum_{i=1}^m b_i^2)$$

The two sides are equal if and only if for all i, $a_i/b_i = c$.

Proof. Using CSI with $a_x = |p(x) - \hat{p_n}(x)|$, b_x=1,

$$\Big(\ell_1(p, \hat{p_n})\Big)^2 \le \Big(\sum_{x\in\mathcal{X}} (p(x) - \hat{p_n}(x))^2\Big) \cdot k$$

If we take expectation for both sides, we have

$$\mathbb{E}\Big[\ell_1(p, \hat{p_n})^2\Big] \le k \cdot \mathbb{E}\Big[\sum_{x\in\mathcal{X}} (\frac{N_x}{n} - p(x))^2\Big] \tag{1}$$

$$= \frac{k}{n^2} \cdot \mathbb{E}\Big[\sum_{x\in\mathcal{X}} (N_x - np(x))^2\Big] \tag{2}$$

$$= \frac{k}{n^2} \cdot \sum_{x\in\mathcal{X}} np(x)(1-p(x)) \tag{3}$$

$$\le \frac{k}{n} \tag{4}$$

The last two lines come from the fact that $N_x \sim \text{Bin}(n, p(x))$. So we have $\mathbb{E}[N_x] = np(x)$ and $\text{Var}(N_x) = np(x)(1 - p(x))$.

Because $f(x) = x^2$ is a convex function, according to Jensen's inequality, we get

$$\mathbb{E}\Big[\ell_1(p, \hat{p_n})\Big]^2 \leq \mathbb{E}\Big[\ell_1(p, \hat{p_n})^2\Big] \leq \frac{k}{n}$$

Lemma 3 (Markov's Inequality)**.** *If X is a nonnegative random variable and $a > 0$, then*

$$\text{Prob}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Using Markov's Inequality,

$$\text{Prob}\Big(\ell_1(p, \hat{p_n}) > \varepsilon\Big) \leq \frac{1}{\varepsilon}\sqrt{\frac{k}{n}}$$

Let $\frac{1}{\varepsilon}\sqrt{\frac{k}{n}} \leq 0.1$, we can get $n \geq 100 \cdot \frac{k}{\varepsilon^2}$. So if we use an empirical estimator, we get an upper bound of $O(\frac{k}{\varepsilon^2})$. □

3 Poisson Sampling

Poisson Sampling is a sampling method that produces independent N_x's without too much loss.

3.1 Properties of Poisson Distribution

If $X \sim \text{Poi}(\lambda_1)$, $Y \sim \text{Poi}(\lambda_2)$

1 PMF: $\text{P}(X = i) = e^{-\lambda_1} \cdot \frac{\lambda_1^i}{i!}$,

2 Mean and Variance: $\mathbb{E}[X] = \text{Var}(X) = \lambda_1$,

3 When $n \cdot p$ is fixed and $p \to 0$, $\text{Bin}(n, p)$ goes to $\text{Poi}(n \cdot p)$. To be specific, when $n \cdot p = \lambda$, $\lim_{p \to 0} \binom{n}{i} \cdot p^i (1-p)^{n-i} = e^{-\lambda} \cdot \frac{\lambda^i}{i!}$

4 $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$

3.2 Procedure for Poisson sampling

Fixed length sampling: We have a fixed sample size n and we draw $X_1, X_2, ..., X_n$ iid samples from distribution p, $N_x \sim \text{Bin}(n, p(x))$

Poisson length sampling:

1 $n' \sim \text{Poi}(n)$

2 Generate n' independent samples from p.

3.2.1 Properties of Poisson Sampling

1 $N'_x \sim \text{Poi}(n \cdot p(x))$.

Proof.

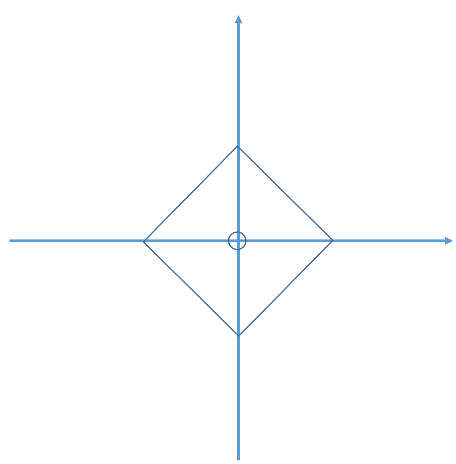
$$\begin{aligned}
\Pr(N'_x = j) &= \sum_{n'} \Pr\left(N'_x = j, n'\right) \\
&= \sum_{n' \geq j} e^{-n} \frac{n^{n'}}{n'!} \binom{n'}{j} (p(x))^j (1-p(x))^{n'-j} \\
&= e^{-n} \frac{(np(x))^j}{j!} \sum_{n' \geq j} \frac{n^{n'-j}(1-p(x))^{n'-j}}{(n'-j)!} \\
&= e^{-n} \frac{(np(x))^j}{j!} \sum_{n' \geq j} \frac{(n(1-p(x)))^{n'-j}}{(n'-j)!} \\
&= e^{-n} \frac{(np(x))^j}{j!} \cdot e^{n(1-p(x))} \\
&= e^{-np(x)} \frac{(np(x))^j}{j!}.
\end{aligned}$$

2 Condition on n', the distribution becomes fixed length with respect to parameter n'.

3 $\text{P}(N'_x = n_x, N'_y = n_y) = \text{P}(N'_x = n_x) \cdot \text{P}(N'_y = n_y)$

4 Testing Problem

Given description of a probability distribution q over $[k]$, parameter ε and n independent samples from an unknown distribution p, we want to know whether $p = q$ or $d_{TV}(p, q) > \varepsilon$. The following picture illustrates the case when $q = u[k]$. We need to distinguish between p is the origin or p lies outside the square.



Now we consider a special case when q is uniform. Given $\varepsilon > 0$ and n independent samples from p, we want to figure out, with probability at least 0.9, whether $p = q$ or $d_{TV}(p, q) > \varepsilon$.

Theorem 4. *Testing uniformity requires $\Omega(\sqrt{k})$ samples for any fixed ε.*

Before we look at the argument for this theorem, let us see the following lemma first.

Lemma 5 (Birthday Paradox)**.** *At least $\Omega(\sqrt{k})$ samples from $u[k]$ are needed before you can find a repeated symbol with some constant probability.*

You can prove this lemma by showing $\mathbb{E}[\#\textit{symbols appear more than } 1 \textit{ time}] < \frac{n^2}{k}$. Don't forget under Poisson Sampling, for every x, $N_x \sim \text{Poi}(n/k)$.

You can also try to prove the following result: At least $\Omega(k^{1-1/\alpha})$ samples from $u[k]$ are needed before you can find a symbol appear α times with some constant probability.

Now let us go back to the theorem. Recall that $P = u[k]$ is the uniform distribution on $[k]$. Let $u[k/2]$ be the collection of all distributions that are uniform over a subset of $k/2$ elements of k. There are $\binom{k}{k/2}$ distributions. Then note that: For any $q \in u[k/2]$, $d_{TV}(q, u[k]) = 0.5$. Let Q be the distribution uniformly drawn from $u[k/2]$. Then if we sample from $P = u[k]$ by $\sqrt{k}/10$ number of samples, all symbols are distinct. The same is true for Q. Hence we can't distinguish between P and Q with a constant probability.

4.1 Goldreich-Ron Algorithm

The algorithm is as follows: Let $T \stackrel{\text{def}}{=} \sum_{i<j} \mathrm{I}\{x_i = x_j\}$. If $T \geq \binom{n}{2}(\frac{1}{k} + \frac{\varepsilon^2}{2k})$, we output $d_{TV}(p, q) > \varepsilon$ else we output $p = q$.

Theorem 6. *The coincidence based test solves uniformity testing problem with $O(\frac{\sqrt{k}}{\varepsilon^4})$*

Proof. When p is a uniform distribution, the expectation of statistics T is:

$$\mathbb{E}[T|p = u] = \binom{n}{2} \cdot \sum_x p^2(x) \tag{5}$$

$$= \binom{n}{2} \cdot \frac{1}{k} \tag{6}$$

When $d_{TV}(p, q) > \varepsilon$, by using Jenson's inequality and Cauchy-Schwarz inequality,

$$\sum_x \left(p(x) - \frac{1}{k}\right)^2 \cdot k \geq \left(\sum_x |p(x) - \frac{1}{k}|^2\right) \geq \varepsilon^2$$

Besides,

$$\sum_x \left(p(x) - \frac{1}{k}\right)^2 = \sum_x p^2(x) - 2\sum_x \frac{p(x)}{k} + \frac{1}{k^2} \tag{7}$$

$$= \sum_x p^2(x) - \frac{1}{k} \tag{8}$$

Then we have

$$\sum_x p^2(x) \geq \frac{1 + \varepsilon^2}{k}$$

So the expectation of the statistics is:

$$\mathbb{E}[T|\ell_1(p,u)] = \binom{n}{2} \cdot \sum_x p^2(x) \tag{9}$$

$$\geq \binom{n}{2} \cdot \frac{1+\varepsilon^2}{k} \tag{10}$$

The following proof about bounding variance and using Chebychev's inequality will be covered in the next lecture. In the next lecture we will look at a statistic that gives an upper bound of $O(\sqrt{k}/\varepsilon^2)$ samples. □

5 Reference

- Mitzenmacher, Michael, and Eli Upfal. Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis.
- Paninski' 08 : http://www.stat.columbia.edu/ liam/research/pubs/sparse-unif-test.pdf