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

Instructor: Jayadev Acharya — Lecture #10-11
Scribe: JA — 27, 29 September, 2016

Please send errors to acharya@cornell.edu

1 Distribution Property Testing

The problem of testing whether a distribution has a property is a very old problem. It has been studied for a long time as Statistical Hypothesis testing `https://en.wikipedia.org/wiki/Statistical_hypothesis_testing`. In the general form, the problem can be stated as following:

- $\mathcal{P}$, and $\mathcal{Q}$ are two collections of distributions, with $\mathcal{P} \cap \mathcal{Q} = \emptyset$.
- Given independent samples from $P \in \mathcal{P} \cup \mathcal{Q}$.
- Design a (possibly randomized) function $C(X_1^n) \to \{\mathcal{P}, \mathcal{Q}\}$.
- The probability of error is defined as:

$$p_e^n = \max_{P \in \mathcal{P} \cup \mathcal{Q}, X_1^n \sim P} \left\{ \Pr\left(C(X_1^n) = \mathcal{Q} | P \in \mathcal{P}\right), \Pr\left(C(X_1^n) = \mathcal{P} | P \in \mathcal{Q}\right) \right\}.$$

In the traditional statistical literature the problem is studied in the following set-up:

- The domain size of the problem is small(eg, small k for discrete setting, or few parameters in continuous settings).
- The number of samples n grows.
- The asymptotic error is studies in this range, namely we characterize:

$$\lim_{n \to \infty} p_e.$$

Note that we are interested in the settings when the domain of the problem is potentially much larger than the number of samples we will get to observe. For example, given the outputs of a few lottery draws, can we say with some confidence that the lottery is actually uniform? This is the framework that was studied by Batu, Kumar, Rubinfeld (who were all at Cornell at the time). One of the problems they studied was the following:

- $\mathcal{P} = \{U[k]\}$, and $\mathcal{Q} = \{P \text{ over } [k] : d_{TV}(P, U[k]) \geq \varepsilon\}$.
- Instead of asymptotic, suppose we are interested in making the error probability less than a constant (say 0.1). What is the number of samples n such that

$$p_e^n \leq 0.1.$$

This is called as the problem of uniformity testing. The problem can be generalized when instead of uniformity, we want to test if a distribution is equal to a known distribution Q, or at least ε far from Q in total variation distance.

Let

$$n^*(\mathcal{P}, \mathcal{Q})$$

be the minimum number of samples such that there exists a testing algorithm with error at most 0.1

A general idea. A general method employed for distribution testing is the following. Design a test statistic $T(X_1^n) \to \mathbb{R}$, such that there is a threshold τ such that

1. If $P \in \mathcal{P}$ then $\Pr\left(T(X_1^n) > \tau\right) < 0.1$,
2. If $P \in \mathcal{Q}$ then $\Pr\left(T(X_1^n) < \tau\right) < 0.1$.

If such a statistic exists, we are done! We simply evaluate the statistic, compare it with τ and output the corresponding answer.

The goal is to find $n^*(\mathcal{P}, \mathcal{Q})$. The vanilla approach is something we already did in class, and is described in the next section.

1.1 Uniformity testing via learning

- Learn the distribution to total variation $\varepsilon/2$ (with probability at least 0.9) to obtain a distribution $\hat{P}$ over k.
- Output

$$C(X_1^n) = \begin{cases} U[k] & \text{if } d_{TV}(\hat{P}, U[k]) < \varepsilon/2 \\ \mathcal{Q} & \text{otherwise.} \end{cases}$$

By the triangle inequality, if $P = U[k]$, with probability at least 0.9, $d_{TV}(\hat{P}, P) < \varepsilon/2$. When $P \in \mathcal{Q}$, $d_{TV}(P, U[k]) > \varepsilon$, and $d_{TV}(P, \hat{P}) < \varepsilon/2$. By the triangle inequality of total variation, $d_{TV}(\hat{P}, U[k]) > \varepsilon/2$. This algorithm tests uniformity with error probability at most 0.1. The number of samples required is the number of samples required to learn a distribution to total variation $\varepsilon/2$. We saw in the first few lectures that this is possible using $O(k/\varepsilon^2)$ samples. Therefore,

Theorem 1. *There is an algorithm that takes $O(k/\varepsilon^2)$ samples and tests uniformity.*

1.1.1 Special case: Binary

In this case $k = 2$. We saw that we can learn a Bernoulli distribution to total variation $\varepsilon/2$ using $O(1/\varepsilon^2)$ samples. Moreover we also proved that to test whether a distribution is $Bern(0.5)$ or $Bern(0.5 + \varepsilon)$ requires at least $\Omega(1/\varepsilon^2)$ samples. Note that Bernoulli distribution are the special case of $k = 2$. Therefore, in this case we obtain:

Theorem 2. *For $k = 2$, and testing uniformity:*

$$n^*(U[k], \mathcal{Q}) = \Theta(1/\varepsilon^2).$$

However, our main interest is in the case when k is large.

1.2 Complexity of testing uniformity: Beating the learning bound

Interestingly, [GR00, BFF$^+$00] showed that using $O(\sqrt{k}/\varepsilon^4)$ samples, one can test uniformity over k elements.

The precise complexity of testing uniformity in the total variation distance was only resolved in 2008 by Paninski [Pan08] who showed that the optimal sample complexity is $O(\sqrt{k}/\varepsilon^2)$.

We will provide arguments that are found in [ADK15, DK16]. Canonne [Can15] provides an excellent survey of the field of distribution property testing, including uniformity testing.

Before we delve into the precise upper bound, let us first look at a simple lower bound, which captures the dependence on k, but not on ε.

1.2.1 A simple lower bound

Theorem 3. *Testing uniformity requires $\Omega(\sqrt{k})$ samples for any fixed ε.*

Suppose $\varepsilon = 0.5$. Consider the following subset of $\mathcal{Q}$.

Recall that $\mathcal{P} = \{U[k]\}$ is the uniform distribution on $[k]$. Let $\mathcal{U}[k/2]$ be the collection of all distributions that are uniform over a subset of $k/2$ elements of $[k]$. There are $\binom{k}{k/2}$ such distributions. Then note that:

Claim 4. *For any $P \in \mathcal{U}[k/2]$,*

$$d_{TV}(P, U[k]) = \frac{1}{2}.$$

Claim 5. *If we sample at most $Poi(n)$ samples for $n \leq k$, the number of samples that appear more than once is bounded by:*

$$\lambda^2.$$

Proof. Recall that $N_x \sim Poi(nP(x))$. Let $\lambda = nP(x)$. Then for $X \sim Poi(\lambda)$,

$$\begin{aligned}
\Pr(X > 1) &= \sum_{j \geq 2} e^{-\lambda} \frac{\lambda^j}{j!} \\
&< \sum_{j \geq 2} e^{-\lambda} \frac{\lambda^j}{2^{j-1}} \\
&= e^{-\lambda} \lambda^2
\end{aligned}$$

□

Let $N_{\geq 2}$ denote the number of symbols that appear more than once in X_1^n when we sample $Poi(n)$ times from $U[k]$. Then

$$\mathbb{E}[N_{\geq 2}] \leq k \cdot e^{-n/k} \left(\frac{n}{k}\right)^2 \leq \frac{n^2}{k} \tag{1}$$

Suppose $n < \sqrt{k}/10$, then $\mathbb{E}[N_{\geq 2}] < 1/100$. Since $N_{\geq 2}$ is an integer, By the Markov Inequality, with probability at least 0.99, $N_{\geq 2} = 0$. For sampling from a distribution in $\mathcal{U}[k/2]$,$\mathbb{E}[N_{\geq 2}] < 1/50$, and with probability at least 0.98, $N_{\geq 2} = 0$.

Therefore, when we sample from the uniform distribution at most $\sqrt{k}/10$ times, all symbols are distinct. The same is true for a random distribution from $\mathcal{U}[k/2]$. Hence, we cannot distinguish between these two cases with probability more than 0.6.

This is also called as the **Birthday Paradox**, since it is also true that once the number of samples is more than $\sqrt{k}$ we can expect to see collisions.

1.2.2 Upper bound

There is inherent difficulty in handling the total variation distance (sums of absolute values). It has been easier to handle ℓ_2 distance in these cases. We first use Cauchy Schwarz inequality to show how ℓ_2 norms and distance can be of interest.

Claim 6. *For P, Q over $[k]$, if $d_{TV}(P,Q) \geq \varepsilon$, then*

$$\sum_{x \in \mathcal{X}} (P(x) - Q(x))^2 \geq \frac{\left(\sum_{x \in \mathcal{X}} |P(x) - Q(x)|\right)^2}{k} \geq \frac{4\varepsilon^2}{k}.$$

For the special case when $P = U[k]$,

$$\sum_{x \in \mathcal{X}} P(x)^2 = \frac{1}{k}, \tag{2}$$

and if $d_{TV}(P, U[k]) \geq \varepsilon$, then

$$\sum_{x \in \mathcal{X}} \left(P(x) - \frac{1}{k}\right)^2 = \sum_{x \in \mathcal{X}} P(x)^2 - \frac{1}{k} \geq \frac{4\varepsilon^2}{k}. \tag{3}$$

With these in mind, we now propose the test statistic:

$$T = \sum_{x \in \mathcal{X}} (N_x - \frac{n}{k})^2 - N_x. \tag{4}$$

While it may seem a little out of the blue, the following result shows that it is not.

Lemma 7. *Suppose we sample P $Poi(n)$ times, and Q is any distribution, then*

$$\mathbb{E}\left[\sum_{x \in \mathcal{X}} (N_x - nQ(x))^2 - N_x\right] = n^2 \sum_{x \in \mathcal{X}} (P(x) - Q(x))^2.$$

Proof. Recall that $N_x \sim Poi(nP(x))$. Therefore,

$$\mathbb{E}\left[\sum_{x \in \mathcal{X}} (N_x - nQ(x))^2 - N_x\right] = \mathbb{E}\left[\sum_{x \in \mathcal{X}} N_x(N_x - 1) - 2nN_xQ(x) + n^2Q(x)^2\right] \tag{5}$$

$$= \sum_{x \in \mathcal{X}} n^2 (P(x) - Q(x))^2. \tag{6}$$

□

Therefore, we have for $P = U[k]$,

$$\mathbb{E}[T] = 0,$$

and for $d_{TV}(P, U[k]) > \varepsilon$,

$$\mathbb{E}[T] = n^2 \sum_{x \in \mathcal{X}} (P(x) - Q(x))^2 \geq \frac{n^2}{k} \cdot 4\varepsilon^2.$$

We are now in a position to state the algorithm.

UNIFORMITY TESTING ALGORITHM.

1. Obtain $Poi(n)$ samples from P.
2. $$Output = \begin{cases} \text{uniform} & \text{if } T < \tau \stackrel{\text{def}}{=} \frac{n^2}{k} \cdot 2\varepsilon^2, \\ \text{not uniform} & \text{if } T \geq \tau. \end{cases}$$

We want to show that with sufficient samples, the test statistic is below the threshold. In particular, we want to show that

1. For uniform P, $\Pr(T \geq \tau) < 0.1$.
2. For non-uniform P, $\Pr(T \leq \tau) < 0.1$.

Case 1: $P = U[k]$ We plan to use Chebychev's inequality. Note that $\mathbb{E}[T] = 0$ in this case. We need to bound the variance. We use the following claim, which we encouraged in class to prove.

Claim 8. *If $X \sim Poi(\lambda)$, and μ is any real number, then*

$$Var\left((X - \mu)^2 - X\right) = 2\lambda^2 + 4\lambda(\lambda - \mu)^2.$$

Therefore, when $P = U[k]$, using $\lambda = \mu = n/k$, and the independence of N_x's,

$$Var(T) = 2 \sum_{x \in \mathcal{X}} (n/k)^2 = 2\frac{n^2}{k}.$$

Applying Chebychev's inequality,

$$\Pr\left(T > \sqrt{10} \cdot \sqrt{2\frac{n^2}{k}}\right) < \frac{1}{10}.$$

Therefore, if $\tau > \sqrt{10} \cdot \sqrt{2\frac{n^2}{k}}$, T does not exceed it with probability at least 0.9. This happens when

$$\frac{n^2}{k} \cdot 2\varepsilon^2 \geq \sqrt{10} \cdot \sqrt{2\frac{n^2}{k}},$$

which holds when $n > \sqrt{5} \cdot \sqrt{k}/\varepsilon^2$.

We will continue the argument next time.

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NIPS*, 2015.

[BFF+00] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing random variables for independence and identity. In *FOCS*, pages 259–269, 2000.

[Can15] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, page 7, 2015.

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. *arXiv preprint arXiv:1601.05557*, 2016.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.