COMS E6998-9: Algorithms for Massive Data (Spring'19) Mar 7, 2019

Lecture 14: Distribution Testing

Instructor: *Alex Andoni* Scribes: *Tianyang Yang*

1 Introduction

Sublinear Time Algorithm: Only Look at a subset of the input:
• take a subset of data and return an approximate output, or
• test a hypothesis under some probability (property tesing).
We will look at the problem of Distribution Testing.

Distribution Testing: We have access to m samples, $x_1, x_2, ..., x_m \sim D$, and want to do some hypothesis test on distribution D with the samples. The goal is to minimize the sample size m.

2 Uniformity Testing

One kind of distribution testing is to test the Uniformity of distribution, i.e., to test whther D is uniform over domain $[n]$. Distribution D is uniform distribution if

$$D_i = \frac{1}{n}, \quad \forall i \in [n].$$

It will require large runtime to exactly distinguish the uniformity, so we turn to use sublinear algorithm for appromate testing. The **Approximate Problem** is to distinguish between

- D is uniform;
- D is "far" from uniform.

Definition 1. *Distributions D and Q are ε-far if*

$$\|D - Q\|_1 \geq \varepsilon.$$

This is equivalent to compute the Total Variance Distance between D and Q (see **Claim 3**).

Definition 2. ***Total Variance Distance*** *between D and Q is define as*

$$TV(D, Q) = \max_{T \subset [n]} |Pr_{x \sim D}[x \in T] - Pr_{x \sim Q}[x \in T]|$$

Claim 3. $TV(D, Q) = \frac{1}{2}\|D - Q\|_1$.

Here is some intuition behind TV Distance. For example, we have only 1 sample to distinguish whether sample x is from distribution D or Q. Let $t(x)$ be the result distribution of our testing, and T be the set where we accept $x \sim D$. Take some intuition from machine learning: we want to accept D as the output distribution when the probability of that $x \sim D$ is larger than that of Q. Without knowledge of prior probability, we should compare the probability $Pr_D(x)$ and $Pr_Q(x)$, and accept D when $Pr_D(x) > Pr_Q(x)$. Thus, we have

$$T = \{x : Pr_D(x) > Pr_Q(x)\}$$

and this is actually equivalent to

$$T \in arg\max_T \Big(Pr_D[x \in T] - Pr_Q[x \in T] \Big)$$

Since the sum of probability over $[n]$ is 1 for both D and Q, we have

$$[n] \backslash T \in arg\max_{T'} \Big(Pr_Q[x \in T'] - Pr_D[x \in T'] \Big)$$

and

$$\begin{aligned} &\max_T \Big(Pr_D[x \in T] - Pr_Q[x \in T] \Big) \\ =&\max_{T'} \Big(Pr_Q[x \in T'] - Pr_D[x \in T'] \Big) \\ =&\max_T |Pr_D[x \in T] - Pr_Q[x \in T]| \end{aligned}$$

Since $\Big(\max_T Pr_D[x \in T] - Pr_Q[x \in T] \Big) + \Big(\max_{T'} Pr_Q[x \in T'] - Pr_D[x \in T'] \Big) = \|D - Q\|_1$, we get

$$\max_T |Pr_D[x \in T] - Pr_Q[x \in T]| = \frac{1}{2}\|D - Q\|_1,$$

that is,

$$TV(D, Q) = \frac{1}{2}\|D - Q\|_1.$$

2.1 Attempt 1

We define **Empirical Distribution** of D on sample $\{x_i\}_{i=1}^m$ as

$$\hat{D}_i = \frac{\sum_j \mathbf{1}[x_j = i]}{m}, \quad \forall i \in [n].$$

Then we test the uiformity hypothesis by

- if $\|U - \hat{D}\| << \varepsilon$, accept that D is uniform;
- otherwise, D is not uiform.

Claim 4. *We can test uniformity with $m >> \Omega_\varepsilon(n \log n)$.*

Proof. A sketch of proof is as below

$$\begin{aligned}
&|\hat{D}_i - D_i| < \frac{\varepsilon}{n} \ , \quad \text{with high prob.} \\
\Rightarrow &\|\hat{D} - D\|_1 = \sum_{i\in[n]} |\hat{D}_i - D_i| \leq \frac{\varepsilon}{3} \ , \quad \text{with high prob.} \\
\Rightarrow &\|\hat{D} - U\| = \|D - U\| \pm \frac{\varepsilon}{3} \ , \quad \text{with high prob..}
\end{aligned}$$

□

Claim 5. *$m = O_\varepsilon(n)$ samples are enough as well for testing uniformity.*

Proof.

$$\begin{aligned}
&E\{\|D - \hat{D}\|_1\} \\
=&\sum_{i\in[n]} E\{|D_i - \hat{D}_i|\} \\
\leq&\sum_{i\in[n]} \Big(E\{|D_i - \hat{D}_i|^2\}\Big)^{\frac{1}{2}} \\
=&\sum_{i\in[n]} \Big(var\{\hat{D}_i\}\Big)^{\frac{1}{2}} \\
=&\sum_{i\in[n]} \Big(var\{\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[x_j = i]\}\Big)^{\frac{1}{2}} \\
=&\sum_{i\in[n]} \Big(\frac{1}{m^2}\sum_{j=1}^{m} var\{\mathbf{1}[x_j = i]\}\Big)^{\frac{1}{2}} \\
\leq&\sum_{i\in[n]} \Big(\frac{1}{m} D_i\Big)^{\frac{1}{2}} \\
=&\frac{1}{\sqrt{m}}\Big(\sum_{i\in[n]} D_i\Big)^{\frac{1}{2}}\sqrt{n} \\
\leq&\frac{1}{\sqrt{m}}\Big(\sum (D_i^{\frac{1}{2}})^2\Big)^{\frac{1}{2}}\sqrt{n} \\
=&\sqrt{\frac{n}{m}}\Big(\sum_{i\in[n]} D_i\Big)^{\frac{1}{2}} \\
=&\sqrt{\frac{n}{m}}
\end{aligned}$$

We want $\sqrt{\frac{n}{m}} < \frac{\varepsilon}{30}$, therefore

$$m = \Omega(\frac{n}{\varepsilon^2})$$

□

2.2 Attempt 2

Let $C := \#\{i < j : x_1 = x_j\}$, C is the collision count. We test the uniformity by

- if $\frac{C}{\binom{m}{2}} < \frac{\alpha}{n}$ for some constant $\alpha = \alpha(\varepsilon)$, then D is uniform;
- otherwise, D is ε-far from uniform.

Analysis: We analyze the l-2 distance:

- if $D = U$, we have $\|D - U\|_1 = 0$ and $\|D - U\|_2 = 0$;
- if $\|D - U\|_1 \geq \varepsilon$, we have $\|D - U\|_2 \geq \frac{\|D-U\|_1}{\sqrt{n}} \geq \frac{\varepsilon}{\sqrt{n}}$, then $\|D - U\|_2^2 \geq \frac{\varepsilon^2}{n}$.

Claim 6. $\|D - U\|_2^2 = \|D\|_2^2 - \frac{1}{n}$.

Proof.

$$
\begin{aligned}
&\|D - U\|_2^2 \\
=&\sum_i (D_i - \frac{1}{n})^2 \\
=&\sum_i (D_1^2 - \frac{2D_i}{n} + \frac{1}{n^2}) \\
=&\|D\|_2^2 - \frac{1}{n}.
\end{aligned}
$$

□

Now our problem becomes to distinguish between
• $\|D\|_2^2 = \frac{1}{n}$, then uniform;
• $\|D\|_2^2 \geq \frac{1}{n} + \frac{\varepsilon}{n}$, then ε-far from uniform.
The following claims will show the correctness to use $\frac{C}{\binom{m}{2}}$ for testing.

Claim 7. $E(\frac{C}{\binom{m}{2}}) = \|D\|_2^2$.

Proof.

$$
\begin{aligned}
E(C) &= \sum_{i<j} Pr[x_i = x_j] \\
&= \sum_{i<j} \sum_{k\in[n]} D_i^2 \\
&= \binom{m}{2} \|D\|_2^2.
\end{aligned}
$$

That is, we have $E(\frac{C}{\binom{m}{2}}) = \|D\|_2^2$. □