6.889 Sublinear Time Algorithms Thursday, October 8, 2019

Lecture 12: Distribution Testing - Uniformity

Lecturer: Ronitt Rubinfeld *Scribe: Isabella Kang*

1 Introduction to Distribution Testing

In the first half of this class, we've focused mainly on testing graph properties, ie. average degree, bipartiteness, planarity, etc. In this lecture, we introduce property testing of probability *distributions.* We begin with some probability distribution $\mathcal{P}$ over a discrete domain D, where $|D| = n$. We know the size of n, but for all $i \in [n]$ where $[n]$ denotes $\{1, 2, ..., n\}$, we do not know $\Pr(i)$ for the distribution $\mathcal{P}$.

Our new model assumes we have an *oracle* that can sample IID random variables from the probability distribution of interest $\mathcal{P}$. We are interested in learning the shape of $\mathcal{P}$, such as whether the distribution is uniform, monotone increasing, or k-modal, and the properties of this distribution, such as whether it has high entropy or large support (having many distinct elements appearing with a nonzero probability). Our goal is to estimate these properties with a sublinear number of queries to our oracle with respect to the size of n. This lecture focuses on testing whether an unknown distribution is close to the uniform distribution.

2 Testing Uniformity

Given an unknown distribution $\mathcal{P}$ and its domain $D = [n]$, we would like to test whether $\mathcal{P}$ is close to the uniform distribution over D, which we denote $\mathcal{U}_D$. We seek to create a tester with the following properties:

- If $\mathcal{P} = \mathcal{U}_D$, we pass with probability at least $\frac{3}{4}$.
- If $dist(\mathcal{P}, \mathcal{U}_D) > \varepsilon$, we fail.

Note that our tester depends on what metric we choose to use to measure *distance* between $\mathcal{P}$ and $\mathcal{U}_D$, and today we will focus on two metrics, ℓ_1 and ℓ_2 distance.

2.1 ℓ_1 and ℓ_2 Distance

We are given two discrete probability distributions $\mathcal{P}$ and $\mathcal{Q}$, and we assume their domains are both $D = [n]$. Let samples s_P and s_Q be randomly drawn from these distributions, respectively. We will define p_i and q_i as $\Pr(s_P = i)$ and $\Pr(s_Q = i)$. Then we have the following definitions for ℓ_1 and ℓ_2 distance between $\mathcal{P}$ and $\mathcal{Q}$.

Definition 1 (ℓ_1 distance)**.** *We define ℓ_1 distance as*

$$||\mathcal{P} - \mathcal{Q}||_1 = \sum_{i \in D} |p_i - q_i|.$$

Definition 2 (ℓ_2 distance)**.** *We define ℓ_2 distance as*

$$||\mathcal{P} - \mathcal{Q}||_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}.$$

Note that

$$||\mathcal{P} - \mathcal{Q}||_2 \leq ||\mathcal{P} - \mathcal{Q}||_1 \leq \sqrt{n} \cdot ||\mathcal{P} - \mathcal{Q}||_2$$

where the first inequality holds because

$$||\mathcal{P}-\mathcal{Q}||_2^2=\sum_{i=1}^{n}|p_i-q_i|^2\leq\sum_{i=1}^{n}|p_i-q_i|^2+2\sum_{i,j,i<j}|p_i-q_i||p_j-q_j|=\Big(\sum_{i=1}^{n}|p_i-q_i|\Big)^2=||\mathcal{P}-\mathcal{Q}||_1^2$$

and the second inequality holds due to the Cauchy-Schwartz inequality.

Example 1

Consider the probability distributions $\mathcal{P}$ and $\mathcal{Q}$ over $[n]$ as follows:

- $\mathcal{P}=(1,0,0,...,0)$
- $\mathcal{Q}=(\frac{1}{n},\frac{1}{n},...,\frac{1}{n})$

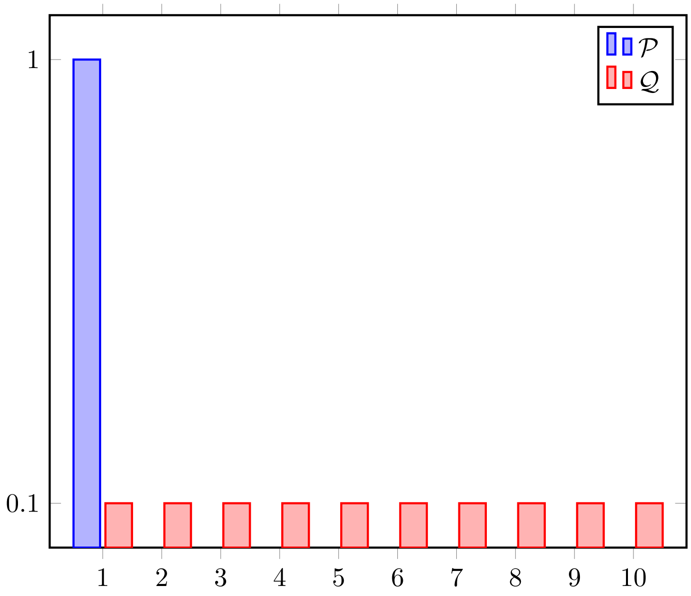

Figure 1: Sample probability distributions when $n=10$.

Then we can calculate the ℓ_1 and ℓ_2 distances as follows:

$$||\mathcal{P}-\mathcal{Q}||_1=(1-\frac{1}{n})+(n-1)\cdot\frac{1}{n}\approx 2$$

$$||\mathcal{P}-\mathcal{Q}||_2=(1-\frac{1}{n})^2+(n-1)\cdot\frac{1}{n^2}\approx 1$$

Example 2

Now consider the disjoint probability distributions $\mathcal{P}$ and $\mathcal{Q}$ over $[n]$:

- $\mathcal{P}=(\frac{2}{n},\frac{2}{n},...,\frac{2}{n},0,0,...,0)$
- $\mathcal{Q}=(0,0,...,0,\frac{2}{n},\frac{2}{n},...,\frac{2}{n})$

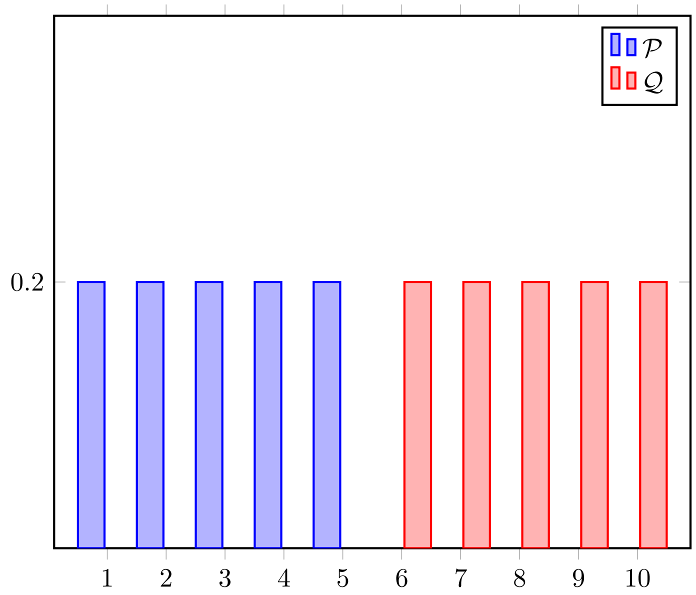

Figure 2: Sample probability distributions when $n = 10$.

Then we can calculate the ℓ_1 and ℓ_2 distances as follows:

$$||\mathcal{P} - \mathcal{Q}||_1 = n \cdot \frac{2}{n} = 1$$

$$||\mathcal{P} - \mathcal{Q}||_2 = \sqrt{n \cdot (\frac{2}{n})^2} = \frac{2}{\sqrt{n}}$$

It is interesting to note that in the second example, the ℓ_2 distance is quite small despite the two distributions being completely disjoint.

3 Plug-In Estimate for ℓ_1 Distance

Our first naive algorithm involves sampling our distribution $\mathcal{P}$, and dividing the number of times we get an element by the total number of samples. These estimates form our sample distribution $\hat{\mathcal{P}}$.

```
Algorithm 1: Plug-In Estimate
Input: $\varepsilon$
Take $m$ samples from $\mathcal{P}$.
Estimate $p_i$ with $\hat{p}_i = \frac{\text{# of times i occurs in sample}}{m}$.
If $\sum_{i=1}^{n} |\hat{p}_i - \frac{1}{n}| > \varepsilon$ we reject.
Otherwise, accept.
```

Naive Analysis

In our first attempt, we will try to pick a number of samples m such that for all elements $i \in D$, we have that $|\hat{p}_i - p_i| < \frac{\varepsilon}{2n}$. Then if we sum over all n elements in our domain, we have that $||\hat{\mathcal{P}} - \mathcal{P}||_1 < \frac{\varepsilon}{2}$. By the triangle inequality, we have that if $||\hat{\mathcal{P}} - \mathcal{P}||_1 < \frac{\varepsilon}{2}$ and $||\hat{\mathcal{P}} - \mathcal{U}_D||_1 < \frac{\varepsilon}{2}$ then $||\mathcal{P} - \mathcal{U}_D||_1 < \varepsilon$. Thus if $||\mathcal{P} - \mathcal{U}_D||_1 > \varepsilon$, we are likely to fail since there is some element that differs significantly from $\mathcal{U}_D$.

How large does m need to be in order for the inequality above to hold? Do we need to see each $i \in D$ at least once? Do we need to see them $\log n$ times? If we need to see each i at least once, we need $\Theta(n \log n)$ samples, but we can actually do much better than that. In fact, we will now show that we only need $O(n)$ samples.

Theorem 3. *We can approximate any distribution to within ε with respect to ℓ_1 distance with high probability in $O(\frac{n}{\varepsilon^2})$ samples.*

Proof. We first show that $\mathbb{E}||\hat{\mathcal{P}} - \mathcal{P}||_1 \leq \sqrt{\frac{n}{m}}$. Then we can simply take $m = \frac{c^2 \cdot n}{\varepsilon^2}$, and our right side becomes $\frac{\varepsilon}{c}$. By Markov's Inequality, we see that

$$Pr(||\hat{\mathcal{P}} - \mathcal{P}||_1 \geq \varepsilon) \leq \frac{\mathbb{E}||\hat{\mathcal{P}} - \mathcal{P}||_1}{\varepsilon}$$

which implies that

$$Pr(||\hat{\mathcal{P}} - \mathcal{P}||_1 < \varepsilon) \geq 1 - \frac{1}{c}$$

and if we choose c to be 4, we get our desired probability of passing/failing correctly with probability at least $\frac{3}{4}$, which would complete our proof of the theorem.

Thus we proceed with showing that $\mathbb{E}||\hat{\mathcal{P}} - \mathcal{P}||_1 \leq \sqrt{\frac{n}{m}}$. We have that

$$\begin{aligned}
\mathbb{E}||\hat{\mathcal{P}} - \mathcal{P}||_1 &= \sum_i \mathbb{E}\big[|\hat{p}_i - p_i|\big] \\
&\leq \sqrt{\sum_i \mathbb{E}(\hat{p}_i - p_i)^2} && \text{(Jensen's Inequality)} \\
&= \sum_i \sqrt{Var(\hat{p}_i)} && (\mathbb{E}(\hat{p}_i) = p_i) \\
&\leq \sum_i \sqrt{\frac{p_i}{m}} \\
&\leq \sqrt{\frac{n}{m}} && \text{(Cauchy-Schwartz Inequality)}
\end{aligned}$$

The second inequality holds because $Var(\hat{p}_i) = \frac{1}{m^2} \cdot m \cdot p_i(1-p_i) = \frac{p_i(1-p_i)}{m} \leq \frac{p_i}{m}$.

Hence we must take $\Theta(\frac{n}{\varepsilon^2})$ samples in order to approximate our distribution with high probability, which is not sublinear in n. □

4 Estimating ℓ_2 Distance

Now we'd like to estimate closeness of our unknown distribution $\mathcal{P}$ to the uniform distribution with respect to ℓ_2 distance. Again, assume that our domain D is $[n]$, and n is known. We can simplify closeness to ℓ_2 distance with the following algebraic manipulation:

$$
\begin{aligned}
||\mathcal{P} - \mathcal{U}_D||^2 &= \sum_{i=1}^{n}(p_i - \frac{1}{n})^2 \\
&= \sum(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2}) \\
&= \sum p_i^2 - \frac{2}{n}\sum p_i + \sum \frac{1}{n^2} \\
&= \sum p_i^2 - \frac{2}{n} + \frac{1}{n} \\
&= \sum p_i^2 - \frac{1}{n}
\end{aligned}
$$

Since we know n, we know what the second term $\frac{1}{n}$ is. Now we look at the first term, $\sum p_i^2$. Note that this term is equivalent to the probability that two samples drawn independently from $\mathcal{P}$ are the same. We define this probability as the *collision probability* of $\mathcal{P}$. Note that $\sum p_i^2 = ||\mathcal{P}||_2^2$ must be at least $\frac{1}{n}$ since we know $||\mathcal{P} - \mathcal{U}_D||^2 \geq 0$, which means that the uniform distribution has the smallest possible collision probability over all distributions.

Our simplified form for ℓ_2 distance naturally proposes an idea for the algorithm where we try to estimate the collision probability $\hat{c}$ of $\mathcal{P}$ from repeated samples from our oracle, then we accept if $\hat{c}$ is within some small δ of the collision probability for the uniform distribution, $\frac{1}{n}$.

How many samples do we need, and how small should we make our δ? We claim that the inequality $||\mathcal{P} - \mathcal{U}_D|| < \varepsilon$ is satisfied when $\hat{c} < \frac{1}{n} + \delta$ and we assume that $|\hat{c} - ||\mathcal{P}||^2| < \delta$, and we choose $\delta = \frac{\varepsilon^2}{2}$.

Assumption 4. *We have taken a large enough number of samples s such that $|\hat{c} - ||\mathcal{P}||^2| < \delta$ holds with probability at least $\frac{3}{4}$.*

We will prove this statement in the next lecture, but assume for now that it holds. Then we can prove the following claim.

Claim 5. *We have that $||\mathcal{P} - \mathcal{U}_D|| < \varepsilon$ is satisfied with high probability when $\hat{c} < \frac{1}{n} + \delta$ and the above assumption holds.*

Proof. If $\mathcal{P} = \mathcal{U}_D$, then $\hat{c} \leq ||\mathcal{P}||_2^2 + \frac{\varepsilon^2}{2} \leq \frac{1}{n} + \frac{\varepsilon^2}{2}$ so we accept with probability at least $\frac{3}{4}$. If $||\mathcal{P} - \mathcal{U}_D|| > \varepsilon$ then $||\mathcal{P} - \mathcal{U}_D||_2^2 > \varepsilon^2$. Since $||\mathcal{P}||^2 = \frac{1}{n} + ||\mathcal{P} - \mathcal{U}_D|| > \frac{1}{n} + \varepsilon^2$ so $\hat{c} > ||\mathcal{P}||^2 - \delta > \varepsilon^2 + \frac{1}{n} - \delta = \frac{1}{n} + \frac{\varepsilon^2}{2}$ and we reject with probability at least $\frac{3}{4}$. □

A naive implementation of estimating $\hat{c}$ involves repeatedly taking pairs of samples and for each of these pairs, counting the number of pairs that collide, and dividing by the total number of pairs. However, if we take k samples, we see only $\Theta(k)$ pairs of collisions, which means that we might need at least $\Omega(n)$ samples in order to see a collision. Thus we'd like to *recycle* by looking at *all* the pairs in a sample, which gives $\Theta(k^2)$ samples that may collide from k samples of $\mathcal{P}$.

Algorithm 2: Recycling Method Estimate

```
Input: $\varepsilon$
$\delta \leftarrow \frac{\varepsilon^2}{2}$
Take $s$ samples from $\mathcal{P}$.
Count the total number of collisions $c$ between any pair of samples.
Put $\hat{c} \leftarrow \frac{c}{\binom{s}{2}}$
If $\hat{c} < \frac{1}{n} + \delta$, accept. Otherwise, fail.
```

Analysis

Define $\sigma_{i,j}$ as 1 if samples s_i and s_j collide, and 0 otherwise. Then we have that

$$\mathbb{E}(\hat{c}) = \frac{\mathbb{E}(\sum_{i<j} \sigma_{i,j})}{\binom{s}{2}} = \frac{\sum_{i<j} \mathbb{E}(\sigma_{i,j})}{\binom{s}{2}} = \frac{\binom{s}{2}}{\binom{s}{2}} \mathbb{E}(\sigma_{i,j}) = Pr(\sigma_{i,j} = 1) = ||\mathcal{P}||^2$$

Then by Chebyshev's Inequality, we have that

$$\Pr(|\hat{c} - ||\mathcal{P}||^2| > \delta) \leq \frac{Var(\hat{c})}{\delta^2}$$

Now we will state another lemma that will be proved in the next lecture, but we'll assume it holds for now.

Lemma 6. $Var(\sum_{i<j} \sigma_{i,j}) \leq 4(\binom{s}{2}||\mathcal{P}||^2)^{3/2}$

We know that $Var(\hat{c}) = \frac{1}{\binom{s}{2}^2} Var(\sum_{i<j} \sigma_{i,j})$ from the way we defined $\hat{c}$ before, so we can combine this with the lemma to get that

$$Var(\hat{c}) \leq \frac{1}{\binom{s}{2}^2} \cdot 4(\binom{s}{2}||\mathcal{P}||^2)^{3/2} = \Theta(\frac{||\mathcal{P}||_2^3}{s})$$

which means that we need to pick s in a way such that it depends on $||\mathcal{P}||_2^3$. In the next lecture, we prove the lemma and show that we only need s to be $O(\frac{1}{\varepsilon^4})$.