CS880: Advanced Learning Theory **Fall 2019**

Lecture 3: Uniformity Testing

Lecturer: Ilias Diakonikolas *Scribes: Nikos Zarifis*

1 Introduction

In the testing setting, we want to test if a distribution has a global property. Properties of interest include uniformity, identity, monotonicity, log-concavity, etc. In this lecture, we are going to develop an algorithm that tests if a discrete distribution is the discrete uniform distribution on its domain. Let $[n]$ be the domain of the underlying distribution p and $\mathcal{U}_n$ denote the uniform distribution on $[n]$. Let $p : [n] \to [0, 1]$ be the probability mass function ($p_i \geq 0$ and $\sum_i p_i = 1$). So, we are looking for an algorithm that outputs *YES* if the distribution is uniform with high probability else if p is far from $\mathcal{U}_n$, then the algorithms output *NO*. More formally we have

Uniformity Testing Problem
Input: set of iid samples from unknown discrete distribution p supported on $[n]$
Output:

- *YES* if $p = \mathcal{U}_n$ with probability $1 - \delta$
- *NO* if $\mathrm{d_{TV}}(p, \mathcal{U}_n) \geq \epsilon$ with probability $1 - \delta$

2 Boosting confidence

Definition 1. *Let $S : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ be a function that given the number of elements and $\epsilon > 0$ outputs the number of samples that needed for a hypothesis testing problem with parameter ϵ and with high constant probability.*

In this section, we are going to show how to increase an algorithm's constant success probability. We are going to assume that $\mathcal{A}(S)$ be an algorithm that given input S gives the correct answer with constant probability $\geq 1 - c$ and the wrong with $c < 1/2$.

Theorem 2. *Let $\mathcal{A}$ be a testing algorithm that uses $S(n,\epsilon)$ samples and with probability at most $c < 1/2$ gives the wrong answer. Then there is an Algorithm $\mathcal{H}$ that uses $O(S(n,\epsilon)\log(1/\delta))$ samples and outputs the wrong answer with probability less than δ*

Proof. We construct $\mathcal{H}$ as follows : run $\mathcal{A}$ t times and let X_i be the result for each time i. If we output the majority, then the probability of failure will be

$$\Pr[\,\mathcal{H}\,] \leq \prod_{i}^{t/2} \Pr[\,X_i\,] \leq c^{t/2} \leq 2^{-t/2}$$

By setting $t = 2\log(1/\delta)$ we have that $\Pr[\,\mathcal{H}\,] \leq \delta$ □

3 Black Box Reduction

We are going to show how we can solve the uniformity testing problem using as a black box routine the Density Estimation. From the previous lecture we know that the Density Estimation given samples from a distribution p, returns an hypothesis $\hat{h}$ such as $\mathrm{d_{TV}}\left(p,\hat{h}\right) \leq \epsilon$ using $O\left(\frac{n}{\epsilon^2}\right)$ samples.

Algorithm Uniformity Testing using density estimation
Input: set of iid samples from unknown discrete distribution p

1. Run Density Estimation on p and output hypothesis $\hat{h}$ such as: $\mathrm{d_{TV}}\left(p,\hat{h}\right) \leq \epsilon/3$
2. Output:
   - *YES* if $\mathrm{d_{TV}}\left(\hat{h},\mathcal{U}_n\right) \leq \epsilon/3$
   - *NO* if $\mathrm{d_{TV}}\left(\hat{h},\mathcal{U}_n\right) \geq \epsilon/2$

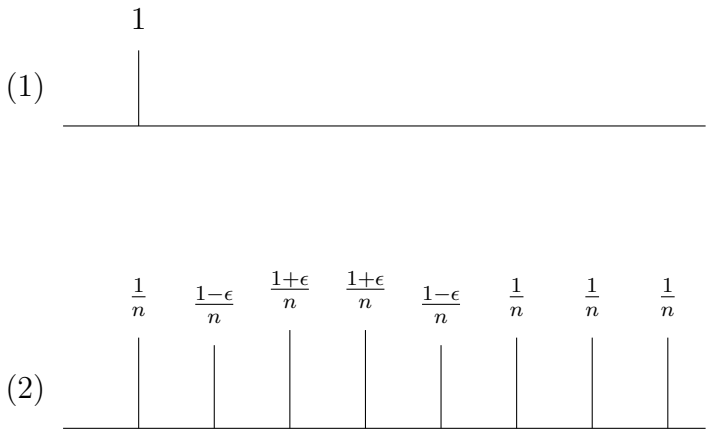
Analysis:

- *Completeness:* If $p = \mathcal{U}_n$ then $\mathrm{d_{TV}}\left(p,\hat{h}\right) \leq \epsilon/3$.
- *Soundness:* If $\mathrm{d_{TV}}(p,\mathcal{U}_n) \geq \epsilon$ then $\mathrm{d_{TV}}\left(\hat{h},\mathcal{U}_n\right) \geq \epsilon/2$ (triangle inequality)

But Uniformity testing needs less information than learning the density of the distribution, so there must be a way to get our answer using fewer samples.

4 Collision-based Uniformity Testing

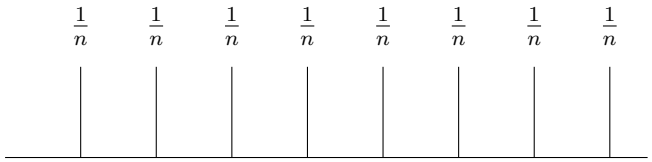
The first question that we need to answer is what property of the uniform distribution can help us develop an algorithm for Uniformity testing. For instance, if we have a distribution that $p_1 = 1$ and $p_k = 0$ for all k (as in Figure 1) then it is easy to determine that this distribution is not uniform as all the samples will be the first element. But what are we going to do if the distribution is close to the uniform? If we have a distribution with some $p_i = \frac{1+\epsilon}{n}$ and some other with $p_i = \frac{1-\epsilon}{n}$, then the problem is more difficult (Figure 2).



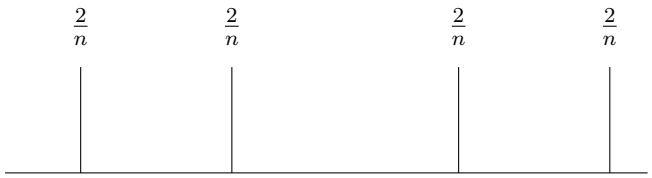
So, it is clear that we need to find a statistic $F(i_1, i_2, ..., i_s)$ that will be symmetric, meaning that only the number of times each element appears matters (i.e., 0, 1, ..., s times). This means that the statistic needs to be invariant under permutations.

Let's say we have two distributions: distribution A which is uniform (as in Figure 3) and distribution B which is supported on a set K that is created by selecting at random half of the domain of A (as in Figure 4). We have that the $\mathrm{d_{TV}}(A, B) = 1/2$ but how many samples are needed to distinguish between these two? The event that two samples drawn according to p are equal, is called collision. It's clear that with high probability we expect more collisions on distribution B than A for the same sample size. Let our statistic F be the number of collisions. We argue that the Uniform minimizes the number of collisions (see Fact 5). By the birthday paradox, we need at least $c\sqrt{n}$ samples to see the same element twice with high probability, thus we need at least $\Omega(\sqrt{n})$ samples to test even for $\epsilon = 1/2$, using a collision tester.

(3)



(4)



Algorithm Uniformity Testing using collisions
Input: s iid samples from unknown discrete distribution p on $[n]$

1. Let $C \leftarrow$ # collisions
2. Let $T(\epsilon, n) \leftarrow \frac{1+\epsilon^2/2}{n}\binom{s}{2}$
3. Output:
   - **If** $C < T(\epsilon, n)$ **then** output *YES*
   - **else** output *NO*

Theorem 3. *The Collision Testing algorithm, when given $m = O(\sqrt{n}/\epsilon^4)$ samples from p over $[n]$ will, with probability at least $3/4$, distinguish the case that $p = \mathcal{U}_n$ from the case that $\mathrm{d_{TV}}(p, \mathcal{U}_n) \geq \epsilon/2$.*

Analysis: The analysis of our algorithm consists of three steps.

1. Analyze the Expected value of C.
2. Analyze the Variance of C.
3. Using 1,2 and Chebyshev inequality to complete our proof.

Let $C = \sum_{i<j}^{s} \sigma_{ij}$, where $\sigma_{ij} = \begin{cases} 1 & \text{if } i\text{-th sample and } j\text{-th sample are equal} \\ 0 & \text{else} \end{cases}$

Lemma 4. *Let p be a discrete probability distribution and C defined as above, then* $\mathbb{E}[C] = \binom{s}{2}\|p\|_2^2$

Proof. We have that

$$\begin{aligned}\mathbb{E}[C] = \sum_{i<j}\mathbb{E}[\sigma_{ij}] &= \sum_{i<j}\Pr[\,ith\ sample = jth\ sample\,] \\ &= \sum_{i<j}\sum_{k=1}^{n}\Pr[\,ith\ sample = k, jth\ sample = k\,] \\ &= \sum_{i<j}\sum_{k=1}^{n}\Pr[\,ith\ sample = k\,]\Pr[\,jth\ sample = k\,] \\ &= \sum_{i<j}\sum_{k=1}^{n}p_k^2 = \sum_{i<j}\|p\|_2^2 = \binom{s}{2}\|p\|_2^2\end{aligned}$$

□

It's clear to see that the distribution that minimizes the expected value will be the one that has the minimal $\|p\|_2$ value.

Fact 5. *For a distribution p we have that: $\|p\|_2 \geq 1/n$ and the minimum is attained by the uniform distribution.*

Proof. Applying the Cauchy-Schwartz inequality, we have that

$$n\|p\|_2^2 \geq (\sum_i p_i)^2 = 1 \Rightarrow \|p\|_2^2 \geq \frac{1}{n}$$

which holds with equality when all the values of p_i are equal, which means that the uniform distribution minimizes the expected value. □

Analyzing each case of output we get

- In the *YES* case we have that : $\mathbb{E}[C] = \frac{\binom{s}{2}}{n}$.
- In the *NO* case we have that : $\|p - \mathcal{U}_n\|_1 \geq \epsilon \Rightarrow \|p - \mathcal{U}_n\|_2^2 \geq \epsilon^2/n$ and using that $\|p\|_2^2 = \|\mathcal{U}_n\|_2^2 + \|p - \mathcal{U}_n\|_2^2$ we get $\|p\|_2^2 \geq \frac{1+\epsilon^2}{n}$ which lead us to $\mathbb{E}[C] \geq \frac{1+\epsilon^2}{n}\binom{s}{2}$

We need to choose the T value in a way that it will separate the *YES* and the *NO* case, we let $T = \frac{1+\epsilon^2/2}{n}\binom{s}{2}$.

Lemma 6. *Let p be a discrete probability distribution, and C as defined above, then* $\mathrm{Var}[C] \leq s^2\|p\|_2^2 + s^3\|p\|_3^3$

Proof. We have that

$$\mathrm{Var}[C] = \mathbb{E}\Big[C^2\Big] - \mathbb{E}[C]^2$$

So we need to bound the $\mathbb{E}[C^2]$ value.

$$\mathbb{E}\Big[C^2\Big] = \mathbb{E}\left[(\sum_{i<j}\sigma_{ij})^2\right] = \mathbb{E}\left[\sum_{i<j}\sigma_{ij}^2 + \sum_{\substack{i_1>j_1\\ i_2>j_2}}\sigma_{i_1j_1}\sigma_{i_2j_2}\right]$$

We have that $\mathbb{E}\Big[\sigma_{ij}^2\Big] = \|p\|_2^2$. For calculating the $\mathbb{E}[\sigma_{i_1j_1}\sigma_{i_2j_2}]$ we distinguish between the following two cases:

1. If the i_1, i_2, j_1, j_2 have all different values, then we are in the independent case, so

$$\mathbb{E}[\sigma_{i_1j_1}\sigma_{i_2j_2}] = \mathbb{E}[\sigma_{i_1j_1}]\,\mathbb{E}[\sigma_{i_2j_2}] = \|p\|_2^4$$

2. If in i_1, i_2, j_1, j_2 two elements have then same value then we are in the dependent case and wlog assume $j_1 = i_2$. Then

$$\begin{aligned}
\mathbb{E}[\sigma_{i_1j_1}\sigma_{j_1j_2}] &= \Pr[\, i_1 th \text{ sample} = j_1 th \text{ sample} = j_2 th \text{ sample}\,] \\
&= \sum_{k=1}^{n} \Pr[\, i_1 th \text{ sample} = k, j_1 th \text{ sample} = k, j_2 th \text{ sample} = k\,] \\
&= \sum_{k=1}^{n} \Pr[\, i_1 th \text{ sample} = k\,]\Pr[\, j_1 th \text{ sample} = k\,]\Pr[\, j_2 th \text{ sample} = k\,] \\
&= \sum_{k=1}^{n} p_k^3 = \|p\|_3^3
\end{aligned}$$

Putting everything together, we have that

$$\mathrm{Var}[C] \leq \binom{s}{2}\|p\|_2^2 + 6\binom{s}{4}\|p\|_2^4 + 6\binom{s}{3}\|p\|_3^3 - \binom{s}{2}^2\|p\|_2^4 \leq s^2\|p\|_2^2 + s^3\|p\|_3^3$$

□

Proposition 7. *We have that:* $\mathrm{Var}[C] \leq 2s^3\|p\|_2^3$

Proof.

$$\mathrm{Var}[C] \leq s^2\|p\|_2^2 + s^3\|p\|_3^3 \leq s^2\|p\|_2^2 + s^3\|p\|_2^3 \leq 2s^3\|p\|_2^3$$

□

Now we are going to bound the sample size that is needed so with probability at least 3/4 we output *YES* when the distribution is uniform. We have:

$$\Pr\Big[\,|C - \mathbb{E}[C]| \geq 2\sqrt{\mathrm{Var}[C]}\,\Big] \leq 1/4$$

Which means that with probability at least 3/4 we have

$$C \leq \mathbb{E}[C] + 2\sqrt{\mathrm{Var}[C]}$$

And we need $C < T$ when $p = \mathcal{U}_n$ (*completeness*), so we have:

$$\begin{aligned} C < T \Rightarrow \mathbb{E}[C] + 2\sqrt{\mathrm{Var}[C]} < T \Rightarrow cs^3\|p\|_2^3 &< \frac{s^4\epsilon^4}{n^2} \\ \Rightarrow c\frac{n^2}{\epsilon^4 n^{3/2}} &< s \\ \Rightarrow c\frac{\sqrt{n}}{\epsilon^4} &< s \end{aligned}$$

We will now deal with the case where $\mathrm{d_{TV}}(p, \mathcal{U}_n) \geq \epsilon$ (*soundness*). We have $\|p\|_2^2 \geq \frac{1+\epsilon^2}{n}$, so we can assume that $\|p\|_2^2 = \frac{1+\epsilon^2+\alpha}{n}$ for $\alpha \geq 0$.
Using Chebyshev as above, we get that with probability at least 3/4 we have

$$C \geq \mathbb{E}[C] - 2\sqrt{\mathrm{Var}[C]}$$

We need $C > T$, so we have

$$\begin{aligned} T < C \Rightarrow \mathbb{E}[C] - 2\sqrt{\mathrm{Var}[C]} < T \Rightarrow c\frac{s^3(1+\epsilon^2+\alpha)^{3/2}}{n^{3/2}} &\leq \frac{s^4(\epsilon^2/2+\alpha)^2}{n^2} \\ \Rightarrow c\frac{\sqrt{n}(1+\epsilon^2+\alpha)^{3/2}}{(\epsilon^2/2+\alpha)^2} &\leq s \end{aligned}$$

Using basic calculus we can show that $\frac{\sqrt{n}(1+\epsilon^2+\alpha)^{3/2}}{(\epsilon^2/2+\alpha)^2}$ is maximized for $\alpha = 0$. So we have:

$$s \geq c\frac{\sqrt{n}}{\epsilon^4}$$

Thus, we need $O\Big(\frac{\sqrt{n}}{\epsilon^4}\Big)$ samples which completes the proof of theorem 3.

5 Further Reading

- In [1] one can find the main theorem and the proof of this lecture.
- In [2] the author proves the information theoretic lower bound $\Omega(\frac{\sqrt{n}}{\epsilon^2})$ and provides a different algorithm that needs $O(\sqrt{n}/\epsilon^2)$ samples but only works when $\epsilon = \Omega(n^{-1/4})$.
- In [3] the authors prove that in fact the collision based testing algorithms are optimal up to constant factors by doing a new more tight analysis.
- Amplifying the success probability of testing uniformity to $1-\delta$ can be done via standard arguments that result in a sample complexity of $O(\sqrt{n}/\epsilon^2 \log(1/\delta))$ (see 2). In [4] the authors show that this dependence of δ is not optimal and prove that with $O\Big(\frac{1}{\epsilon^2}\Big(\sqrt{n\log(1/\delta)} + \log(1/\delta)\Big)\Big)$ samples, we can get confidence $1-\delta$. Considering that $\sqrt{n}$ is usually large this difference in the multiplicative factor may be substantial.

References

1. Goldreich, O. & Ron, D. On Testing Expansion in Bounded-Degree Graphs (2000).
2. Paninski, L. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory* **54,** 4750–4755 (2008).
3. Diakonikolas, I., Gouleakis, T., Peebles, J. & Price, E. Collision-based testers are optimal for uniformity and closeness. *arXiv preprint arXiv:1611.03579* (2016).
4. Diakonikolas, I., Gouleakis, T., Peebles, J. & Price, E. *Sample-optimal identity testing with high probability* in *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)* (2018).