CS 514: Advanced Algorithms II – Sublinear Algorithms **Rutgers: Fall 2021**

Lecture 5

October 05, 2021

Instructor: Sepehr Assadi

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Topics of this Lecture

1 Distribution Testing

The field of distribution testing asks questions about distributions over $[n]$, or in particular functions of the form $\mu : [n] \to \mathbb{R}_{\geq 0}$, with $\sum_{i=1}^{n} \mu(i) = 1$. Some examples of questions that might be asked about the distribution are whether it is uniform, what its mean is, and what its entropy is. Note that for the decision version of these questions, we can again fall back to the paradigm of property testing when needed by asking question of the type: is this distribution uniform or "far" from uniform? (for some reasonable distance measure between distributions – more on this later in the section.)

Distribution testing tend to deviate from previous models we saw for sublinear time algorithms. For a sublinear time algorithm working on distribution μ, we would typically assume query access of the form what is $\mu(i)$ for some given $i \in [n]$. Instead, in distribution testing, the only type of "query" that an algorithm can make is to *sample* an element from the distribution. In this sense, the computational model of distribution testing is much less flexible than that of sublinear time algorithms. A sublinear time algorithm typically has the freedom to choose the type and parameters of the queries it makes. However, the only freedom an algorithm has in the distribution testing model is to choose the *number* of samples it wants to make from the distribution and how to use them to infer the final answer.

1.1 Uniformity Testing

To show case the notion of distribution testing, we are going to focus on one of the first problems studied in this area, namely, the *uniformity testing* problem.

The particular question we will address is whether a distribution μ is uniform or not. We will use U_n to denote the uniform distribution over $[n]$, that is $U_n(i) = \frac{1}{n}$ for all $i \in [n]$. Distribution testing is similar to property testing in the sense that it does not lend itself well to asking extremely precise questions. So instead we will attempt to find an ε-tester for this question; in particular we wish to describe an algorithm that outputs *Yes* if the distribution it is sampling from is U_n, and *No* if the distribution is "ε-far" from U_n. Both of these conditions are of course only required with high probability.

In the previous problem statement, we are still missing a definition of what it means for one distribution to be"ε-far" from another. While there are a variety of possible choices here, we will use the notion of **Total Variation Distance** between two distributions which will be defined formally in the next subsection. Under this definition, we can define our problem formally as follows:

Problem 1 (**Uniformity Testing**)**.** Given sample access to a distribution μ on domain $[n]$:

- if $\mu = U_n$, output *Yes*;
- if $\Delta_{\text{tvd}}(\mu, U_n) \geq \varepsilon$, output *No*, where $\Delta_{\text{tvd}}(\mu, U_n)$ denotes the total variation distance (see Section 1.2).

The answer can be arbitrary in other case.

We start with a detour on total variation distance in the next subsection, and then present an algorithm for uniformity testing in Section 2.

1.2 Detour: Total Variation Distance

The total variation distance between any two distributions μ and ν over the same (discrete) domain $[n]$ is defined as:

$$\Delta_{\text{tvd}}(\mu, \nu) := \frac{1}{2} \cdot ||\mu - \nu||_1 = \frac{1}{2} \cdot \sum_{i=1}^{n} |\mu(i) - \nu(i)|. \tag{1}$$

(Here, $||\cdot||_1$ is the (standard) ℓ_1 norm of the given vector – recall that we can treat a distribution μ as simply a vector in $\mathbb{R}^n$ with non-negative entries and $||\mu||_1 = 1$.)

Total variation distance is a widely used distance metric between two distributions. In the following, we are going to establish some illustrating properties of this distance[1].

The first property is that total variation distance between two distributions is equal to the maximum gap possible between probability assigned to any event under these two distributions. Formally,

Proposition 1. *The following is true for any pairs of distributions μ and ν on $[n]$:*

$$\Delta_{tvd}(\mu, \nu) = \max_{\Omega \subseteq [n]} \{\mu(\Omega) - \nu(\Omega)\}.$$

Proof. Let $\Omega \subseteq [n]$ be $\arg\max \{\mu(\Omega) - \nu(\Omega)\}$. Note that for any $i \in \Omega$, $\mu(i) \geq \nu(i)$ as otherwise the event $\Omega \setminus \{i\}$ will have a higher gap in probabilities between μ and ν, contradicting the choice of Ω. Moreover,

$$\sum_{i \in \Omega} \mu(i) - \nu(i) = \sum_{i \notin \Omega} \nu(i) - \mu(i),$$

as $\sum_i \mu(i) = \sum_i \nu(i) = 1$. This way, by Eq (1), we have,

$$\Delta_{\text{tvd}}(\mu, \nu) = \frac{1}{2} \cdot \left(\sum_{i \in \Omega} \mu(i) - \nu(i) + \sum_{i \notin \Omega} \nu(i) - \mu(i) \right) = \sum_{i \in \Omega} \mu(i) - \nu(i) = \mu(\Omega) - \nu(\Omega),$$

as desired. □

Proposition 1 is particularly useful when one wants to bound the probability of an event over a "complicated" probability distribution which is *close* in total variation distance to some "nicer" probability distribution. We can bound the probability of the event under the nicer distribution and get a bound for the original distribution as well which can only be looser compared to the actual bound by the distance between the two distributions. The following proposition extend this reasoning to random variables as well.

[1]These properties are *not* needed for our uniformity testing algorithm and thus this part can be skipped by a reader solely interested in distribution testing.

Proposition 2. *The following is true for any pairs of distributions μ and ν on $[n]$, and any random variable X with domain $[n]$:*

$$|\mathop{\mathbb{E}}_{\mu}[X] - \mathop{\mathbb{E}}_{\nu}[X]| \leq 2 \cdot \Delta_{tvd}(\mu, \nu) \cdot \max |X|.$$

Proof. Recall that a random variable X is simply a function from $[n]$ to range $[-R, R]$ for $R := \max |X|$. We thus have,

$$|\mathop{\mathbb{E}}_{\mu}[X] - \mathop{\mathbb{E}}_{\nu}[X]| = |\sum_{i=1}^{n}(\mu(i) - \nu(i)) \cdot X(i)| \leq \max |X| \cdot \sum_{i=1}^{n} |\mu(i) - \nu(i)| = |X| \cdot 2\,\Delta_{\text{tvd}}(\mu, \nu),$$

as desired. $\square$

Finally, we show that total variation distance is also a good measure for distinguishing two probability distributions given a single sample from them.

Proposition 3. *Let μ and ν be two probability distributions over $[n]$. Suppose one is given a sample s chosen with probability $1/2$ from μ and with probability $1/2$ from ν. Then, the best probability of success in deciding whether s was sampled from μ or ν is exactly $1/2 + 1/2 \cdot \Delta_{tvd}(\mu, \nu)$.*

Proof. Firstly, it is easy to see that the best estimator for the origin of s is the *maximum likelihood estimator* (MLE) that returns μ if $\mu(s) > \nu(s)$, ν if $\nu(s) > \mu(s)$, and either of μ or ν arbitrarily whenever $\mu(s) = \nu(s)$ (for simplicity, we assume in this case also, the algorithm always return ν). We omit the proof.

Let $\Omega \subseteq [n]$ be the set of elements with $\mu(s) > \nu(s)$. Whenever we get $s \in \Omega$, the estimator returns μ and otherwise it returns ν. Thus, the error happens when $s \in \Omega$ is sampled from ν or $s \notin \Omega$ is sampled from μ. The probabilities of these events can be calculated as:

$$\text{for } s \in \Omega: \qquad \Pr(s \text{ chosen from } \nu \mid s) = \Pr(s \mid s \text{ chosen from } \nu) \cdot \frac{\Pr(s \text{ chosen from } \nu)}{\Pr(s)} = \nu(s) \cdot \frac{1}{2 \cdot \Pr(s)}.$$

$$\text{for } s \notin \Omega: \qquad \Pr(s \text{ chosen from } \mu \mid s) = \Pr(s \mid s \text{ chosen from } \mu) \cdot \frac{\Pr(s \text{ chosen from } \mu)}{\Pr(s)} = \mu(s) \cdot \frac{1}{2 \cdot \Pr(s)},$$

where in both case we use the Bayes' rule. As a result, we have

$$\begin{aligned}
\Pr(\text{error}) &= \sum_{s=1}^{n} \Pr(\text{error} \mid s) \cdot \Pr(s) = \frac{1}{2} \cdot \left(\sum_{s \in \Omega} \nu(s) + \sum_{s \in [n] \setminus \Omega} \mu(s) \right) \\
&= \frac{1}{2} \cdot \sum_{s=1}^{n} \min\{\mu(s), \nu(s)\} = \frac{1}{2} \cdot \left(\sum_{s=1}^{n} \frac{\mu(s) + \nu(s) - |\mu(s) - \nu(s)|}{2} \right) \qquad (\text{as } \min\{a, b\} = \tfrac{a+b-|a-b|}{2}) \\
&= \frac{1}{2} \cdot (1 - \Delta_{\text{tvd}}(\mu, \nu)). \qquad (\text{as } \textstyle\sum_{s=1}^{n} \mu(s) = \sum_{s=1}^{n} \nu(s) = 1 \text{ and by Eq (1)})
\end{aligned}$$

This concludes the proof. $\square$

2 A Uniformity Testing Algorithm

We now design a uniformity testing algorithm for Problem 1 and prove the following theorem.

Theorem 4. *For any $\varepsilon \in (0, 1)$, there is a distribution testing algorithm for testing uniformity (under total variation distance) that outputs the correct answer with probability at least $1 - \delta$ using $O(\frac{1}{\varepsilon^4} \cdot \sqrt{n} \cdot \log(1/\delta))$ samples.*

This theorem was first proved by Goldreich and Ron [6] (implicitly and in a different context), and then was reproved in the above form by Batu *et.al.* in [3] and [2]. These bounds were later improved to $O(\sqrt{n}/\varepsilon^2)$ samples by Paninski [7] (for not-too-small ε) and then by Chan *et.al.* [4], Diakonikolas *et.al.* [5], and Acharya *et.al.* [1] (for all range of $\varepsilon > 0$). The $\Omega(\sqrt{n}/\varepsilon^2)$ number of samples was also shown to be necessary by Paninski [7].

The intuition behind the tester is the following. Consider picking two samples from distribution μ. If $\mu = U_n$, then the probability that these samples are equal, i.e., we have a *collision*, is $\frac{1}{n}$. It is also easy to see that in any other distribution, this probability can only be larger (we will prove this formally below). In fact, we are going to show that if $\Delta_{\text{tvd}}(\mu, U_n) \geq \varepsilon$, then this probability is considerably larger than $\frac{1}{n}$ by a factor of $(1 + O(\varepsilon^2))$. We can then design an algorithm to estimate the collision probability and use it to distinguish between the two cases for distribution μ.

While the total variation distance is the metric we use in the problem statement, we will also repeatedly make use of the ℓ_2 or **Euclidean norm**. For some distribution μ, we define the square of the ℓ_2 norm to be

$$||\mu||_2^2 = \sum_{i=1}^{n} \mu(i)^2.$$

This is the same as calculating the ℓ_2 norm of the vector in $\mathbb{R}^n$ that represents this distribution. Note that this definition has an intuitive interpretation: It is precisely the probability that two independent samples from μ have the same value, i.e.,

$$\Pr_{x,y\sim\mu}(x = y) = ||\mu||_2^2. \tag{2}$$

We start by showing that any distribution that is far from uniform will also have a sufficiently larger collision probability (or ℓ_2-norm by Eq (2)).

Lemma 5. *If μ is a distribution with $\Delta_{tvd}(\mu, U_n) \geq \varepsilon$, then $||\mu||_2^2 \geq (1 + 4\varepsilon^2) \cdot \frac{1}{n}$.*

Proof. We will begin by considering $||\mu - U_n||_2^2$. The subtraction here is performed on the distributions interpreted as vectors. We have

$$\begin{aligned}
||\mu - U_n||_2^2 &= \sum_{i=1}^{n} (\mu(i) - \frac{1}{n})^2 \\
&= \sum_{i=1}^{n} \mu(i)^2 - 2\sum_{i=1}^{n} \frac{\mu(i)}{n} + \sum_{i=1}^{n} \frac{1}{n^2} \\
&= ||\mu||_2^2 - \frac{2}{n} + \frac{1}{n} \\
&= ||\mu||_2^2 - \frac{1}{n}.
\end{aligned}$$

The knowledge about $\Delta_{\text{tvd}}(\mu, U_n) = (1/2) \cdot ||\mu - U_n||_1$ allows us to argue about $||\mu - U_n||_2^2$ as well. In particular, we will use the following standard connection between ℓ_2 and ℓ_1 norms (the proof is simply by Cauchy-Schwartz inequality).

Fact 6. *For any vector $x \in \mathbb{R}^n$, $||x||_1 \leq \sqrt{n} \cdot ||x||_2$.*

Based on this, we have,

$$\begin{aligned}
||\mu||_2^2 = ||\mu - U_n||_2^2 + \frac{1}{n} &\geq \frac{1}{n} \cdot ||\mu - U_n||_1^2 + \frac{1}{n} && \text{(by Fact 6)} \\
&\geq \frac{\varepsilon^2}{n} + \frac{1}{n}, && \text{(by the bound } ||\mu - U_n||_1 = 2 \cdot \Delta_{\text{tvd}}(\mu, \nu) \geq 2\varepsilon)
\end{aligned}$$

proving the result. ∎

Estimating $||\mu||_2^2$ as a way to test uniformity. By Lemma 5, when $\mu = U_n$ in the *Yes* case, we have $||\mu||_2^2 = \frac{1}{n}$, while when $\Delta_{\text{tvd}}(\mu, U_n) \geq \varepsilon$ in the *No* case, we have $||\mu||_2^2 \geq (1+4\varepsilon^2)/n$. As such, if we can get a $(1 \pm \varepsilon^2)$ approximation to value of μ we can distinguish between these two cases. In particular,

- In the *Yes* case, even if we completely overestimate the value of $||\mu||_2^2$ within the approximation range, the answer would be at most $(1+\varepsilon^2) \cdot \frac{1}{n}$.
- On the other hand, in the *No* case, even if we completely underestimate the value of $||\mu||_2^2$ within the approximation range, the answer would be at least $(1-\varepsilon^2) \cdot (1+4\varepsilon^2) \cdot \frac{1}{n}$.
- Considering,
$$(1-\varepsilon^2) \cdot (1+4\varepsilon^2) = (1+3\varepsilon^2 - 4\varepsilon^4) \geq (1+2\varepsilon^2) \qquad \text{(for } \varepsilon \in (0, 1/2]\text{)}$$
we have that the largest overestimation in the *Yes* case is still smaller than the smallest underestimation in the *No* case (whenever $\varepsilon \geq 1/2$, we can simply replace it with $1/2$ and run the algorithm for this smaller ε to get the same asymptotic sample complexity bound as desired in Theorem 4).
- Finally, we can distinguish between the two cases as follows: if the estimate of $||\mu||_2^2 \leq (1+\varepsilon^2) \cdot \frac{1}{n}$, we output the distribution as uniform, namely, *Yes* case, and otherwise output the answer is *No*.

Hence, our goal now is to design a $(1 \pm \varepsilon^2)$-approximation algorithm for estimating the value of $||\mu||_2^2$. By the above discussion, this suffices to obtain our tester for this distribution testing algorithm.

2.1 An Algorithm for Estimating $||\mu||_2^2$

We propose the following algorithm for estimating $||\mu||_2^2$ to within a factor of $(1 \pm \gamma)$ for $\gamma = \varepsilon^2$.

An algorithm for estimating $||\mu||_2^2$:

1. Sample $x_1, \ldots, x_k$ from μ for $k := \frac{24}{\gamma^2} \cdot \sqrt{n}$.
2. For $i < j$, let $Y_{ij} \in \{0, 1\}$ be the indicator random variable that is 1 if $x_i = x_j$ and 0 otherwise.
3. Return $X := \frac{\sum_{i<j} Y_{ij}}{\binom{k}{2}}$.

The sample complexity of this algorithm is $O(\sqrt{n}/\varepsilon^4)$ by the choice of k and γ. It can also be shown that this algorithm can be implemented in $O(\sqrt{n}/\varepsilon^4)$ by sorting the samples using radix-sort and then counting the collisions (as the main focus in distribution testing is the sample complexity, we do not get into the details of the runtime further). We now prove the correctness of the algorithm in the following lemma.

Lemma 7. $\Pr\left(|X - ||\mu||_2^2| \geq \gamma \cdot ||\mu||_2^2\right) \leq 1/3$.

Similar to some of our other sublinear algorithms, the proof of this lemma is by first showing that X is in expectation equal to the desired number, namely, $||\mu||_2^2$, and then bound the variance of X. This allows us to apply Chebyshev's inequality to obtain that X with sufficiently large probability is close to its expected value and thus $||\mu||_2^2$, and prove the lemma.

Claim 8 (Expectation bound)**.** $\mathbb{E}[X] = ||\mu||_2^2$.

Proof. We have,

$$\mathbb{E}[X] = \frac{\mathbb{E}[\sum_{i<j} Y_{ij}]}{\binom{k}{2}}$$

$$
\begin{aligned}
&= \frac{\sum_{i<j} \mathbb{E}[Y_{i,j}]}{\binom{k}{2}} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i<j} \Pr_{x_i,x_j\sim\mu}(x_i = x_j)}{\binom{k}{2}} && \text{(by definition of } Y_{ij}\text{)} \\
&= \frac{\sum_{i<j} ||\mu||_2^2}{\binom{k}{2}} && \text{(by Eq (2))} \\
&= ||\mu||_2^2, && \text{(as there are } \tbinom{k}{2} \text{ choices of } i<j\in[k]\text{)}
\end{aligned}
$$

concluding the proof. □

Claim 9 (Variance bound)**.** $\mathrm{Var}[X] \leq 4 \cdot ||\mu||_2^2/k^2 + 4 \cdot ||\mu||_3^3/k$.

Proof. As $\mathrm{Var}[\alpha \cdot A] = \alpha^2 \cdot \mathrm{Var}[A]$ for any constant α and random variable A, we have,

$$
\mathrm{Var}[X] = \mathrm{Var}\left[\frac{\sum_{i<j} Y_{ij}}{\binom{k}{2}}\right] = \frac{1}{\binom{k}{2}^2} \cdot \mathrm{Var}\left[\sum_{i<j} Y_{ij}\right]. \tag{3}
$$

Thus, we only need to focus on variance of $\sum_{i<j} Y_{ij}$. We are going to use the following fact in the proof. Recall that for any two random variables A, B, *covariance* of A, B is $\mathrm{Cov}[A,B] := \mathbb{E}[AB] - \mathbb{E}[A]\,\mathbb{E}[B]$.

Fact 10. *For random variables* $A_1, \ldots, A_n$*,* $\mathrm{Var}[\sum_{i=1}^n A_i] = \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}[A_i, A_j]$.

By Fact 10, we have,

$$
\begin{aligned}
\mathrm{Var}\left[\sum_{i<j} Y_{ij}\right] &= \sum_{i<j, s<t} \mathrm{Cov}[Y_{ij}, Y_{st}] \\
&= \sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=2}} \mathrm{Cov}[Y_{ij}, Y_{st}] + \sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=3}} \mathrm{Cov}[Y_{ij}, Y_{st}] + \sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=4}} \mathrm{Cov}[Y_{ij}, Y_{st}]. \qquad (4)
\end{aligned}
$$

We bound each of the terms in Eq (4) in the following.

1. For the first term,

$$
\begin{aligned}
\sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=2}} \mathrm{Cov}[Y_{ij}, Y_{st}] &= \sum_{i<j} \mathrm{Cov}[Y_{ij}, Y_{ij}] && \text{(because } i=s \text{ and } j=t \text{ when } |\{i,j,s,t\}|=2\text{)} \\
&= \sum_{i<j} \mathrm{Var}[Y_{ij}] \\
& && \text{(because for any random variable } A,\ \mathrm{Cov}[A,A] = \mathbb{E}[A^2] - \mathbb{E}[A]^2 = \mathrm{Var}[A]\text{)} \\
&\leq \sum_{i<j} \mathbb{E}[Y_{ij}^2] = \sum_{i<j} \mathbb{E}[Y_{ij}] \\
& && \text{(as } Y_{ij} \text{ is an indicator random variable and thus } Y_{ij}^2 = Y_{ij}\text{)} \\
&= \binom{k}{2} \cdot ||\mu||_2^2. && \text{(by Claim 8)}
\end{aligned}
$$

2. For the second term,

$$
\sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=3}} \mathrm{Cov}[Y_{ij}, Y_{st}] \leq \sum_{\substack{i<j,s<t \\ |\{i,j,s,t\}|=3}} \mathbb{E}[Y_{ij} Y_{st}]
$$

$$= \sum_{\substack{i<j, s<t \\ |\{i,j,s,t\}|=3}} \Pr_{x,y,z \sim \mu} (x = y = z)$$

(when $|\{i, j, s, t\}| = 3$, $Y_{ij} Y_{st} = 1$ if the three corresponding samples are all equal and is zero otherwise)

$$= \sum_{\substack{i<j, s<t \\ |\{i,j,s,t\}|=3}} \sum_{x=1}^{n} \mu(x)^3$$

$$= \sum_{\substack{i<j, s<t \\ |\{i,j,s,t\}|=3}} ||\mu||_3^3$$

$$= 6 \cdot \binom{k}{3} \cdot ||\mu||_3^3,$$

where the last equality follows from a simple combinatorics exercise in counting number of ways of choosing $i < j, s < t$ from $[k]$ so that $|\{i, j, s, t\}| = 3$ (proof: we first pick distinct x, y, z from $[k]$ with $\binom{k}{3}$ choices; for each such choice, we pick one of x, y, z to be repeated twice among i, j, s, t with 3 choices, and finally, we can assign the remaining indices i, j, s, t in two ways, resulting in $6 \cdot \binom{k}{3}$ choices).

3. For the third term,

$$\sum_{\substack{i<j, s<t \\ |\{i,j,s,t\}|=4}} \text{Cov}[Y_{ij}, Y_{st}] = 0.$$

(since $i \neq s$ and $j \neq t$ and thus $Y_{ij} \perp Y_{st}$; also $\text{Cov}[A, B] = 0$ for *independent* A, B)

Putting all these together in Eq (4), we get that,

$$\text{Var}\left[\sum_{i<j} Y_{ij}\right] \leq \binom{k}{2} \cdot ||\mu||_2^2 + 6 \cdot \binom{k}{3} \cdot ||\mu||_3^3 + 0.$$

Finally, we plug in this bound in Eq (3) and obtain that,

$$\text{Var}[X] \leq \frac{||\mu||_2^2}{\binom{k}{2}} + \frac{k \cdot (k-1) \cdot (k-2) \cdot ||\mu||_3^3}{k^2 \cdot (k-1)^2/4}$$

$$\leq \frac{4 \cdot ||\mu||_2^2}{k^2} + \frac{4 \cdot ||\mu||_3^3}{k}. \qquad \text{(as } \tbinom{x}{y} \geq (x/y)^y\text{)}$$

This concludes the proof. □

Remark. Prior to the proof of Lemma 7, we have only worked with bounding variance of sum of random variables in the case they were *independent*. Under this assumption, Fact 10 simplifies to variance of the sum is equal to sum of individual variances. In the proof of Lemma 7 however, we no longer had the independence assumption but similar to what we saw in the proof, in many scenarios bounding variance of the sum is still manageable by considering covariance terms explicitly.

We are now ready to finalize the proof of Lemma 7. First, let us recall a simple fact about norms.

Fact 11. *For any vector $x \in \mathbb{R}^n$ and integers $p \leq q$, $||x||_q \leq ||x||_p$.*

Finally, we present the proof of Lemma 7.

Proof of Lemma 7. By Chebyshev's inequality,

$$\begin{aligned}
\Pr\left(|X - \mathbb{E}[X]| \geq \gamma \cdot ||\mu||_2^2\right) &\leq \frac{\mathrm{Var}[X]}{\gamma^2 \cdot ||\mu||_2^4} \\
&\leq \frac{4 \cdot ||\mu||_2^2}{k^2 \cdot \gamma^2 \cdot ||\mu||_2^4} + \frac{4 \cdot ||\mu||_3^3}{k \cdot \gamma^2 \cdot ||\mu||_2^4} && \text{(by Claim 9)} \\
&\leq \frac{4}{k^2 \cdot \gamma^2 \cdot ||\mu||_2^2} + \frac{4}{k \cdot \gamma^2 \cdot ||\mu||_2} && \text{(by Fact 11, } ||\mu||_3 \leq ||\mu||_2) \\
&\leq \frac{4n}{k^2 \cdot \gamma^2} + \frac{4\sqrt{n}}{k \cdot \gamma^2} && \text{(as } ||\mu||_2^2 \geq \tfrac{1}{n} \text{ for any distribution } \mu) \\
&\leq \frac{4\varepsilon^4 \cdot n}{24^2 \cdot (\varepsilon^2)^2 \cdot n} + \frac{4\varepsilon^4 \cdot \sqrt{n}}{24 \cdot (\varepsilon^2)^2 \cdot \sqrt{n}} && \text{(by the choices of } k \text{ and } \gamma) \\
&= \frac{1}{144} + \frac{1}{6} < \frac{1}{3}.
\end{aligned}$$

As by Claim 8, $\mathbb{E}[X] = ||\mu||_2^2$, we get the final result. □

By Lemma 7, we obtain an algorithm for uniformity testing with probability of success at least 2/3. We can now use our standard trick of running the algorithm $O(\log(1/\delta))$ times in parallel and return the majority answer for uniformity testing (or median answer for estimating $||\mu||_2^2$), which gives us an algorithm with probability of success $1 - \delta$. We can thus conclude the proof of Theorem 4.

References

[1] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3591–3599, 2015. 4

[2] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 442–451, 2001. 4

[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269, 2000. 4

[4] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1193–1203, 2014. 4

[5] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1183–1202, 2015. 4

[6] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000. 4

[7] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Information Theory*, 54(10):4750–4755, 2008. 4