Electronic Colloquium on Computational Complexity, Report No. 178 (2016)

# Collision-based Testers are Optimal for Uniformity and Closeness

Ilias Diakonikolas* Themis Gouleakis† John Peebles‡ Eric Price

USC MIT MIT UT Austin

November 10, 2016

**Abstract**

We study the fundamental problems of (i) uniformity testing of a discrete distribution, and (ii) closeness testing between two discrete distributions with bounded $\ell_2$-norm. These problems have been extensively studied in distribution testing and sample-optimal estimators are known for them [Pan08, CDVV14, VV14, DKN15b].

In this work, we show that the original collision-based testers proposed for these problems [GR00, BFR$^+$00] are sample-optimal, up to constant factors. Previous analyses showed sample complexity upper bounds for these testers that are optimal as a function of the domain size $n$, but suboptimal by polynomial factors in the error parameter $\epsilon$. Our main contribution is a new tight analysis establishing that these collision-based testers are information-theoretically optimal, up to constant factors, both in the dependence on $n$ and in the dependence on $\epsilon$.

## 1 Introduction

**1.1 Background and Our Results** The generic inference problem in *distribution property testing* [BFR$^+$00, BFR$^+$13] (also see, e.g., [Rub12, Can15, Gol16b]) is the following: given sample access to one or more unknown distributions, determine whether they satisfy some global property or are "far" from satisfying the property. During the past couple of decades, distribution testing – whose roots lie in statistical hypothesis testing [NP33, LR05] – has developed into a mature field. One of the most fundamental tasks in this field is deciding whether an unknown discrete distribution is approximately uniform on its domain, known as the problem of *uniformity testing*. Formally, we want to design an algorithm that, given independent samples from a discrete distribution $p$ over $[n]$ and a parameter $\epsilon > 0$, distinguishes (with high probability) the case that $p$ is uniform from the case that $p$ is $\epsilon$-far from uniform, i.e., the total variation distance between $p$ and the uniform distribution over $[n]$ is at least $\epsilon$.

Uniformity testing was the very first problem considered in this line of work: Goldreich and Ron [GR00], motivated by the question of testing the expansion of graphs, proposed a simple and natural uniformity tester that relies on the *collision probability* of the unknown distribution. The collision probability of a discrete distribution $p$ is the probability that two samples drawn according to $p$ are equal. The key intuition here is that the uniform distribution has the minimum collision probability among all distributions on the same domain, and that any distribution that is $\epsilon$-far

*Part of this research was performed when the author was at the University of Edinburgh, and while visiting MIT. Supported in part by a Marie Curie Career Integration grant and an EPSRC grant.

†This material is based upon work supported by the NSF under Grant No. 1420692.

‡This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. 1122374, and by the NSF under Grant No. 1065125.

ISSN 1433-8092

from uniform has noticeably larger collision probability. Formalizing this intuition, Goldreich and Ron [GR00] showed that the collision-based uniformity tester succeeds after drawing $O(n^{1/2}/\epsilon^4)$ samples from the unknown distribution. An information-theoretic lower bound of $\Omega(n^{1/2})$ on the number of samples required by any uniformity tester follows from a simple birthday-paradox argument [GR00, BFF$^+$01], even for constant values of the parameter $\epsilon$. In subsequent work, Paninski [Pan08] showed an information-theoretic lower bound of $\Omega(n^{1/2}/\epsilon^2)$, and also provided a matching upper bound of $O(n^{1/2}/\epsilon^2)$ that holds under the assumption that $\epsilon = \Omega(n^{-1/4})$[1]. This lower bound assumption on $\epsilon$ is not inherent: As shown in a number of recent works [VV14, DKN15b] (see also [ADJ$^+$12, CDVV14]), a variant of Pearson's $\chi^2$-tester can test uniformity with $O(n^{1/2}/\epsilon^2)$ samples for all values of $n, \epsilon > 0$. The "chi-squared type" testers of [CDVV14, VV14] are simple, but are also arguably slightly less natural than the original collision-based uniformity tester [GR00].

Perhaps surprisingly, prior to this work, the sample complexity of the collision uniformity tester was not fully understood. In particular, it was not known whether the sample upper bound of $O(n^{1/2}/\epsilon^4)$ – established in [GR00] – is tight for this tester, or there exists an improved analysis that can give a better upper bound. As our first main contribution (Theorem 1), we provide a new analysis of the collision uniformity tester establishing a tight $O(n^{1/2}/\epsilon^2)$ upper bound on its sample complexity. That is, we show that the originally proposed uniformity tester is in fact sample-optimal, up to constant factors.

A related testing problem of central importance in the field is the following: Given samples from two unknown distributions $p, q$ over $[n]$ with the promise that $\max\{\|p\|_2^2, \|q\|_2^2\} \leq b$, distinguish between the cases that $\|p-q\|_2 \leq \epsilon/2$ and $\|p-q\|_2 \geq \epsilon$. That is, we want to test the closeness between two unknown distributions with small $\ell_2$-norm. (We remark here that the assumption that both $p$ and $q$ have small $\ell_2$-norm is critical in this context.) The seminal work of Batu *et al.* [BFR$^+$00] gave a collision-based tester for this problem that uses $O(b^2/\epsilon^4 + b^{1/2}/\epsilon^2)$ samples. Subsequent work by Chan, Diakonikolas, Valiant, and Valiant [CDVV14] gave a different "chi-squared type" tester that uses $O(b^{1/2}/\epsilon^2)$; this sample bound was shown [CDVV14, VV14] to be optimal, up to constant factors.

Similarly to the case of uniformity testing, prior to this work, it was not known whether the analysis of the collision-based closeness tester in [BFR$^+$00] is tight. As our second contribution, we show (Theorem 8) that (essentially) the collision-based tester of [BFR$^+$00] succeeds with $O(b^{1/2}/\epsilon^2)$ samples, i.e., it is sample-optimal, up to constants, for the corresponding problem.

**Remark.** Uniformity testing has been a useful algorithmic primitive for several other distribution testing problems as well [BFF$^+$01, DDS$^+$13, DKN15b, DKN15a, CDGR16, Gol16a]. Notably, Goldreich [Gol16a] recently showed that the more general problem of testing the identity of any explicitly given distribution can be reduced to uniformity testing with only a constant factor loss in sample complexity.

The problem of $\ell_2$ closeness testing for distributions with small $\ell_2$ norm has been identified as an important algorithmic primitive since the original work of Batu *et al.* [BFR$^+$00] who exploited it to obtain the first $\ell_1$ closeness tester. Recently, Diakonikolas and Kane [DK16] gave a collection of reductions from various distribution testing problems to the above $\ell_2$ closeness testing problem. The approach of [DK16] shows that one can obtain sample-optimal testers for a range of different properties of distributions by applying an optimal tester for the above problem as a black-box.

[1] The uniformity tester of [Pan08] relies on the number of *unique elements*, i.e., the elements that appear in the sample set exactly once. Such a tester is only meaningful in the regime that the total number of samples is smaller than the domain size.

**1.2 Overview of Analysis** We now provide a brief summary of previous analyses and a comparison with our work. The canonical way to construct and analyze distribution property testers roughly works as follows: Given $m$ independent samples $s_1, \ldots, s_m$ from our distribution(s), we consider an appropriate random variable (statistic) $F(s_1, \ldots, s_m)$. If $F(s_1, \ldots, s_m)$ exceeds an appropriately defined threshold $T$, our tester rejects; otherwise, it accepts. The canonical analysis proceeds by bounding the expectation and variance of $F$ for the case that the distribution(s) satisfy the property (completeness), and the case they are $\epsilon$-far from satisfying the property (soundness), followed by an application of Chebyshev's inequality.

The main difficulty is choosing the statistic $F$ appropriately so that the expectations for the completeness and soundness cases are sufficiently separated after a small number of samples, and at the same time the variance of the statistic is not "too large". Typically, the challenging step in the analysis is bounding from above the variance of $F$ in the soundness case. Our analysis follows this standard framework. Roughly speaking, for both problems we consider, we provide a tighter analysis of the variance of the corresponding estimators, that in turn leads to the optimal sample complexity upper bound.

More specifically, for the case of uniformity testing, the argument of [GR00] proceeds by showing that the collision tester yields a $(1+\gamma)$-multiplicative approximation of the $\ell_2^2$-norm of the unknown distribution with $O(n^{1/2}/\gamma^2)$ samples. Setting $\gamma = \epsilon^2$ gives a uniformity testing under the $\ell_1$ distance that uses $O(n^{1/2}/\epsilon^4)$ samples. We note that the quadratic dependence on $1/\gamma$ in the multiplicative approximation of the $\ell_2$ norm is tight in general. (For an easy example, consider the case that our distribution is either uniform over two elements, or assigns probability mass $1/2-\gamma, 1/2+\gamma$ to the elements.) Roughly speaking, we show that we can do better when the $\ell_2$ norm of the distribution in question is small. More specifically, the collision uniformity tester can distinguish between the case that $\|p\|_2^2 \leq (1+\gamma/2)/n$ and $\|p\|_2^2 \geq (1+\gamma)/n$ with $O(n^{1/2}/\gamma)$ samples. This immediately yields the desired $\ell_1$ guarantee.

For the closeness testing problem (under our bounded $\ell_2$ norm assumption), Batu *et al.* [BFR[+]00] construct a statistic whose expectation is proportional to the square of the $\ell_2$ distance between the two distributions $p$ and $q$. This statistic has three terms whose expectations are proportional to $\|p\|_2^2$, $\|q\|_2^2$, and $2p \cdot q$ respectively. Specifically, the first term is obtained by considering the number of self-collisions of a set of samples from $p$. Similarly, the second term is proportional to the number of self-collisions of a set of samples from $q$. The third term is obtained by considering the number of "cross-collisions" between some samples from $p$ and $q$. In order to simplify the analysis, [BFR[+]00] uses a separate set of fresh samples for the cross-collisions term. This set is independent of the set of samples used for the two self-collisions terms. While this choice makes the analysis cleaner, it ends up increasing the variance of the estimator too much leading to a sub-optimal sample upper bound. We show that by reusing samples to calculate the number of cross-collisions, one achieves sufficiently good variance to get optimal sample complexity. This comes at the cost of a more complicated analysis involving a very careful calculation of the variance.

**1.3 Notation** We write $[n]$ to denote the set $\{1, \ldots, n\}$. We consider discrete distributions over $[n]$, which are functions $p : [n] \to [0,1]$ such that $\sum_{i=1}^n p_i = 1$. We use the notation $p_i$ to denote the probability of element $i$ in distribution $p$. We will denote by $U_n$ the uniform distribution over $[n]$.

For $r \geq 1$, the $\ell_r$–norm of a distribution is identified with the $\ell_r$–norm of the corresponding vector, i.e., $\|p\|_r = (\sum_{i=1}^n |p_i|^r)^{1/r}$. The $\ell_1$ (resp. $\ell_2$) distance between distributions $p$ and $q$ is defined as the the $\ell_1$ (resp. $\ell_2$) norm of the vector of their difference, i.e., $\|p-q\|_1 = \sum_{i=1}^n |p_i - q_i|$ and $\|p-q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$.

## 2 Testing Uniformity via Collisions

In this section, we show that the natural collision uniformity tester proposed in [GR00] is sample-optimal up to constant factors. More specifically, we are given $m$ samples from a probability distribution $p$ over $[n]$, and we wish to distinguish (with high constant probability) between the cases that $p$ is uniform versus $\epsilon$-far from uniform in $\ell_1$-distance. The main result of this section is that the collision-based uniformity tester succeeds in this task with $m = O(n^{1/2}/\epsilon^2)$ samples.

In fact, we prove the following stronger $\ell_2$-guarantee for the collisions tester: With $m = O(n^{1/2}/\epsilon^2)$ samples, it distinguishes between the cases that $\|p - U_n\|_2^2 \leq \epsilon^2/(2n)$ (completeness) versus $\|p - U_n\|_2^2 \geq \epsilon^2/n$ (soundness). The desired $\ell_1$ guarantee follows from this $\ell_2$ guarantee by an application of the Cauchy-Schwarz inequality in the soundness case.

Formally, we analyze the following tester:

**Algorithm** TEST-UNIFORMITY-COLLISIONS$(p, n, \epsilon)$
Input: sample access to a distribution $p$ over $[n]$, and $\epsilon > 0$.
Output: "YES" if $\|p - U_n\|_2^2 \leq \epsilon^2/(2n)$; "NO" if $\|p - U_n\|_2^2 \geq \epsilon^2/n$.

1. Draw $m$ iid samples from $p$.
2. Let $\sigma_{ij}$ be an indicator variable which is 1 if samples $i$ and $j$ are the same and 0 otherwise.
3. Define the random variable $s = \sum_{i<j} \sigma_{ij}$ and the threshold $t = \binom{m}{2} \cdot \frac{1+3\epsilon^2/4}{n}$
4. If $s \geq t$ return "NO"; otherwise, return "YES".

The following theorem characterizes the performance of the above estimator:

**Theorem 1.** *The above estimator, when given $m$ samples drawn from a distribution $p$ over $[n]$ will, with probability at least $3/4$, distinguish the case that $\|p - U_n\|_2^2 \leq \epsilon^2/(2n)$ from the case that $||p - U_n||_2^2 \geq \epsilon^2/n$ provided that $m \geq 3200 n^{1/2}/\epsilon^2$.*

The rest of this section is devoted to the proof of Theorem 1. Note that the condition of the theorem is equivalent to testing whether $\|p\|_2^2 \leq \frac{1+\epsilon^2/2}{n}$ versus $\|p\|_2^2 \geq \frac{1+\epsilon^2}{n}$. Our tester takes $m = \frac{3200 n^{1/2}}{\epsilon^2}$ samples from $p$ and distinguishes between the two cases with probability at least $3/4$.

### 2.1 Analysis of TEST-UNIFORMITY-COLLISIONS

The analysis proceeds by bounding the expectation and variance of the estimator for the completeness and soundness cases, and applying Chebyshev's inequality. The novelty here is a tight analysis of the variance which leads to the optimal sample bound.

We start by recalling the following simple closed formula for the expected value:

**Lemma 2.** *We have that $\mathbb{E}[s] = \binom{m}{2} \|p\|_2^2$ .*

*Proof.* For any $i, j$, the probability that samples $i$ and $j$ are equal is $\|p\|_2^2$. By this and linearity of expectation, we get

$$\mathbb{E}[s] = \mathbb{E}\left[\sum_{ij} \sigma_{ij}\right] = \sum_{ij} \mathbb{E}[\sigma_{ij}] = \sum_{ij} \|p\|_2^2 = \binom{m}{2} \|p\|_2^2.$$

□

Thus, we see that in the completeness case the expected value is at most $\binom{m}{2} \cdot \frac{1+\epsilon^2/2}{n}$. In the soundness case, the expected value is at least $\binom{m}{2} \cdot \frac{1+\epsilon^2}{n}$. This motivates our choice of the threshold $t$ halfway between these expected values.

In order to argue that the statistic will be close to its expected value, we bound its variance from above and use Chebyshev's inequality. We bound the variance in two steps. First, we obtain the following bound:

**Lemma 3.** *We have that* $\mathrm{Var}[s] \le m^2 \cdot \|p\|_2^2 + m^3 \cdot (\|p\|_3^3 - \|p\|_2^4)$.

*Proof.* The lemma follows from the following chain of (in-)equalities:

$$\begin{aligned}
\mathrm{Var}[s] &= \mathbb{E}[s^2] - \mathbb{E}[s]^2 \\
&= \mathbb{E}\left[\sum_{i<j}\sum_{k<\ell} \sigma_{ij}\sigma_{k\ell}\right] - \binom{m}{2}^2 \|p\|_2^4 \\
&= \mathbb{E}\left[\sum_{\substack{i<j;\, k<\ell \\ \text{all distinct}}} \sigma_{ij}\sigma_{k\ell} + 2\sum_{i<j<\ell} \sigma_{ij}\sigma_{j\ell} + 2\sum_{\substack{i,k<j \\ i\neq k}} \sigma_{ij}\sigma_{kj} + \sum_{i<j}\sigma_{ij}^2\right] - \binom{m}{2}^2 \|p\|_2^4 \\
&= \binom{m}{2}\binom{m-2}{2}\|p\|_2^4 + 2\cdot\binom{m}{3}\|p\|_3^3 + 4\cdot\binom{m}{3}\|p\|_3^3 + \binom{m}{2}\|p\|_2^2 - \binom{m}{2}^2\|p\|_2^4 \\
&= \binom{m}{2}\cdot(\|p\|_2^2 - \|p\|_2^4) + m(m-1)(m-2)\cdot(\|p\|_3^3 - \|p\|_2^4) \\
&\le m^2 \cdot \|p\|_2^2 + m^3 \cdot (\|p\|_3^3 - \|p\|_2^4).
\end{aligned}$$

□

**Remark.** We note that the upper bound of the previous lemma is tight, up to constant factors. The $-m^3\|p\|_2^4$ term is critical for getting the optimal dependence on $\epsilon$ in the sample bound.

Continuing the analysis, we now derive an upper bound on the number of samples that suffices for the tester to have the desired success probability of $3/4$.

**Lemma 4.** *Let $\alpha$ satisfy $\|p\|_2^2 = \frac{1+\alpha}{n}$ and $\sigma$ be the standard deviation of $s$. The number of samples required by* Test-Uniformity-Collisions *is at most*

$$m \le \sqrt{\frac{5\sigma n}{|\alpha - 3\epsilon^2/4|}},$$

*in order to get error probability at most* $1/4$.

*Proof.* By Chebyshev's inequality, we have that

$$\Pr\left[\left|s - \binom{m}{2}\|p\|_2^2\right| \ge k\sigma\right] \le \frac{1}{k^2}$$

where $\sigma \triangleq \sqrt{\mathrm{Var}[s]}$.

We want $s$ to be closer to its expected value than the threshold is to its expected value because when this occurs, the tester outputs the right answer. Furthermore, to achieve our desired probability of error of at most $1/4$, we want this to happen with probability at least $3/4$. So, we set $k = 2$, and then we want

$$k\sigma \leq |\mathbb{E}[s] - t| = \left| \binom{m}{2} |\|p\|_2^2 - (1 + 3\epsilon^2/4)/n \right| = \binom{m}{2} |\alpha - 3\epsilon^2/4|/n$$

It suffices for the number of samples $m$ to satisfy the slightly stronger condition that

$$\sigma \leq m^2 \cdot \frac{|\alpha - 3\epsilon^2/4|}{5n}.$$

So, it suffices to have

$$m \geq \sqrt{\frac{5\sigma n}{|\alpha - 3\epsilon^2/4|}}.$$

We might as well take the smallest number of samples $m$ for which the tester works, which implies the desired inequality. □

We are now ready to show an upper bound on the number of samples in the completeness case, i.e., when $p$ is the uniform distribution.

**Lemma 5.** *In the completeness case, the required number of samples is at most*

$$m \leq \frac{6n^{1/2}}{\epsilon^2} \ ,$$

*in order to get error probability* $1/4$.

*Proof.* It is easy to see that $\|p\|_2^2 = 1/n$ and $\|p\|_3^3 = \|p\|_2^4 = 1/n^2$. Thus, by Lemma 3, $\sigma \leq m/n^{1/2}$. Also, we know $\alpha = 0$ when $p$ is uniform. Substituting these two facts into Lemma 4 and solving for $m$ gives

$$m \leq \frac{6n^{1/2}}{\epsilon^2}.$$

□

We now turn to the soundness case, where $p$ is far from uniform. By Lemma 4, it suffices to bound from above the variance $\sigma^2$. We proceed by a case analysis based on whether the term $m^2\|p\|_2^2$ or $m^3(\|p\|_3^3 - \|p\|_2^4)$ contributes more to the variance.

### 2.1.1 Case when $m^2\|p\|_2^2$ is Larger

**Lemma 6.** *Consider the soundness case, where* $\|p\|_2^2 = (1+\alpha)/n$ *for* $\alpha \geq \epsilon^2$. *If* $m^2\|p\|_2^2$ *contributes more to the variance, i.e., if* $m^2\|p\|_2^2 \geq m^3(\|p\|_3^3 - \|p\|_2^4)$, *then the required number of samples is at most*

$$m \leq \frac{48n^{1/2}}{\epsilon^2}$$

*in order to get error probability* $1/4$.

*Proof.* Suppose that $m^2\|p\|_2^2 \geq m^3(\|p\|_3^3 - \|p\|_2^4)$. Then $\sigma^2 \leq 2m^2\|p\|_2^2 = 2m^2(1+\alpha)/n$. Substituting this into Lemma 4 and solving for $m$ gives that the necessary number of samples is at most

$$m \leq 8n^{1/2} \cdot \frac{\sqrt{1+\alpha}}{(\alpha - 3\epsilon^2/4)}.$$

Using calculus to maximize this expression by varying $\alpha$, one gets that $\alpha = \epsilon^2$ maximizes the expression. Thus,

$$m \leq 32n^{1/2} \cdot \frac{\sqrt{1+\epsilon^2}}{\epsilon^2} \leq 32n^{1/2} \cdot \frac{\sqrt{2}}{\epsilon^2} < \frac{48n^{1/2}}{\epsilon^2}.$$

□

### 2.1.2 Case when $m^3(\|p\|_3^3 - \|p\|_2^4)$ is Larger

**Lemma 7.** *Consider the soundness case, where $\|p\|_2^2 = (1+\alpha)/n$ for $\alpha \geq \epsilon^2$. If $m^3(\|p\|_3^3 - \|p\|_2^4)$ contributes more to the variance, i.e., if $m^3(\|p\|_3^3 - \|p\|_2^4) \geq m^2\|p\|_2^2$, then the required number of samples is at most*

$$m \leq \frac{3200n^{1/2}}{\epsilon^2}$$

*in order to get error probability $\leq 1/4$.*

*Proof.* Suppose that $m^3(\|p\|_3^3 - \|p\|_2^4) \geq m^2\|p\|_2^2$. Then $\sigma^2 \leq 2m^3(\|p\|_3^3 - \|p\|_2^4)$. Substituting this into Lemma 4 and solving for $m$ gives that the necessary number of samples is at most

$$m \leq 50n^2 \cdot \frac{\|p\|_3^3 - \|p\|_2^4}{(\alpha - 3\epsilon^2/4)^2}.$$

Let us parameterize $p$ as $p_i = 1/n + a_i$ for some vector $a$. Then we have $\|a\|_2^2 = \alpha/n$, and we can write

$$\begin{aligned}
m &\leq 50n^2 \cdot \frac{\|p\|_3^3 - \|p\|_2^4}{(\alpha - 3\epsilon^2/4)^2} \leq 50n^2 \cdot \frac{\|p\|_3^3 - \|p\|_2^4}{(\alpha/4)^2} && (\text{since } \epsilon^2 \leq \alpha) \\
&\leq 50n^2 \cdot \frac{\|p\|_3^3 - \frac{1}{n^2}}{(\alpha/4)^2} = 50n^2 \cdot \frac{\left[\sum_{i=1}^n (1/n + a_i)^3\right] - \frac{1}{n^2}}{(\alpha/4)^2} \\
&= 50n^2 \cdot \frac{\left[\frac{1}{n^2} + \frac{3}{n^2}\sum_{i=1}^n a_i + \frac{3}{n}\sum_{i=1}^n a_i^2 + \sum_{i=1}^n a_i^3\right] - \frac{1}{n^2}}{(\alpha/4)^2} \\
&= 50n^2 \cdot \frac{\frac{3}{n^2}\sum_{i=1}^n a_i + \frac{3}{n}\sum_{i=1}^n a_i^2 + \sum_{i=1}^n a_i^3}{(\alpha/4)^2} \\
&= 50n^2 \cdot \frac{\frac{3}{n}\sum_{i=1}^n a_i^2 + \sum_{i=1}^n a_i^3}{(\alpha/4)^2} && \left(\text{since } \sum_{i=1}^n a_i = 0\right) \\
&\leq 50n^2 \cdot \frac{\frac{3}{n}\|a\|_2^2 + \|a\|_3^3}{(\alpha/4)^2} \leq 50n^2 \cdot \frac{\frac{3}{n}\|a\|_2^2 + \|a\|_2^3}{(\alpha/4)^2} \\
&= 50n^2 \cdot \frac{\frac{3}{n}(\alpha/n) + (\alpha/n)^{3/2}}{(\alpha/4)^2} = \frac{2400}{\alpha} + \frac{800n^{1/2}}{\sqrt{\alpha}} \\
&\leq \frac{2400}{\epsilon^2} + \frac{800n^{1/2}}{\sqrt{\epsilon^2}} && (\text{since } \epsilon^2 \leq \alpha) \\
&\leq \frac{3200n^{1/2}}{\epsilon^2}.
\end{aligned}$$

□

Note that, as mentioned earlier, if we had ignored the $-\|p\|_2^4$ term, we would have had an $\Omega(1/\epsilon^4)$ term in our bound, which would have given us the wrong dependence on $\epsilon$.

Theorem 1 now follows as an immediate consequence of these last three lemmas.

**Remark.** It is worth noting that the collisions statistic analyzed in this section is very similar to the chi-squared-like uniformity tester in [DKN15b] – itself a simplification of similar testers in [CDVV14, VV14] – which also achieves the optimal sample complexity of $O(n^{1/2}/\epsilon^2)$. Specifically, if $X_i$ denotes the number of times we see the $i$-th domain element in the sample, the [DKN15b] statistic is $\sum_i (X_i - m/n)^2 - X_i = \sum_{i<j} \sigma_{ij} - 2\frac{m}{n}\sum_i X_i + \frac{m^2}{n}$. We note that the [DKN15b] analysis uses Poissonization; i.e., instead of drawing $m$ samples from the distribution, we draw $\mathrm{Poi}(m)$ samples. Without Poissonization, the aforementioned statistic simplifies to $s - \frac{m^2}{n}$, where $s$ is the collisions statistic. While the non-Poissonized versions of the two testers are equivalent, the Poissonized versions are not. Specifically, the Poissonized version of the [DKN15b] uniformity tester has sufficiently good variance to yield the sample-optimal bound. On the other hand, the Poissonized version of the collisions statistic does not have good variance: Specifically, its variance does not have the $-\|p\|_2^4$ term which – as noted earlier – is necessary to get the optimal $\epsilon$ dependence.

## 3 Testing Closeness via Collisions

Given samples from two unknown distributions $p, q$ over $[n]$ with the promise that $\max\{\|p\|_2^2, \|q\|_2^2\} \leq b$, we want to distinguish between the cases that $\|p - q\|_2 \leq \epsilon/2$ versus $\|p - q\|_2 \geq \epsilon$. We show that a natural collisions-based tester succeeds in this task with $O(b^{1/2}/\epsilon^2)$ samples. The estimator we analyze is a slight variant of the $\ell_2$ tester in [BFR$^+$00], described in pseudocode below.

We define the number of self-collisions in a sequence of samples from a distribution as $\sum_{i<j} \sigma_{ij}$, where $\sigma_{ij}$ is the indicator variable denoting whether samples $i$ and $j$ are the same. Similarly, we define the number of cross-collisions between two sequences of samples as $\sum_{i,j} \ell_{ij}$, where $\ell_{ij}$ is the indicator variable denoting whether sample $i$ from the first sequence is the same as sample $j$ from the second sequence.

**Algorithm** TEST-CLOSENESS-COLLISIONS$(p, q, n, b, \epsilon)$
Input: sample access to distribution $p, q$ over $[n]$, $\epsilon, b > 0$.
Output: "YES" if $\|p - q\|_2 \leq \epsilon/2$; "NO" if $\|p - q\|_2 \geq \epsilon$.

1. Draw two multisets $S_p, S_q$ of $m$ iid samples from $p, q$. Let $C_1$ denote the number of self-collisions of $S_p$, $C_2$ denote the number of self-collisions of $S_q$, and $C_3$ denote the number of cross-collisions between $S_p$ and $S_q$.
2. Define the random variable $Z = C_1 + C_2 - \frac{m-1}{m} \cdot C_3$ and the threshold $t = \binom{m}{2}\epsilon^2/2$.
3. If $Z \geq t$ return "NO"; otherwise, return "YES".

The following theorem characterizes the performance of the above estimator:

**Theorem 8.** *There exists an absolute constant $c$ such that the above estimator, when given $m$ samples drawn from each of two distributions, $p, q$ over $[n]$ will, with probability at least $3/4$, distinguish the case $||p - q||_2 \leq \epsilon/2$ from the case that $||p - q||_2 \geq \epsilon$ provided that $m \geq c \cdot \frac{b^{1/2}}{\epsilon^2}$, where $b$ is an upper bound on $||p||_2^2, ||q||_2^2$.*

**3.1 Analysis of Test-Closeness-Collisions** Let $X_i, Y_i$ be the number of times we see the element $i$ in each set of samples $S_p$ and $S_q$, respectively. The above random variables are distributed as follows: $X_i \sim Bin(m, p_i), Y_i \sim Bin(m, q_i)$. Note that the statistic $Z$ can be written as

$$Z = \frac{m-1}{2m} \sum_{i=1}^{n} \left[(X_i - Y_i)^2 - X_i - Y_i\right] + \frac{1}{2m} \sum_{i=1}^{n} [X_i(X_i-1) + Y_i(Y_i-1)] = \frac{m-1}{2m} A + \frac{1}{2m} B ,$$

where $A = \sum_{i=1}^{n} \left[(X_i - Y_i)^2 - X_i - Y_i\right]$ and $B = \sum_{i=1}^{n} [X_i(X_i-1) + Y_i(Y_i-1)]$. Note that

$$\mathrm{Var}[Z] \le 4 \cdot \max\left\{ \frac{(m-1)^2}{4m^2} \mathrm{Var}[A], \frac{1}{4m^2} \mathrm{Var}[B] \right\} .$$

Note that $B$ essentially corresponds to the number of collisions within two disjoint sets of samples, hence we already have an upper bound on its variance. The bulk of the analysis goes into bounding from above the variance of $A = \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} \left[(X_i - Y_i)^2 - X_i - Y_i\right]$.

**Remark.** The $\ell_2$ collision-based tester we analyze here is closely related to the $\ell_2$-tester of [CDVV14]. Specifically, the $A$ term in the expression for $Z$ has the same formula as the $\ell_2$-tester of [CDVV14]. However, a key difference is that the statistic of [CDVV14] is Poissonized, which is crucial for its analysis.

We now proceed to analyze the collision-based closeness tester. We start with a simple formula for its expectation:

**Lemma 9.** *For the expectation of the statistic $Z$ in the closeness tester, we have:*

$$\mathbb{E}[Z] = \binom{m}{2} \|p - q\|_2^2 . \tag{1}$$

*Proof.* Viewing $p$ and $q$ as vectors, we have

$$\mathbb{E}[Z] = \mathbb{E}[C_1 + C_2 - \frac{m-1}{m} \cdot C_3] = \binom{m}{2}(p \cdot p) + \binom{m}{2}(q \cdot q) - \frac{m-1}{m} \cdot m^2 (p \cdot q) = \binom{m}{2} \|p - q\|_2^2.$$

□

For the variance, we show the following upper bound:

**Lemma 10.** *For the variance of the statistic $Z$ in the closeness tester, we have:*

$$\mathrm{Var}[Z] \le 116 m^2 b + 16 m^3 \|p - q\|_4^2 b^{1/2} .$$

To prove this lemma, we will use the following proposition, whose proof is deferred to the following subsection.

**Proposition 11.** *We have that* $\mathrm{Var}[A] \le 100 m^2 b + 8 m^3 \sum_i (p_i - q_i)(p_i^2 - q_i^2)$.

*Proof of Lemma 10.* Recall that by Lemma 3 we have

$$\mathrm{Var}[B] \le 4m^2(\|p\|_2^2 + \|q\|_2^2) + 4m^3(\|p\|_3^3 - \|p\|_2^4 + \|q\|_3^3 - \|q\|_2^4) .$$

Combined with Proposition 11, we obtain:

$$\begin{aligned}\mathrm{Var}[Z] &\le 4\cdot\max\left\{\frac{(m-1)^2}{4m^2}\mathrm{Var}[A], \frac{1}{4m^2}\mathrm{Var}[B]\right\}\\ &\le \max\{100m^2b + 8m^3\sum_i (p_i-q_i)(p_i^2-q_i^2),\\ &\qquad 4(\|p\|_2^2+\|q\|_2^2) + 4m(\|p\|_3^3 - \|p\|_2^4 + \|q\|_3^3 - \|q\|_2^4)\}\ .\end{aligned}$$

The second term in the max statement is at most $16mb$. Thus, we have

$$\begin{aligned}\mathrm{Var}[Z] &\le 116(m-1)^2b + 8m(m-1)^2\sum_i (p_i-q_i)(p_i^2-q_i^2)\\ &\le 116m^2b + 8m^3\sum_i (p_i-q_i)^2(p_i+q_i)\\ &\le 116m^2b + 8m^3\sqrt{\sum_i (p_i-q_i)^4\sum_i (p_i+q_i)^2} && \text{(by the Cauchy-Schwarz inequality)}\\ &\le 116m^2b + 16m^3\|p-q\|_4^2 b^{1/2} && (\text{since } \sum_i (p_i+q_i)^2 \le 4b)\ .\end{aligned}$$

□

**3.2 Proof of Theorem 8** By Lemma 10, we have that

$$\mathrm{Var}[Z] \le 116m^2b + 16m^3\|p-q\|_4^2 b^{1/2} \le 116m^2b + 16m^3\|p-q\|_2^2 b^{1/2}.$$

We wish to show we can distinguish the completeness case (i.e., $\|p-q\|_2 \le \epsilon/2$) from the soundness case (i.e., $\|p-q\|_2 \ge \epsilon$). Set $\alpha = \|p-q\|_2^2$. Then we are promised that either $\alpha \ge \epsilon^2$ or $\alpha \le \epsilon^2/4$. Recall we chose $t = \frac{\binom{m}{2}\epsilon^2}{2}$ and that Lemma 9 says that $\mathbb{E}[Z] = \binom{m}{2}\alpha$.

Since

$$\mathbb{E}[Z|\text{completeness case}] \le t \le \mathbb{E}[Z|\text{soundness case}],$$

the only way we fail to distinguish the completeness and soundness cases is if $Z$ deviates from its expectation additively by at least

$$|t - \mathbb{E}[Z]| = \left|\frac{\binom{m}{2}\epsilon^2}{2} - \binom{m}{2}\alpha\right| \ge \binom{m}{2}\max\{\alpha,\epsilon^2\}/4,$$

where the last inequality follows by the promise on $\alpha$ in the completeness and soundness cases.[2] By Chebyshev's inequality, the probability this happens is at most

$$\begin{aligned}\Pr[\ |Z-\mathbb{E}[Z]| \ge \binom{m}{2}\max\{\alpha,\epsilon^2\}/4\ ] &\le \frac{\mathrm{Var}[Z]}{[t-\mathbb{E}[Z]]^2} \le \frac{116m^2b + 16m^3\alpha b^{1/2}}{[\binom{m}{2}\max\{\alpha,\epsilon^2\}/4]^2}\\ &\le \frac{32768\cdot b}{m^2\epsilon^4} + \frac{4096\cdot b^{1/2}}{m}\cdot\min\left\{\frac{1}{\alpha},\frac{\alpha}{\epsilon^4}\right\}\\ &\le \frac{32768\cdot b}{m^2\epsilon^4} + \frac{4096\cdot b^{1/2}}{m\epsilon^2},\end{aligned}$$

where we simplified using the assumption that $m \ge 2$. Thus, if we set $m = O(\frac{b^{1/2}}{\epsilon^2})$, we get a constant probability of error in both cases as desired. □

[2] In the completeness case where $\alpha \le \varepsilon^2/4$ and $\mathbb{E}[Z] = \binom{m}{2}\alpha$, $Z$ has to deviate by at least $\binom{m}{2}\varepsilon^2/4 \ge \binom{m}{\epsilon}^2\alpha$ to cross the threshold $t = \binom{m}{2}\varepsilon^2/2$. In the soundness case where $\alpha \ge \varepsilon^2$, $Z$ has to deviate by at least $\binom{m}{2}\alpha/2 \ge \epsilon/2$ to cross the threshold $t$.

### 3.3 Proof of Proposition 11

Recall that $A = \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} \left[(X_i - Y_i)^2 - X_i - Y_i\right]$, hence $\mathrm{Var}(A) = \sum_{i=1}^{n} \mathrm{Var}(A_i) + \sum_{i \neq j} \mathrm{Cov}(A_i, A_j)$. We proceed to bound from above the individual variances and covariances via a sequence of elementary but quite tedious calculations.

#### 3.3.1 Bounding $\mathrm{Var}(A_i)$:

Since

$$A_i = (X_i - Y_i)^2 - X_i - Y_i = X_i^2 + Y_i^2 - 2X_iY_i - X_i - Y_i \ ,$$

we can write:

$$\begin{aligned}\mathrm{Var}(A_i) &= \mathrm{Var}(X_i^2) + \mathrm{Var}(Y_i^2) + 4\mathrm{Var}(X_iY_i) + \mathrm{Var}(X_i) + \mathrm{Var}(Y_i) \\ &\quad + 2 \cdot [-2\,\mathrm{Cov}(X_i^2, X_iY_i) - \mathrm{Cov}(X_i^2, X_i) - 2\,\mathrm{Cov}(Y_i^2, X_iY_i) - \mathrm{Cov}(Y_i^2, Y_i) \\ &\quad + 2\,\mathrm{Cov}(X_iY_i, X_i) + 2\,\mathrm{Cov}(X_iY_i, Y_i)] \ .\end{aligned}$$

We proceed to calculate the individual quantities:

(a)

$$\begin{aligned}\mathrm{Cov}(X_i^2, X_i) &= \sum_{r,s,t \in [m]} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_t = i]) \\ &= \sum_{r \in [m]} \mathrm{Cov}([\sigma_r = i], [\sigma_r = i]) + 2 \sum_{r,s \in [m],\ r \neq s} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_r = i]) \\ &= mp_i(1 - p_i) + 2(m^2 - m)(\mathbb{E}[[\sigma_r = \sigma_s = i] \cdot [\sigma_r = i]] - p_i^2 p_i) \\ &= mp_i(1 - p_i) + 2(m^2 - m)(p_i^2 - p_i^3) \\ &= mp_i(1 - p_i)[1 + 2p_i(m - 1)] \\ &= mp_i(1 - p_i)[1 - 2p_i + 2p_i m] \\ &= mp_i(1 - p_i)(1 - 2p_i) + 2m^2 p_i^2 (1 - p_i) \ .\end{aligned}$$

(b)

$$\mathrm{Cov}(X_i^2, X_iY_i) = \mathbb{E}[X_i^3 Y_i] - \mathbb{E}[X_i^2] \cdot \mathbb{E}[X_iY_i] = \mathrm{Cov}(X_i^2, X_i) \cdot \mathbb{E}[Y_i] = m^2 p_i q_i (1 - p_i)(1 - 2p_i) + 2m^3 p_i^2 q_i (1 - p_i) \ .$$

(c)

$$\mathrm{Cov}(X_i, X_iY_i) = \mathrm{Var}(X_i) \cdot \mathbb{E}[Y_i] = m^2 p_i (1 - p_i) q_i \ .$$

(d)

$$\begin{aligned}\mathrm{Var}(X_i^2) =& \mathbb{E}[X_i^4] - (\mathbb{E}[X_i^2])^2 \\ =& mp_i(1 - 7p_i + 7mp_i + 12p_i^2 - 18mp_i^2 + 6m^2p_i^2 - 6p_i^3 \\ & + 11mp_i^3 - 6m^2p_i^3 + m^3p_i^3) - (mp_i - mp_i^2 + m^2p_i^2)^2 \\ =& mp_i - 7mp_i^2 + 7m^2p_i^2 + 12mp_i^3 - 18m^2p_i^3 + 6m^3p_i^3 - 6mp_i^4 + 11m^2p_i^4 - 6m^3p_i^4 + m^4p_i^4 \\ & - (m^2p_i^2 + m^2p_i^4 + m^4p_i^4 - 2m^2p_i^3 + 2m^3p_i^3 - 2m^3p_i^4) \\ =& mp_i - 7mp_i^2 + 6m^2p_i^2 + 12mp_i^3 - 16m^2p_i^3 + 4m^3p_i^3 - 6mp_i^4 + 10m^2p_i^4 - 4m^3p_i^4 \\ =& mp_i - 7mp_i^2 + 6m^2p_i^2 + 12mp_i^3 - 6mp_i^4 - 16m^2p_i^3 + 4m^3p_i^3 + 10m^2p_i^4 - 4m^3p_i^4 \ .\end{aligned}$$

(e)

$$\begin{aligned}\mathrm{Var}(X_iY_i) &= \mathbb{E}[X_i^2Y_i^2] - (\mathbb{E}[X_iY_i])^2 = \mathbb{E}[X_i^2]\mathbb{E}[Y_i^2] - (\mathbb{E}[X_i]\mathbb{E}[Y_i])^2 \\ &= (mp_i - mp_i^2 + m^2p_i^2)\cdot(mq_i - mq_i^2 + m^2q_i^2) - m^4p_i^2q_i^2 \\ &= m^2p_iq_i + m^2p_i^2q_i^2 - m^2(p_iq_i^2 + p_i^2q_i) + m^3(p_iq_i^2 + p_i^2q_i) - 2m^3p_i^2q_i^2 \ .\end{aligned}$$

So, we get:

$$\begin{aligned}\mathrm{Var}(A_i) &= mp_i - 7mp_i^2 + 6m^2p_i^2 + 12mp_i^3 - 6mp_i^4 - 16m^2p_i^3 + 4m^3p_i^3 + 10m^2p_i^4 - 4m^3p_i^4 \\ &+ mq_i - 7mq_i^2 + 6m^2q_i^2 + 6mq_i^3 - 16m^2q_i^3 + 4m^3q_i^3 + 10m^2q_i^4 - 4m^3q_i^4 \\ &+ 4(m^2(p_iq_i + p_i^2q_i^2 - p_iq_i^2 - p_i^2q_i) + m^3(p_iq_i^2 + p_i^2q_i) - 2m^3p_i^2q_i^2) \\ &+ mp_i(1-p_i) + mq_i(1-q_i) - 4(m^2p_i(1-p_i)(1-2p_i) + 2m^3p_i^2(1-p_i))q_i \\ &- 2(mp_i(1-p_i)(1-2p_i) + 4m^2p_i^2(1-p_i)) - 4(m^2q_i(1-q_i)(1-2q_i) + 2m^3q_i^2(1-q_i))p_i \\ &- 2mq_i(1-q_i)(1-2q_i) - 4m^2q_i^2(1-q_i) + 4m^2p_i(1-p_i)q_i + 4m^2q_i(1-q_i)p_i \\ &= m[p_i - 7p_i^2 + 12p_i^3 - 6p_i^4 + q_i - 7q_i^2 + 12q_i^3 - 6q_i^4 + p_i - p_i^2 + q_i - q_i^2 \\ &\quad - 2p_i(1-p_i)(1-2p_i) - 2q_i(1-q_i)(1-2q_i)] \\ &+ m^2[-4p_iq_i(1-p_i)(1-2p_i) - 4p_iq_i(1-q_i)(1-2q_i) + 4p_iq_i(2-p_i-q_i) \\ &\quad + 6p_i^2 - 16p_i^3 + 10p_i^4 + 6q_i^2 - 16q_i^3 + 10q_i^4 + 4p_iq_i(1 + p_iq_i - p_i - q_i) \\ &\quad - 4p_i^2 + 4p_i^3 - 4q_i^2 + 4q_i^3] \\ &+ m^3[4p_i^3 - 4p_i^4 + 4q_i^3 - 4q_i^4 + 4p_iq_i(p_i+q_i) - 8p_i^2q_i^2 - 8p_1^2q_i - 8p_iq_i^2 + 8p_1^3q_i + 8p_iq_i^3] \\ &= m[-2p_i^2 + 8p_i^3 - 6p_i^4 - 2q_i^2 + 8q_i^3 - 6q_i^4] \\ &+ m^2[2(p_i+q_i)^2 - 12p_i^3 + 10p_i^4 + 4p_i^2q_i - 8p_i^3q_i + 4p_iq_i^2 + 4p_i^2q_i^2 - 12q_i^3 - 8p_iq_i^3 + 10q_i^4)] \\ &+ 4m^3(p_i-q_i)^2[p_i(1-p_i) + q_i(1-q_i)] \\ &\le 8m(p_i^3+q_i^3) + 12m^2(p_i+q_i)^2 + 4m^3(p_i-q_i)^2\,(p_i+q_i) \\ &\le 20m^2(p_i+q_i)^2 + 4m^3(p_i-q_i)^2(p_i+q_i) \ .\end{aligned}$$

**3.4 Bounding the Covariances** It suffices to show that the covariances of $A_i$ and $A_j$, for $i \neq j$, are appropriately bounded from above. Let $i \neq j$. Note that if $\sigma_r$ is the result of sample $k$, we have:

$$\mathrm{Cov}(X_i, X_j) = \sum_{r,u\in[m]} \mathrm{Cov}([\sigma_r = i], [\sigma_u = j]) = \sum_{r\in[m]} \mathrm{Cov}([\sigma_r = i], [\sigma_r = j]) = -mp_ip_j \ .$$

Similarly,

$$\begin{aligned}\mathrm{Cov}(X_i^2, X_j) &= \sum_{r,s,t\in[m]} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_t = j]) \\ &= \sum_{r\in[m]} \mathrm{Cov}([\sigma_r = i], [\sigma_r = j]) + 2\sum_{r,s\in[m],\, r\neq s} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_r = j]) \\ &= -mp_ip_j - 2m(m-1)p_i^2p_j.\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{Cov}(X_i^2, X_j^2) = \sum_{r,s,t,u\in[m]} &\mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_t = \sigma_u = j]) \\
= 4 &\sum_{\text{unique } r,s,u\in[m]} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_r = \sigma_u = j]) \\
&+ 2 \sum_{\text{unique } r,s\in[m]} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_r = \sigma_s = j]) \\
&+ 2 \sum_{\text{unique } r,s\in[m]} \mathrm{Cov}([\sigma_r = \sigma_s = i], [\sigma_r = j]) \\
&+ 2 \sum_{\text{unique } r,t\in[m]} \mathrm{Cov}([\sigma_r = i], [\sigma_r = \sigma_t = j]) \\
&+ \sum_{r\in[m]} \mathrm{Cov}([\sigma_r = i], [\sigma_r = j]) \\
&= -mp_ip_j - 2m(m-1)(p_i^2 p_j + p_i p_j^2 + p_i^2 p_j^2) - 4m(m-1)(m-2)p_i^2 p_j^2 \\
&= -mp_ip_j - 2m(m-1)(p_i^2 p_j + p_i p_j^2) - 2m(m-1)(2m-3)p_i^2 p_j^2.
\end{aligned}$$

And,

$$\begin{aligned}
\mathrm{Cov}(X_iY_i, X_jY_j) &= \mathbb{E}[X_iY_iX_jY_j] - \mathbb{E}[X_iY_i]\mathbb{E}[X_jY_j] \\
&= \mathbb{E}[X_iX_j]\mathbb{E}[Y_iY_j] - \mathbb{E}[X_i]\mathbb{E}[Y_i]\mathbb{E}[X_j]\mathbb{E}[Y_j] \\
&= (\mathrm{Cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j]) \cdot (\mathrm{Cov}(Y_i, Y_j) + \mathbb{E}[Y_i]\mathbb{E}[Y_j]) - \mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[Y_i]\mathbb{E}[Y_j] \\
&= (m^2 - 2m^3)p_ip_jq_iq_j.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{Cov}(X_iY_i, X_j) &= \mathbb{E}[X_iY_iX_j] - \mathbb{E}[X_iY_i]\mathbb{E}[X_j] \\
&= \mathbb{E}[X_iX_j]\mathbb{E}[Y_i] - \mathbb{E}[X_i]\mathbb{E}[Y_i]\mathbb{E}[X_j] \\
&= (\mathrm{Cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j]) \cdot \mathbb{E}[Y_i] - \mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[Y_i] \\
&= \mathrm{Cov}(X_i, X_j)\mathbb{E}[Y_i]\ .
\end{aligned}$$

Similar equations hold if we swap $i$ and $j$ and/or we swap $X$ and $Y$. Because covariance is bilinear, this gives us all the information we need in order to exactly compute $\mathrm{Cov}(A_i, A_j)$. In particular, by setting $W_i = X_i - Y_i$, we have:

$$\begin{aligned}
\mathrm{Cov}(A_i, A_j) &= \mathrm{Cov}(W_i^2 - X_i - Y_i, W_j^2 - X_j - Y_j) \\
&= \mathrm{Cov}(X_i, X_j) + \mathrm{Cov}(Y_i, Y_j) + \mathrm{Cov}(X_i, Y_j) + \mathrm{Cov}(X_j, Y_i) - \mathrm{Cov}(W_i^2, X_j) \\
&\quad - \mathrm{Cov}(W_i^2, Y_j) - \mathrm{Cov}(W_j^2, X_i) - \mathrm{Cov}(W_j^2, Y_i) + \mathrm{Cov}(W_i^2, W_j^2)\ .
\end{aligned}$$

For the summands we have:

(a)

$$\begin{aligned}
\mathrm{Cov}(W_i^2, X_j) &= \mathrm{Cov}((X_i - Y_i)^2, X_j) = \mathrm{Cov}(X_i^2, X_j) - 2\,\mathrm{Cov}(X_iY_i, X_j) \\
&= -mp_ip_j - 2m(m-1)p_i^2 p_j + 2m^2 p_ip_jq_i \\
&= -mp_ip_j(1 - 2p_i) + 2m^2 p_ip_j(q_i - p_i)\ .
\end{aligned}$$

(b)

$$\mathrm{Cov}(W_i^2, Y_j) = -mq_iq_j(1-2q_i) + 2m^2q_iq_j(p_i - q_i)\ .$$

(c)

$$\mathrm{Cov}(W_j^2, X_i) = -mp_ip_j(1-2p_j) + 2m^2p_ip_j(q_j - p_j)\ .$$

(d)

$$\mathrm{Cov}(W_j^2, Y_i) = -mq_iq_j(1-2q_j) + 2m^2q_iq_j(p_j - q_j)\ .$$

(e)

$$\begin{aligned}
\mathrm{Cov}(W_i^2, W_j^2) =& \mathrm{Cov}(X_i^2, X_j^2) + \mathrm{Cov}(Y_i^2, Y_j^2) + 4\,\mathrm{Cov}(X_iY_i, X_jY_j) \\
& - 2\,\mathrm{Cov}(X_i^2, X_jY_j) - 2\,\mathrm{Cov}(X_j^2, X_iY_i) - 2\,\mathrm{Cov}(Y_i^2, X_jY_j) - 2\,\mathrm{Cov}(Y_j^2, X_iY_i) \\
=& \mathrm{Cov}(X_i^2, X_j^2) + \mathrm{Cov}(Y_i^2, Y_j^2) + 4\,\mathrm{Cov}(X_iY_i, X_jY_j) \\
& - 2\,\mathrm{Cov}(X_i^2, X_j)\mathbb{E}[Y_j] - 2\,\mathrm{Cov}(X_j^2, X_i)\mathbb{E}[Y_i] \\
& - 2\,\mathrm{Cov}(Y_i^2, Y_j)\mathbb{E}[X_j] - 2\,\mathrm{Cov}(Y_j^2, Y_i)\mathbb{E}[X_i] \\
=& - mp_ip_j - 2m(m-1)(p_i^2p_j + p_ip_j^2) - 2m(m-1)(2m-3)p_i^2p_j^2 \\
& - mq_iq_j - 2m(m-1)(q_i^2q_j + q_iq_j^2) - 2m(m-1)(2m-3)q_i^2q_j^2 \\
& + 4(m^2 - 2m^3)p_ip_jq_iq_j + 2m^2p_ip_jq_j + 4m^2(m-1)p_i^2p_jq_j \\
& + 2m^2p_ip_jq_i + 4m^2(m-1)p_j^2p_iq_i + 2m^2q_iq_jp_j + 4m^2(m-1)q_i^2q_jp_j \\
& + 2m^2q_iq_jp_i + 4m^2(m-1)q_j^2q_ip_i\ .
\end{aligned}$$

By substituting, we get:

$$\begin{aligned}
\mathrm{Cov}(A_i, A_j) =& \mathrm{Cov}(X_i, X_j) + \mathrm{Cov}(Y_i, Y_j) - \mathrm{Cov}(W_i^2, X_j) \\
& - \mathrm{Cov}(W_i^2, Y_j) - \mathrm{Cov}(W_j^2, X_i) - \mathrm{Cov}(W_j^2, Y_i) + \mathrm{Cov}(W_i^2, W_j^2) \\
=& - m(p_ip_j + q_iq_j) \\
& + mp_ip_j(1-2p_i) - 2m^2p_ip_j(q_i - p_i) + mq_iq_j(1-2q_i) - 2m^2q_iq_j(p_i - q_i) \\
& + mp_ip_j(1-2p_j) - 2m^2p_ip_j(q_j - p_j) + mq_iq_j(1-2q_j) - 2m^2q_iq_j(p_j - q_j) \\
& + \mathrm{Cov}(W_i^2, W_j^2) \\
=& - 2m^2[p_ip_j(q_i + q_j) + q_iq_j(p_i + p_j)] + 2m^2[p_ip_j(q_i + q_j) + q_iq_j(p_i + p_j)] \\
& - 2m(m-1)(2m-3)p_i^2p_j^2 - 2m(m-1)(2m-3)q_i^2q_j^2 \\
& + 4(m^2 - 2m^3)p_ip_jq_iq_j + 4m^2(m-1)(p_iq_j + p_jq_i)(p_ip_j + q_iq_j) \\
=& - 6m(p_i^2p_j^2 + q_i^2q_j^2) \\
& + m^2[10(p_i^2p_j^2 + q_i^2q_j^2) + 4p_ip_jq_iq_j - 4(p_iq_j + p_jq_i)(p_ip_j + q_iq_j)] \\
& - m^3[4(p_i^2p_j^2 + q_i^2q_j^2) + 8p_ip_jq_iq_j - 4(p_iq_j + p_jq_i)(p_ip_j + q_iq_j)] \\
=& - 6m(p_i^2p_j^2 + q_i^2q_j^2) \\
& + 2m^2[5(p_i^2p_j^2 + q_i^2q_j^2) + 2p_ip_jq_iq_j - 2(p_iq_j + p_jq_i)(p_ip_j + q_iq_j)] \\
& - 4m^3[(p_ip_j + q_iq_j)^2 - (p_iq_j + p_jq_i)(p_ip_j + q_iq_j)]\ .
\end{aligned}$$

In summary,

$$\begin{aligned}\mathrm{Cov}(A_i, A_j) =& -6m(p_i^2 p_j^2 + q_i^2 q_j^2) \\ & + 2m^2[(5p_i^2 p_j^2 + 5q_i^2 q_j^2) - 6p_i p_j q_i q_j - 2p_i q_i (p_j - q_j)^2 - 2p_j q_j (p_i - q_i)^2] \\ & - 4m^3 (p_i - q_i)(p_j - q_j)(p_i p_j + q_i q_j) \; .\end{aligned}$$

The total contribution of the covariances to the variance for all $i \neq j$ is $\sum_{i \neq j} \mathrm{Cov}(A_i, A_j)$. We consider the coefficients on each of the powers of $m$ separately. We have:

$$\begin{aligned}[m^3] \sum_{i \neq j} \mathrm{Cov}(A_i, A_j) &= -4 \sum_{i \neq j} (p_i - q_i)(p_j - q_j)(p_i p_j + q_i q_j) \\ &= 4 \sum_i (p_i - q_i)^2 (p_i^2 + q_i^2) - 4 \sum_{i,j} (p_i - q_i)(p_j - q_j)(p_i p_j + q_i q_j) \\ &\leq 4 \sum_i (p_i - q_i)^2 (p_i + q_i) - 4 \sum_{i,j} (p_i - q_i)(p_j - q_j)(p_i p_j + q_i q_j) \\ &= 4 \sum_i (p_i - q_i)^2 (p_i + q_i) - 4(p - q)^\intercal (pp^\intercal + qq^\intercal)(p - q) \\ &\leq 4 \sum_i (p_i - q_i)^2 (p_i + q_i).\end{aligned}$$

Also, $[m] \sum_{i \neq j} \mathrm{Cov}(A_i, A_j) \leq 0$.
Finally, we have

$$\begin{aligned}[m^2] \sum_{i \neq j} \mathrm{Cov}(A_i, A_j) &= 2 \sum_{i \neq j} [(5p_i^2 p_j^2 + 5q_i^2 q_j^2) - 6p_i p_j q_i q_j - 2p_i q_i (p_j - q_j)^2 - 2p_j q_j (p_i - q_i)^2] \\ &\leq 10 \sum_{i \neq j} (p_i^2 p_j^2 + q_i^2 q_j^2) \\ &\leq 10 \sum_{i,j} (p_i^2 p_j^2 + q_i^2 q_j^2) \\ &= 10[p^\intercal (pp^\intercal) p + q^\intercal (qq^\intercal) q] \\ &= 10[(p^\intercal p)(p^\intercal p) + (q^\intercal q)(q^\intercal q)] \\ &= 10\|p\|_2^4 + 10\|q\|_2^4 \\ &\leq 20b^2 \leq 20b.\end{aligned}$$

### 3.5 Completing the Proof

$$\begin{aligned}\mathrm{Var}[A] &= \sum_{i=1}^n \mathrm{Var}[A_i] + \sum_{i \neq j} \mathrm{Cov}(A_i, A_j) \\ &\leq \sum_{i=1}^n 80m^2 \left( \frac{p_i + q_i}{2} \right)^2 + 4m^3 (p_i - q_i)^2 (p_i + q_i) \\ &\quad + 20m^2 b + 4m^3 \sum_i (p_i - q_i)^2 (p_i + q_i) \\ &\leq 100m^2 b + 8m^3 \sum_i (p_i - q_i)(p_i^2 - q_i^2) \; .\end{aligned}$$

□

## References


[ADJ+12] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh. Competitive classification and closeness testing. In *COLT*, 2012.

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[Can15] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

[CDGR16] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS*, pages 25:1–25:14, 2016.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing $k$-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[DK16] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. *CoRR*, abs/1601.05557, 2016. In FOCS'16.

[DKN15a] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015.

[DKN15b] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, 2015.

[Gol16a] O. Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:15, 2016.

[Gol16b] O. Goldreich. Lecture Notes on Property Testing of Distributions. Available at http://www.wisdom.weizmann.ac.il/ oded/PDF/pt-dist.pdf, March, 2016.

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

[LR05] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

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

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.




ECCC ISSN 1433-8092

http://eccc.hpi-web.de