On the optimal analysis of the collision probability tester (an exposition)

Oded Goldreich

December 5, 2019

Abstract

The collision probability tester, introduced by Goldreich and Ron (*ECCC*, TR00-020, 2000), distinguishes the uniform distribution over $[n]$ from any distribution that is ϵ-far from this distribution using $\text{poly}(1/\epsilon) \cdot \sqrt{n}$ samples. While the original analysis established only an upper bound of $O(1/\epsilon)^4 \cdot \sqrt{n}$ on the sample complexity, a recent analysis of Diakonikolas, Gouleakis, Peebles, and Price (*ECCC*, TR16-178, 2016) established the optimal upper bound of $O(1/\epsilon)^2 \cdot \sqrt{n}$. In this note we survey their analysis, while highlighting the sources of improvement. Specifically:

1. While the original analysis reduces the testing problem to approximating the collision probability of the unknown distribution up to a $1+\epsilon^2$ factor, the improved analysis capitalizes on the fact that the latter problem needs only be solved "at the extreme" (i.e., it suffices to distinguish the uniform distribution, which has collision probability $1/n$, from any distribution that has collision probability exceeding $(1+4\epsilon^2)/n$).
2. While the original analysis provides an almost optimal analysis of the variance of the estimator when $\epsilon = \Omega(1)$, a more careful analysis yields a significantly better bound for the case of $\epsilon = o(1)$, which is the case that is relevant here.

A preliminary version of this exposition was posted in September 2017 as Comment Nr. 1 on TR16-178 of *ECCC*. The current revision is quite minimal, although some typos were fixed and some of the discussions were improved.

1 Introduction

We consider the task of testing whether an unknown distribution X, which ranges over $[n]$, equals the uniform distribution over $[n]$, denoted U_n. On input n, a proximity parameter $\epsilon > 0$, and $s = s(n, \epsilon)$ samples of a distribution $X \in [n]$, the tester should accept (with probability at least $2/3$) if $X \equiv U_n$ and reject (with probability at least $2/3$) if the statistical distance between X and U_n exceeds ϵ. (This testing task is a central problem in "distribution testing" (see, e.g., [9, Chap. 11]), which in turn is part of property testing [9].)[1]

[1] Although testing properties of distributions was briefly discussed in [10, Sec. 3.4.3], its study was effectively initiated in [4]. The starting point of [4] was a test of uniformity, which was implicit in [11], where it is applied to test the distribution of the endpoint of a relatively short random walk on a bounded-degree graph. Generalizing this tester of uniformity, Batu *et al.* [4, 3] presented testers for the property consisting of pairs of identical distributions as well as for all properties consisting of any single distribution (where the property $\{U_n\}$ is a special case).

The collision probability tester [11] is such a tester. It operates by counting the number of (pairwise) collisions between the s samples that it is given, and accepts if and only if the count exceeds $\frac{1+2\epsilon^2}{n} \cdot \binom{s}{2}$. Specifically, this tester estimates the collision probability of X, and accepts if and only if the estimate exceeds $\frac{1+2\epsilon^2}{n}$. An estimate that is at distance at most $2\epsilon^2/n$ from the correct value (with probability at least 2/3) suffices, since the collision probability of U_n equals $1/n$, whereas the collision probability of any distribution that is ϵ-far from U_n must exceed $\frac{1+4\epsilon^2}{n}$.

The initial analysis of this tester, presented in [11], showed that the collision probability of X can be estimated to within a deviation of $\eta > 0$ using $O(\sqrt{n}/\eta^2)$ samples. This yields a tester with sample complexity $O(\sqrt{n}/\epsilon^4)$, where $\epsilon > 0$ is the proximity parameter. Subsequently, it was shown that closely related testers use $O(\sqrt{n}/\epsilon^2)$ samples, and that this upper bound is optimal [13].[2] The fact that $O(\sqrt{n}/\epsilon^2)$ samples actually suffice for the collision probability tester was recently established by Diakonikolas *et al.* [8], and the current note surveys their proof.

The analysis of Diakonikolas *et al.* [8] is based on (1) observing that approximating the collision probability is easier when its value is extremely small, and (2) providing a more tight analysis of the variance of the (empirical) count (i.e., number of collision). The "take home messages" correspond to these two steps: Firstly, one should bear in mind (the well-known fact) that, in many settings, approximating a value is easier when the value is at an extreme (e.g., it is easier to distinguish the cases $\mathbf{Pr}[Y=1] = 1$ and $\mathbf{Pr}[Y=1] = 1-\epsilon$ than to distinguish the cases $\mathbf{Pr}[Y=1] = 0.5$ and $\mathbf{Pr}[Y=1] = 0.5-\epsilon$). Secondly, it often pays to obtain a tighter analysis. Furthermore, a bound that is essentially optimal in general may be sub-optimal in extreme cases, which may actually be the cases we care about. (Indeed, this is exactly what happens in the current setting.)

To illustrate and motivate the analysis recall that the s samples of X yield $m = \binom{s}{2}$ votes regarding the collision probability of X, where each vote correspond to a pair of samples. That is, the $(j,k)^{\text{th}}$ vote it 1 if and only if the j^{th} sample yields the same value as the k^{th} sample. Clearly, the expected value of each vote equals the collision probability of X, and having $m = O(n/\eta^2)$ pairwise independent votes would have sufficed for approximating the collision probability of X up to a multiplicative factor of $1+\eta$, which would have allowed using $s = O(\sqrt{m}) = O(\sqrt{n}/\eta)$ samples. The problem is that, in general, the votes are not pairwise independent (i.e., the $(j,k)^{\text{th}}$ vote is not independent of the $(k,\ell)^{\text{th}}$ vote), and this fact increases the varaince of the count (i.e., number of collision) and leads to the weaker bound of [11]. However, when $X \equiv U_n$, the votes are pairwise independent (e.g., the value of the $(j,k)^{\text{th}}$ vote does not condition the k^{th} sample, and so the value of the $(k,\ell)^{\text{th}}$ vote is statistically independent of the former value). Furthermore, in general, the variance of the count can be upper-bounded by $I+E$, where I represents the value in the ideal case in which the votes are pairwise independent and E is an error term that depends on the difference between the collision probability of X and $1/n$. It turns out that the dependence of E on the latter difference is good enough to yield the desired result (see Section 3).

2 Preliminaries (partially reproduced from [9])

The collision probability of a distribution X is the probability that two samples drawn according to X are equal; that is, the collision probability of X is $\mathbf{Pr}_{i,j\sim X}[i=j]$, which equals $\sum_{i\in[n]} \mathbf{Pr}[X=i]^2$.

[2] Alternative proofs of these bounds can be found in [5] (see also [7, Apdx.]) and [6, Sec. 3.1.1], respectively.

For example, the collision probability of U_n is $1/n$. Letting $p(i) = \mathbf{Pr}[X = i]$, observe that

$$\sum_{i\in[n]} p(i)^2 = \frac{1}{n} + \sum_{i\in[n]} \left(p(i) - n^{-1}\right)^2, \tag{1}$$

which means that the collision probability of X equals the sum of the collision probability of U_n and the square of the $\mathcal{L}_2$-norm of $X - U_n$ (viewed as a vector, i.e., $\|X - U_n\|_2^2 = \sum_{i\in[n]} |p(i) - u(i)|^2$, where $u(i) = \mathbf{Pr}[U_n = i] = 1/n$).

The key observation is that, while the collision probability of U_n equals $1/n$, *the collision probability of any distribution that is ϵ-far from U_n is greater than $\frac{1}{n} + \frac{4\epsilon^2}{n}$*. To see the latter claim, let p denote the corresponding probability function, and note that if $\sum_{i\in[n]} |p(i) - n^{-1}| > 2\epsilon$, then

$$\begin{aligned} \sum_{i\in[n]} \left(p(i) - n^{-1}\right)^2 &\geq \frac{1}{n} \cdot \left(\sum_{i\in[n]} \left|p(i) - n^{-1}\right|\right)^2 \\ &> \frac{(2\epsilon)^2}{n} \end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality.[3] Indeed, using Eq. (1), we get $\sum_{i\in[n]} p(i)^2 > \frac{1}{n} + \frac{(2\epsilon)^2}{n}$. Hence, *testing whether an unknown distribution $X \in [n]$ equals U_n or is ϵ-far from U_n reduces to distinguishing the case that the collision probability of X equals $1/n$ from the case that the collision probability of X exceeds $\frac{1}{n} + \frac{4\epsilon^2}{n}$.*

In light of the above, we focus on approximating the collision probability of the unknown distribution X. This yields the following test, where the sample size, denoted s, is intentionally left as a free parameter.

Algorithm 1 (the collision probability tester): *On input $(n, \epsilon; i_1, ..., i_s)$, where $i_1, ..., i_s$ are drawn from a distribution X, compute $c \leftarrow |\{j < k : i_j = i_k\}|$, and accept if and only if $\frac{c}{\binom{s}{2}} \leq \frac{1+2\epsilon^2}{n}$. We call c the* empirical collision count.

Algorithm 1 approximates the collision probability of the distribution X from which the sample is drawn, and the issue at hand is the quality of this approximation (as a function of s, or rather how to set s so to obtain good approximation). The key observation is that each pair of sample points provides an unbiased estimator[4] of the collision probability (i.e., for every $j < k$ it holds that $\mathbf{Pr}_{i_j, i_k \sim X}[i_j = i_k] = \sum_{i\in[n]} \mathbf{Pr}[X = i]^2$), and that these $\binom{s}{2}$ pairs are "almost pairwise independent".

Recalling that the collision probability of $X \in [n]$ is at least $1/n$, it follows that a sample of size $O(\sqrt{n})$ (which "spans" $O(n)$ pairs) provides a "good approximation" of the collision probability

[3] That is, use $\sum_{i\in[n]} |p(i) - n^{-1}| \cdot 1 \leq \left(\sum_{i\in[n]} |p(i) - n^{-1}|^2\right)^{1/2} \cdot \left(\sum_{i\in[n]} 1^2\right)^{1/2}$.

[4] A random variable X (resp., an algorithm) is called an unbiased estimator of a quantity v if $\mathbb{E}[X] = v$ (resp., the expected value of its output equals v). Needless to say, the key question with respect to the usefulness of such an estimator is the magnitude of its variance (and, specifically, the relation between its variance and the square of its expectation). For example, for any NP-witness relation $R \subseteq \bigcup_{n\in\mathbb{N}}(\{0,1\}^n \times \{0,1\}^{p(n)})$, the (trivial) algorithm that on input x selects at random $y \in \{0,1\}^{p(|x|)}$ and outputs $2^{p(|x|)}$ if and only if $(x, y) \in R$, is an unbiased estimator of the number of witnesses for x, whereas counting the number of NP-witnesses is notoriously hard. The catch is, of course, that this estimation has a huge variance; letting $\rho(x) > 0$ denote the fraction of witnesses for x, this estimator has expected value $\rho(x) \cdot 2^{p(|x|)}$ whereas its variance is $(\rho(x) - \rho(x)^2) \cdot 2^{2 \cdot p(|x|)}$, which is typically much larger than the expectation squared (i.e., when $0 < \rho(x) \ll 1/\text{poly}(|x|)$).

of X in the sence that, with probability at least 2/3, the value of $c/\binom{s}{2}$ approxinates the collision probability up to a multiplicative factor of 1.01. Furthermore, using $s = O(\eta^{-2}\sqrt{n})$ samples suffice for approxinating the collision probability up to a factor of $1+\eta$. Recalling that testing requires approxinating the collision probability up to a factor of $1+\epsilon^2$. this yield an upper bound of $O(\epsilon^{-4}\sqrt{n})$ on the number of samples.

The better analysis presented next (in Section 3) capitalizes on the fact that we do not need to approximate the collision probability of any distribution up to a factor of $1+\eta$, but rather to distinguish the case that the collision probability equals $1/n$ from the case that the collision probability exceeds $\frac{1}{n} + \frac{2\eta}{n}$. In addition, it uses a more refined analysis of the variance of the count c computed by Algorithm 1, which is presented in Lemma 2.

3 The actual analysis

The core of the analysis is captured by the following lemma, where μ denotes the collision probability of X, and δ its deviation from the collision probability of U_n (i.e., $\delta = \mu - \frac{1}{n}$). The upper bound on the variance of the empirical collision count provided by this lemma improves over the standard upper bound of $O(s^3\mu^{3/2})$, where the improvement is in the dominant term of $O(s^3\delta^{3/2})$. This improvement is significant in case $\delta = o(\mu)$, which is the case that we are interested in.

Lemma 2 (the variance of the collision counter): *Let μ denote the collision probability of X, and let Z denote the empirical collision count; that is, $Z = |\{1 \leq j < k \leq s : i_j = i_k\}|$, where $i_1, ..., i_s$ are drawn from a distribution X. Then, $\mathbb{E}[Z] = \binom{s}{2} \cdot \mu$ and $\mathbb{V}[Z] = O(s^2 \cdot \mu) + O(s^3) \cdot (\delta^{3/2} + \frac{\delta}{n})$, where $\delta = \mu - \frac{1}{n}$.*

The standard upper bound is $\mathbb{V}[Z] = O(s^3\mu^{3/2})$, and it can be obtained by degenerating the refined analysis presented below (as indicated in a couple of notes). Evidently, Lemma 2 implies the standard upper bound: The key point is that $\delta < \mu$ implies that $s^3 \cdot \delta^{3/2} < s^3\mu^{3/2}$, whereas $s^2 \cdot \mu + s^3 \cdot (\delta/n) = O(s^3) \cdot \delta^{3/2}$ holds since $\mu/n < \mu^{3/2}$ (equiv., $\mu^{1/2} > 1/n$)) and $s = \Omega(1/\sqrt{\mu})$. Note that the tighter bound (of Lemma 2) coincides with the standard one when $\delta = \Omega(\mu)$, but we are interested in smaller δ (i.e., $\delta \ll \mu$). For example, when $\delta = 0$ (i.e., $X \equiv U_n$), we get an upper bound asserting $\mathbb{V}[Z] = O(s^2 \cdot \mu)$, which is much better than $\mathbb{V}[Z] = O(s^3 \cdot \mu^{3/2}) = O(s^2 \cdot \mu)^{3/2}$ (assuming $s = \omega(1/\sqrt{\mu})$).

Proof: As noted before, each pair of samples provides an unbiased estimator of μ, and so $\mathbb{E}[Z] = \binom{s}{2} \cdot \mu$. If these pairs of samples were pairwise independent, then $\mathbb{V}[Z] = \binom{s}{2} \cdot (\mu - \mu^2)$ would have followed. But the pairs are not pairwise independent, although they are close to being so in the sense that almost all pairs of samples (i.e., quadruples of samples) are independent (i.e., (i_j, i_k) and $(i_{j'}, i_{k'})$ are independent if $|\{j, k, j', k'\}| = 4$). Hence, the desired bound is obtained by carefully examining the contribution of pairs of samples that are independent and the contribution of pairs of samples that are (potentially) dependent.

Specifically, we consider $m = \binom{s}{2}$ random variables $\zeta_{j,k}$ that represent the possible collision events; that is, for $j, k \in [s]$ such that $j < k$, let $\zeta_{j,k} = 1$ if the j^{th} sample collides with the k^{th} sample (i.e., $i_j = i_k$) and $\zeta_{j,k} = 0$ otherwise. Then, $\mathbb{E}[\zeta_{j,k}] = \sum_{i \in [n]} \mathbf{Pr}[i_j = i_k = i] = \mu$ and

$\mathbb{V}[\zeta_{j,k}] = \mathbb{E}[\zeta_{j,k}^2] - \mu^2 = \mu - \mu^2$. Letting $\overline{\zeta}_{i,j} \overset{\text{def}}{=} \zeta_{i,j} - \mu$ (and using $\mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$), we get:

$$\begin{aligned} \mathbb{V}[Z] &= \mathbb{E}\left[\left(\sum_{j<k} \overline{\zeta}_{j,k}\right)^2\right] \\ &= \sum_{j_1<k_1, j_2<k_2} \mathbb{E}\left[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}\right]. \end{aligned}$$

We partition the terms in the last sum according to the number of distinct indices that occur in them such that, for $t \in \{2,3,4\}$, we let $(j_1,k_1,j_2,k_2) \in S_t \subseteq [s]^4$ if and only if $|\{j_1,k_1,j_2,k_2\}| = t$ (and $j_1 < k_1 \wedge j_2 < k_2$). Hence,

$$\mathbb{V}[Z] = \sum_{t\in\{2,3,4\}} \sum_{(j_1,k_1,j_2,k_2)\in S_t} \mathbb{E}\left[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}\right]. \tag{2}$$

The contribution of each quadruple in S_4 to the sum is zero, since the four samples are independent and so $\mathbb{E}[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}] = \mathbb{E}[\overline{\zeta}_{j_1,k_1}] \cdot \mathbb{E}[\overline{\zeta}_{j_2,k_2}] = 0$. Each quadruple in S_2 (which necessarily satisfies $(j_1,k_1) = (j_2,k_2)$) contributes $\mathbb{E}[\overline{\zeta}_{j_1,k_1}^2] = \mathbb{V}[\zeta_{j_1,k_1}] \le \mu$ to the sum, and there are exactly m such quadruples, so their total contribution is at most $m \cdot \mu$. Turning to S_3, we note that each of its $\Theta(s^3)$ quadruples contributes

$$\begin{aligned} \mathbb{E}[\overline{\zeta}_{1,2}\overline{\zeta}_{2,3}] &= \mathbb{E}[\zeta_{1,2}\zeta_{2,3}] - \mathbb{E}[\zeta_{1,2}] \cdot \mathbb{E}[\zeta_{2,3}] \\ &= \sum_{i\in[n]} \mathbf{Pr}[X=i]^3 \ - \mu^2. \end{aligned}$$

Letting $\tau = \sum_{i\in[n]} \mathbf{Pr}[X=i]^3$ denote the three-way collision probability of X, the total contribution of the quadruples of S_3 is $\Theta(s^3) \cdot (\tau - \mu^2)$. Plugging all of this into Eq. (2), we get

$$\mathbb{V}[Z] = \Theta(s^2) \cdot \mu + \Theta(s^3) \cdot (\tau - \mu^2). \tag{3}$$

(*The standard bound of* $\mathbb{V}[Z] = O(s^3\mu^{3/2})$ *is obtained by giving-up on the* μ^2 *term and using* $\tau \le \mu^{3/2}$, *while assuming* $s = \Omega(1/\sqrt{\mu})$; specifically, note that $\tau = \sum_{i\in[n]} \mathbf{Pr}[X=i]^3$ is upper-bounded by $\max_{i\in[n]}\{\mathbf{Pr}[X=i]\} \cdot \sum_{i\in[n]} \mathbf{Pr}[X=i]^2 \le \sqrt{\mu} \cdot \mu$.)[5]

Letting $p_i \overset{\text{def}}{=} \mathbf{Pr}[X=i]$, we upper-bound $\mathbb{V}[Z] = \Theta(s^2) \cdot \mu + \Theta(s^3) \cdot (\tau - \mu^2)$ by upper-bounding τ as follows:

$$\begin{aligned} \tau &= \sum_{i\in[n]} p_i^3 \\ &= \sum_{i\in[n]} \left(\left(p_i - \frac{1}{n}\right) + \frac{1}{n}\right)^3 \\ &= \sum_{i\in[n]} \left(p_i - \frac{1}{n}\right)^3 + \frac{3}{n} \cdot \sum_{i\in[n]} \left(p_i - \frac{1}{n}\right)^2 + \frac{3}{n^2} \cdot \sum_{i\in[n]} \left(p_i - \frac{1}{n}\right) + \frac{n}{n^3} \end{aligned}$$

[5] In fact, one typically derives the standrad bound earlier by using $\mathbb{E}[\overline{\zeta}_{1,2}\overline{\zeta}_{2,3}] \le \mathbb{E}[\zeta_{1,2}\zeta_{2,3}] = \tau$ (instead of $\mathbb{E}[\overline{\zeta}_{1,2}\overline{\zeta}_{2,3}] = \tau - \mu^2$), and noting that $\tau \le \mu^{3/2}$.

$$
\begin{aligned}
&\le \left(\sum_{i\in[n]}\left(p_i-\frac{1}{n}\right)^2\right)^{3/2}+\frac{3}{n}\cdot\delta+0+\mu^2\\
&= \delta^{3/2}+3\cdot(\delta/n)+\mu^2
\end{aligned}
$$

where the inequality uses $\sum_i a_i^3 \le \left(\sum_i a_i^2\right)^{3/2}$ as well as $\sum_{i\in[n]}\left(p_i-\frac{1}{n}\right)^2=\mu-\frac{1}{n}=\delta$ and $\mu\ge 1/n$. Hence, $\mathbb{V}[Z]=\Theta(s^2)\cdot\mu+\Theta(s^3)\cdot(\tau-\mu^2)$ is upper-bounded by $O(s^2\cdot\mu)+O(s^3)\cdot(\delta^{3/2}+(\delta/n))$. ■

Theorem 3 (distinguishing U_n from X of higher collision probability): *For any $\eta\in(0,1]$ and sufficiently large $s=O(\sqrt{n}/\eta)$, the following holds.*

1. *If $X\equiv U_n$, then $\mathbf{Pr}\left[\frac{Z}{\binom{s}{2}}>\frac{1+\eta}{n}\right]\;<\;1/3$.*

2. *If the collision probability of X exceeds $\frac{1}{n}+\frac{2\eta}{n}$, then $\mathbf{Pr}\left[\frac{Z}{\binom{s}{2}}\le\frac{1+\eta}{n}\right]\;<\;1/3$.*

where Z is as in Lemma 2.

Hence, for sufficiently large $s=O(\sqrt{n}/\eta)$, with probability at least $2/3$, the empirical collision count distinguishes U_n from X having collision probability exceeding $(1+2\eta)/n$. It follows that, for $s=O(\sqrt{n}/\epsilon^2)$, with probability at least $2/3$, Algorithm 1 distinguishes U_n from any distribution that is ϵ-far from U_n.

Proof: Combining Chebyshev's Inequality with Lemma 2 (while letting $m=\binom{s}{2}$), we get:

$$
\begin{aligned}
\mathbf{Pr}\left[\left|\frac{Z}{m}-\mu\right|>\gamma\right] &< \frac{\mathbb{V}[Z]}{m^2\cdot\gamma^2}\\
&= \frac{O(s^2)\cdot\mu+O(s^3)\cdot(\delta^{3/2}+(\delta/n))}{s^4\gamma^2}
\end{aligned}
$$

where $\mu=\mathbb{E}[Z]/m$ and $\delta=\mu-(1/n)$. In the case of $X\equiv U_n$ (where $\mu=1/n$ and $\delta=0$), we get

$$
\begin{aligned}
\mathbf{Pr}\left[\frac{Z}{m}>(1+\eta)/n\right] &\le \mathbf{Pr}\left[\left|\frac{Z}{m}-\mu\right|>\eta/n\right]\\
&< \frac{O(s^2)\cdot\mu}{s^4\cdot(\eta/n)^2}\\
&= \frac{O(1/n)}{s^2\cdot(\eta/n)^2}\\
&= \frac{O(1)}{s^2\cdot\eta^2/n}
\end{aligned}
$$

which is upper bounded by $1/3$ provided that $s=O(\sqrt{n}/\eta)$ is sufficiently large. Turning to the case that the collision probability of X exceeds $\frac{1}{n}+\frac{2\eta}{n}$ (i.e., $\delta>2\eta/n$), we get

$$
\mathbf{Pr}\left[\frac{Z}{m}\le(1+\eta)/n\right]\;\le\;\mathbf{Pr}\left[\left|\frac{Z}{m}-\left(\frac{1}{n}+\delta\right)\right|>\delta-\frac{\eta}{n}\right]
$$

$$
\begin{aligned}
&\leq \mathbf{Pr}\left[\left|\frac{Z}{m}-\mu\right| > \delta/2\right] \\
&< \frac{O(s^2)\cdot\mu + O(s^3)\cdot(\delta^{3/2}+(\delta/n))}{s^4\cdot\delta^2} \\
&= \left(\frac{O(1/n)}{s^2\cdot\delta^2}+\frac{O(\delta)}{s^2\cdot\delta^2}\right)+\frac{O(1)}{s\cdot\delta^{1/2}}+\frac{O(1)}{s\cdot\delta\cdot n} \\
&= \frac{O(1)}{s^2\cdot\delta^2\cdot n}+\frac{O(1)}{s^2\cdot\delta}+\frac{O(1)}{s\cdot\delta^{1/2}}+\frac{O(1)}{s\cdot\delta\cdot n}
\end{aligned}
$$

which is upper bounded by 1/3 provided that $s = O(\sqrt{n}/\eta)$ is sufficiently large.[6] ■

Comments. The proof of Theorem 3 can be easily adapted to show that *if the collision probability of* X *is at most* $\frac{1}{n}+\frac{\eta}{2n}$*, then* $\mathbf{Pr}[Z > (1+\eta)/n] < 1/3$. This implies that, using $s = O(\sqrt{n}/\eta)$, the empirical collision count distinguishes distributions having collision probability at most $(1+0.5\eta)/n$ from distributions having collision probability exceeding $(1+2\eta)/n$. (Note that this does *not* yield an algorithm that, using $O(\sqrt{n}/\epsilon^2)$ samples, distinguishes distributions that are $0.1\cdot\epsilon$-close to U_n from distributions that are ϵ-far from U_n, since a distribution that is 0.1ϵ-close to U_n may have collision probability greater than $0.1\epsilon = \omega(1/n)$.)[7] We note that the proof of Theorem 3 would remain intact if we replaced the bound of Lemma 2 (i.e., $\mathbb{V}[Z] = O(s^2\cdot\mu)+O(s^3)\cdot(\delta^{3/2}+\frac{\delta}{n})$) by the weaker $\mathbb{V}[Z] = O(s^2\cdot\mu)+O(s^3)\cdot(\delta^{3/2}+\frac{\delta}{\sqrt{n}})$.

Acknowledgements

I am grateful to Ryan O'Donnell for many helpful discussions regarding the result of [8]. Ryan, in turn, claims to have been benefitting from his collaborators on [2, 12], and was also inspired by [1]. Hence, my thanks are extended to all contributors to these works as well as to the contributers to [8].

References

[1] J. Acharya, C. Daskalakis, and G. Kamath. Optimal Testing for Properties of Distributions. `arXiv:1507.05952 [cs.DS]`, 2015.

[2] C. Badescu, R. O'Donnell, and J. Wright. Quantum state certification. `arXiv:1708.06002 [quant-ph]`, 2017.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *42nd FOCS*, pages 442–451, 2001.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that Distributions are Close. In *41st FOCS*, pages 259–269, 2000.

[6] Let $s = c\cdot\sqrt{n}/\eta$ for some constant c. Then, when upper-bounding the first and last terms, use $s^2\cdot\delta^2\cdot n > c^2\cdot(n/\eta^2)\cdot(2\eta/n)^2\cdot n = 4c^2$. When upper-bounding the second and third terms, use $s^2\cdot\delta > c^2\cdot(n/\eta^2)\cdot(2\eta/n) \geq 2c^2$, where the last inequality uses $\eta \leq 1$.

[7] Actually, the foregoing "tolerant testing" task has sample complexity $\Omega(n/\log n)$; see [14].

[5] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *25th ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2014.

[6] I. Diakonikolas and D. Kane. A New Approach for Testing Properties of Discrete Distributions. `arXiv:1601.05557 [cs.DS]`, 2016.

[7] I. Diakonikolas, D. Kane, V. Nikishkin. Testing Identity of Structured Distributions. In *26th ACM-SIAM Symposium on Discrete Algorithms*, pages 1841–1854, 2015.

[8] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based Testers are Optimal for Uniformity and Closeness. *ECCC*, TR16-178, 2016.

[9] O. Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[10] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, pages 653–750, July 1998. Extended abstract in *37th FOCS*, 1996.

[11] O. Goldreich and D. Ron. On Testing Expansion in Bounded-Degree Graphs. *ECCC*, TR00-020, 2000.

[12] R. O'Donnell and J. Wright. A Primer on the Statistics of Longest Increasing Subsequences and Quantum States. To appear in *SIGACT News*.

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

[14] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. *ECCC*, TR10-179, 2010.