Identity Testing Under Label Mismatch

Clément L. Canonne
The University of Sydney, Australia

Karl Wimmer
Duquesne University, Pittsburgh, PA, USA

Abstract

Testing whether the observed data conforms to a purported model (probability distribution) is a basic and fundamental statistical task, and one that is by now well understood. However, the standard formulation, *identity testing*, fails to capture many settings of interest; in this work, we focus on one such natural setting, *identity testing under promise of permutation.* In this setting, the unknown distribution is assumed to be equal to the purported one, up to a relabeling (permutation) of the model: however, due to a systematic error in the reporting of the data, this relabeling may not be the identity. The goal is then to test identity under this assumption: equivalently, whether this systematic labeling error led to a data distribution statistically far from the reference model.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms; Mathematics of computing → Hypothesis testing and confidence interval computation

Keywords and phrases distribution testing, property testing, permutations, lower bounds

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2021.55

Related Version *Full Version*: https://arxiv.org/abs/2105.01856

1 Introduction

Imagine you painstakingly gathered observations, data point after data point, and managed to form an accurate estimate of the data distribution; unfortunately, you did not record the labels correctly, and due to a systematic error the data labels have been permuted in an unknown and arbitrary way. You did make your best educated guess to fix this though, and are confident the data, once carefully relabeled, *should* reflect the reality. Can you check this, without having to go through the whole process of obtaining an entirely new dataset?

In this paper, we are concerned with a variant of identity testing which captures the above scenario, where one is promised that the unknown distribution is equal to the reference distribution $\mathbf{q}$ *up to a permutation of the domain.* Formally, the algorithm has access to i.i.d. samples from a probability distribution $\mathbf{p}$ over a finite domain $[n] := \{1, 2, \ldots, n\}$ such that $\mathbf{p} \circ \pi = \mathbf{q}$ for some (unknown) $\pi \in \mathcal{S}_n$, and, on input $0 \leq \varepsilon' < \varepsilon \leq 1$, must output yes or no such that

- if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon'$, then the algorithm outputs yes with probability at least $2/3$;
- if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, then the algorithm outputs no with probability at least $2/3$.

When $\varepsilon' = 0$, the task is termed *identity testing (under promise of permutation)*; otherwise, it is *tolerant identity testing.* It is worth noting that this permutation promise fundamentally changes the problem, and makes it incomparable to the standard identity testing problem. As an illustrative example, it is known that *uniformity* testing, where the reference distribution $\mathbf{q}$ is uniform over $[n]$, is the "hardest" case of identity testing, with sample complexity $\Theta(\sqrt{n})$ and $\Theta(n/\log n)$ for the testing and tolerant testing versions, respectively [18, 22, 23, 13]. However, it is easy to see that under the permutation promise, uniformity testing is a trivial problem which can be solved with *zero* samples: any permutation of the uniform distribution is itself the uniform distribution.

Our results further demonstrate this stark difference, showing how the difficulty of testing and tolerant testing differ under this promise. In particular, we show an exponential gap between the sample complexities of non-tolerant and tolerant identity testing under this promise: to the best of our knowledge, this constitutes the first example of such a gap between the tolerant and non-tolerant version of a problem in distribution testing.

1.1 Our results

Our results show that, quite surprinsingly, the promise of equality up to permutation of the domain fundamentally changes the sample complexity landscape, and is both qualitatively and quantitatively different from what one could expect from the known bounds on identity and tolerant identity testing without this promise.

Our first set of results indeed establishes that, in contrast to the $\Theta(\sqrt{n})$ sample complexity of "regular" identity testing, identity testing under promise of permutation has sample complexity merely *polylogarithmic* in the domain size:

▶ **Theorem 1** (Theorems 5 and 8, (Informal)). *Identity testing under promise of permutation has sample complexity* $\Theta\big(\log^2 n\big)$*, where* n *is the domain size.*

Given the fact that (regular) tolerant identity testing has sample complexity nearly quadratically higher than (regular) identity testing, one could conjecture that the sample complexity tolerant testing under our promise remains polylogarithmic. Our next set of results shows that this is far from being the case: instead, allowing for some noise tolerance makes the promise of equality up to permutation essentially useless, as the sample complexity blows up *exponentially*, growing from poylogarithmic to nearly linear in the domain size:

▶ **Theorem 2** (Theorems 8 and 9, (Informal)). *Tolerant identity testing under promise of permutation has sample complexity* $\Theta\big(n^{1-o(1)}\big)$*, where* n *is the domain size.*

We also show that relaxing the tolerance allowed from additive (as in the usual tolerant testing setting) to multiplicative in the distance parameter does not really help, as the sample complexity still remains polynomial:

▶ **Theorem 3** (Theorem 17, (Informal)). *Multiplicative-factor tolerant identity testing under promise of permutation, where one needs to distinguish between* ε*-close and* $C\varepsilon$*-far, has sample complexity* $\Omega(\sqrt{n})$ *for any constant factor* $C > 1$*, where* n *is the domain size.*

We emphasize once more that those results, and in particular the lower bounds, do not follow from the known results on standard identity testing, as the promise of equality up to permutation, by strenghtening the premise, drastically changes the problem. In particular, the case where the reference $\mathbf{q}$ is uniform, while known to be the hardest case for identity and tolerant identity testing, is actually a trivially easy case under our promise (as any distribution promised to be a permutation of the uniform distribution is, of course, the uniform distribution itself.)

1.2 Previous work

Distribution testing has a long history in Statistics, that one can trace back to the work of Pearson [12]. More recently, from the computer science perspective, Goldreich, Goldwasser, and Ron initiated the field of property testing [14]; of which distribution testing emerged through the seminal work of Batu, Fortnow, Rubinfeld, Smith, and White [2]. We refer the read to the survey [5] for a review of the area of distribution testing.

Among the problems tackled in this field, *identity testing* (also known as goodness-of-fit or one-sample testing), in which the goal is to decide whether an unknown probability distribution $\mathbf{p}$ is equal to a purported model $\mathbf{q}$, has received significant attention. It is known that for identity testing with any reference distribution $\mathbf{q}$ over a domain of size n, $\Theta(\sqrt{n})$ samples are necessary and sufficient [18, 7, 1, 23]; moreover, the exact asymptotic dependence on the distance parameter and the probability of error of the test [15, 9], as well as some good understanding of the dependence on the reference distribution $\mathbf{q}$ itself [23, 4], are now understood. Further, we also have tight bounds for the harder problem where one seeks to allow for some noise in the data (i.e., perform *tolerant* identity testing, where the algorithm has to accept distributions sufficient close to the reference $\mathbf{q}$): $\Theta(n/\log n)$ samples, a nearly linear dependence on the domain size, are known to be necessary and sufficient [20, 21, 22, 16].

However, how the identity testing problem changes under natural constraints on the input data, or under some variations of the formulation, remains largely unexplored. Among the works concerned with such problems, [3, 8] consider identity testing under monotonicity or k-modality constraints; and [10] focuses on a broad class of shape constraints on the density. Finally, [6] focuses on a variant of identity testing, "identity up to binning," where two distributions are considered equal if some binning of the domain can make them coincide. To the best of our knowledge, the question considered in the present work, albeit arguably quite natural, has not been previously considered in the Statistics or distribution testing literature.

Organization. We provide in Section 3.1 our algorithm for testing identity under promise of permutation, before complementing it in Section 3.2 by our matching lower bound. Section 4 is then concerned with the upper and lower bounds for the tolerant version of the problem; the bulk of which lies in proving the two lower bounds.

2 Preliminaries

Let $\mathcal{S}_n$ denote the set of permutations of $[n] := \{1, 2, \ldots, n\}$. We identify a probability distribution $\mathbf{p}$ over $[n]$ with its probability mass function (pmf), that is, a function $\mathbf{p}\colon [n] \to [0,1]$ such that $\sum_{i=1}^{n} \mathbf{p}(i) = 1$. For a subset $S \subseteq [n]$, we then write $\mathbf{p}(S) = \sum_{i \in S} \mathbf{p}(i)$ for the probability mass assigned to S by $\mathbf{p}$. Given two probability distributions $\mathbf{p}, \mathbf{q}$ over $[n]$, their total variation distance is

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq [n]} (\mathbf{p}(S) - \mathbf{q}(S)) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 \tag{1}$$

where $\|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |\mathbf{p}(i) - \mathbf{q}(i)|$ is the ℓ_1 distance between the two pmfs. In what follows, given a probability distribution $\mathbf{q}$ over $[n]$, we define

$$\Pi_n(\mathbf{q}) := \{\, \mathbf{q} \circ \pi : \pi \in \mathcal{S}_n \,\}, \tag{2}$$

the set of distributions equal to $\mathbf{q}$ up to permutation of the domain.

Finally, we will rely on the so-called DKW inequality, which roughly states that $O(1/\varepsilon^2)$ samples from any univariate distribution suffice to learn it to Kolmogorov distance ε with high probability: this is a result due to Dvoretzky, Kiefer, and Wolfowitz from 1956 [11] (with the optimal constant due to Massart, in 1990 [17]).

▶ **Theorem 4** (DKW Inequality). *Let $\hat{\mathbf{p}}$ denote the empirical distribution on m i.i.d. samples from an* arbitrary *distribution $\mathbf{p}$ on $\mathbb{R}$. Then, for every $\varepsilon > 0$,*

$$\Pr[\, \mathrm{d}_{\mathrm{K}}(\hat{\mathbf{p}}, \mathbf{p}) > \varepsilon \,] \leq 2e^{-2m\varepsilon^2},$$

where, for two univariate distributions $\mathbf{p}, \mathbf{q}$, $\mathrm{d}_{\mathrm{K}}(\mathbf{p}, \mathbf{q}) = \sup_{x \in \mathbb{R}} |\mathbf{p}((-\infty, x]) - \mathbf{q}((-\infty, x])|$ denotes the Kolmogorov distance between $\mathbf{p}$ and $\mathbf{q}$.

3 Testing

In this section, we establish our matching upper and lower bounds for testing under promise of permutation, Theorems 8 and 9.

3.1 Upper bound

We begin by proving our $O(\log^2 n)$ upper bound for identity testing under promise of permutation.

▶ **Theorem 5.** *There exists an algorithm (Algorithm 1) which, for any reference distribution* $\mathbf{q}$ *over* $[n]$ *and any* $0 < \varepsilon \leq 1$*, given* $O\left(\frac{\log^2 n}{\varepsilon^4}\right)$ *samples from an unknown distribution* $\mathbf{p} \in \Pi_n(\mathbf{q})$*, distinguishes with probability at least* $2/3$ *between (i)* $\mathbf{p} = \mathbf{q}$ *and (ii)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$*.*

Proof. We first partition the domain into $L := O(\log(n/\varepsilon)/\varepsilon)$ *buckets* $B_1, \dots, B_L$, where

$$B_\ell := \left\{ i \in [n] : \frac{1}{(1+\varepsilon/4)^\ell} < \mathbf{q}(i) \leq \frac{1}{(1+\varepsilon/4)^{\ell-1}} \right\}, \qquad 1 \leq \ell \leq L-1 \tag{3}$$

and $B_L := \left\{ i \in [n] : \mathbf{q}(i) \leq \frac{1}{(1+\varepsilon/4)^{L-1}} \right\}$. Note that since $\mathbf{q}$ is known, we can exactly compute the partition $B_1, \dots, B_L$, and in particular those L sets can be efficiently obtained.

Algorithm 1 Algorithm for identity testing under promise of permutation.

Require: Reference distribution $\mathbf{q}$, distance parameter $\varepsilon \in (0, 1]$, sample access to $\mathbf{p} \in \Pi_n(\mathbf{q})$

1: Set $L \leftarrow 1 + \left\lceil \frac{\log(4n/\varepsilon)}{\log(1+\varepsilon/4)} \right\rceil = O\left(\frac{\log(n/\varepsilon)}{\varepsilon}\right)$, $\delta \leftarrow \frac{\varepsilon}{4(L-1)}$
2: Compute the bucketing $B_1, \dots, B_L$, as in (3)
3: Using $O(1/\delta^2)$ samples from $\mathbf{p}$, use the empirical estimator to learn the distribution

$$\bar{\mathbf{p}} := (\mathbf{p}(B_1), \dots, \mathbf{p}(B_L))$$

over $[L]$ to Kolmogorov distance $\frac{\delta}{3}$, with probability of error $1/10$. Let $\hat{\mathbf{p}}$ be the output.
4: **if** $\hat{\mathbf{p}}(B_L) > \frac{3\varepsilon}{8}$ or there exists ℓ^* such that $|\hat{\mathbf{p}}(\{\ell^*, \dots, L-1\}) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)| > \frac{\delta}{3}$ **then**
5: **return** no
6: **else**
7: **return** yes
8: **end if**

For our choice of L, $\frac{1}{(1+\varepsilon/4)^{L-1}} \leq \frac{\varepsilon}{4n}$, so the last bucket B_L has small probability mass under the reference distribution: $\mathbf{q}(B_L) \leq \frac{\varepsilon}{4}$. Now, distinguishing with high constant probability between $\mathbf{p}(B_L) \leq \frac{\varepsilon}{4}$ and $\mathbf{p}(B_L) \geq \frac{\varepsilon}{2}$ can be done with $O(1/\varepsilon)$ samples, so we can detect a discrepancy in B_L with high probability if there is one (we will argue this part formally at the end of the proof). Consequently, we hereafter assume that $\mathbf{p}(B_L) < \frac{\varepsilon}{2}$.

If $\mathbf{p} = \mathbf{q}$, clearly $\mathbf{p}(B_L) \leq \frac{\varepsilon}{4}$ (so the first check above passes) and $\mathbf{q}(B_\ell) = \mathbf{p}(B_\ell)$ for all $1 \leq \ell \leq L-1$. However, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, then $\sum_{\ell=1}^{L-1} \sum_{i \in B_\ell} |\mathbf{p}(i) - \mathbf{q}(i)| > 2\varepsilon - \frac{3\varepsilon}{4} = \frac{5}{4}\varepsilon$. Moreover, letting $\pi \in \mathcal{S}_n$ be the permutation such that $\mathbf{p} = \mathbf{q} \circ \pi$, consider the set $S \subseteq [n]$ of elements which π maps to an element from the same bucket:

$$S := \{ i \in [n] \setminus B_L : \exists \ell \in [L-1], i \in B_\ell, \pi(i) \in B_\ell \} .$$

For each such element i, by definition of the bucketing, $|\mathbf{p}(i) - \mathbf{q}(i)| = |\mathbf{q}(i) - \mathbf{q}(\pi(i))| \leq \frac{\varepsilon}{4}\mathbf{q}(i)$. It follows that the elements from S amount for a total ℓ_1 distance of at most $\frac{\varepsilon}{4}$, and therefore a constant fraction of the distance between $\mathbf{p}$ and $\mathbf{q}$ comes from the set $T := [n] \setminus (B_L \cup S)$ of elements that π "moves to a different bucket:"

$$\frac{5}{4}\varepsilon < \sum_{i\in S} |\mathbf{p}(i) - \mathbf{q}(i)| + \sum_{i\in T} |\mathbf{p}(i) - \mathbf{q}(i)| \leq \frac{\varepsilon}{4}\mathbf{q}(S) + \sum_{i\in T} |\mathbf{p}(i) - \mathbf{q}(i)| \leq \frac{\varepsilon}{4} + \sum_{i\in T} |\mathbf{p}(i) - \mathbf{q}(i)|$$

that is, $\sum_{i\in T} |\mathbf{p}(i) - \mathbf{q}(i)| > \varepsilon$.

Partition the set T by setting $T_\ell := T \cap B_\ell$, for $\ell \in [L-1]$. Rewriting the above inequality, we obtained that

$$\sum_{i\in T} |\mathbf{p}(i) - \mathbf{q}(i)| = \sum_{\ell=1}^{L-1} \sum_{i\in T_\ell} |\mathbf{p}(i) - \mathbf{q}(i)| > \varepsilon\,. \tag{4}$$

We will use this to prove the following result.

▷ Claim 6. Suppose that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$. Then there exists some $\ell^* \in [L-1]$ such that $\left|\mathbf{p}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)\right| > \delta$, where $\delta = \frac{\varepsilon}{4(L-1)}$.

Proof. We note that since $\mathbf{p}(B_\ell) = \mathbf{p}(S_\ell) + \mathbf{p}(T_\ell)$ and that $\mathbf{p}(S_\ell) = \mathbf{q}(S_\ell)$ (by definition of $S_\ell \subseteq S$)[1] for every ℓ, it suffices to prove the statement for T_ℓ, that is, that there exists ℓ^* such that

$$\left|\mathbf{p}\Big(\bigcup_{\ell=\ell^*}^{L-1} T_\ell\Big) - \mathbf{q}\Big(\bigcup_{\ell=\ell^*}^{L-1} T_\ell\Big)\right| > \delta\,.$$

The key property we will use is that, for every $\ell < \ell'$, we have $\mathbf{q}(i) \geq \mathbf{q}(j)$ for every $i \in B_\ell, j \in B_{\ell'}$. This property, which follows from the definition of bucketings, guarantees that if π maps an element $i \in T_{\ell'}$ to element $\pi(i) \in T_\ell$, then $\mathbf{p}(i) \geq \mathbf{q}(i)$.

Let $U, V \subseteq [L-1]$ be the buckets whose probability mass under $\mathbf{p}$ is greater than or equal to (resp., less than or equal to) the probability mass under $\mathbf{q}$, i.e.,

$$U := \{\, \ell \in [L-1] :\ \mathbf{p}(T_\ell) \geq \mathbf{q}(T_\ell) \,\}\,, \quad V := \{\, \ell \in [L-1] :\ \mathbf{p}(T_\ell) \leq \mathbf{q}(T_\ell) \,\}$$

This lets us rewrite (4) as

$$\varepsilon < \sum_{\ell\in U} \sum_{i\in T_\ell} |\mathbf{p}(i) - \mathbf{q}(i)| + \sum_{\ell\in V} \sum_{i\in T_\ell} |\mathbf{p}(i) - \mathbf{q}(i)|$$

and so at least one of the two terms in the RHS must exceed $\frac{\varepsilon}{2}$. Without loss of generality, suppose $\sum_{\ell\in U} \sum_{i\in T_\ell} |\mathbf{p}(i) - \mathbf{q}(i)| > \frac{\varepsilon}{2}$. This implies there exists $\ell^* \in U$ such that $\sum_{i\in T_{\ell^*}} |\mathbf{p}(i) - \mathbf{q}(i)| > \frac{\varepsilon}{2(L-1)}$; we will focus on this ℓ^*.

Partition T_{ℓ^*} further into $T^+_{\ell^*}$ and $T^-_{\ell^*}$, where $T^+_{\ell^*}$ (resp. $T^-_{\ell^*}$) is the set of elements $i \in T_{\ell^*}$ such that $\pi(i)$ belongs to a bucket B_ℓ with $\ell < \ell^*$ (resp., $\ell > \ell^*$). Note that, for any $i \in T^+_{\ell^*}$, we then have $\mathbf{p}(i) = \mathbf{q}(\pi(i)) \geq \mathbf{q}(i)$, and conversely for $i \in T^-_{\ell^*}$: so that we can rewrite the above as

$$\frac{\varepsilon}{2(L-1)} < (\mathbf{p}(T^+_{\ell^*}) - \mathbf{q}(T^+_{\ell^*})) + (\mathbf{q}(T^-_{\ell^*}) - \mathbf{p}(T^-_{\ell^*}))$$

[1] Indeed, we have $\mathbf{p}(S_\ell) = \sum_{i\in S_\ell} \mathbf{p}(i) = \sum_{i\in S_\ell} \mathbf{q}(\pi(i)) = \mathbf{q}(\pi^{-1}(S_\ell))$, and $\pi(S_\ell) = S_\ell$ by definition.

Now, since $\mathbf{p}(T_{\ell^*}) \geq \mathbf{q}(T_{\ell^*})$ (as $\ell^* \in U$), we have $\mathbf{p}(T^+_{\ell^*}) - \mathbf{q}(T^+_{\ell^*}) \geq \mathbf{q}(T^-_{\ell^*}) - \mathbf{p}(T^-_{\ell^*})$ and therefore $\mathbf{p}(T^+_{\ell^*}) - \mathbf{q}(T^+_{\ell^*}) > \frac{\varepsilon}{4(L-1)}$. This implies the claim: indeed, we then have

$$\mathbf{p}(\cup_{\ell=\ell^*}^{L-1} T_\ell) > \mathbf{q}(\cup_{\ell=\ell^*}^{L-1} T_\ell) + \frac{\varepsilon}{4(L-1)}$$

since T_{ℓ^*} "receives" a difference of at least $\frac{\varepsilon}{4(L-1)}$ probability mass from lower-index buckets, and besides this the total probability mass of the suffix of buckets $\cup_{\ell=\ell^*}^{L-1} T_\ell$ cannot decrease by any internal swap of elements. ◀

With the above claim in hand, we can conclude the analysis. Indeed, as the sets $B_1, \ldots, B_L$ are known, one can estimate the induced probability distribution $\bar{\mathbf{p}} := (\mathbf{p}(B_1), \ldots, \mathbf{p}(B_L))$ to Kolmogorov distance $\frac{\delta}{3}$ (with probability at least 9/10) using $O(1/\delta^2) = O(L^2/\varepsilon^2) = O(\log^2(n/\varepsilon)/\varepsilon^4)$ samples (this follows from Theorem 4). Let $\hat{\mathbf{p}}$ be the resulting distribution over $[L]$. Whenever this step is successful (i.e., with probability at least 9/10, the following holds.

- If $\mathbf{p} = \mathbf{q}$, then $|\hat{\mathbf{p}}(L) - \mathbf{q}(B_L)| \leq \frac{\delta}{3}$, so $\hat{\mathbf{p}}(B_L) \leq \frac{\varepsilon}{4} + \frac{\delta}{3} \leq \frac{3}{8}\varepsilon$; and $|\hat{\mathbf{p}}(\{\ell^*, \ldots, L-1\}) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)| \leq \frac{\delta}{3}$ for all ℓ. Thus, the test accepts.
- If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, then either
  - $\mathbf{p}(B_L) > \frac{\varepsilon}{2}$, in which case $\hat{\mathbf{p}}(L) \geq \frac{\varepsilon}{2} - \frac{\delta}{3} > \frac{3}{8}\varepsilon$ and the test rejects; or
  - $\mathbf{p}(B_L) \leq \frac{\varepsilon}{2}$, in which case by Claim 6 there exists some $\ell^* \in [L-1]$ such that $\left|\mathbf{p}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)\right| > \delta$. Then,

$$\left|\hat{\mathbf{p}}(\{\ell^*, \ldots, L\}) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)\right| \geq \left|\mathbf{p}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell) - \mathbf{q}(\bigcup_{\ell=\ell^*}^{L-1} B_\ell)\right| - \frac{\delta}{3} > \frac{2}{3}\delta$$

    and the test rejects.

This concludes the proof of correctness of the algorithm. The claimed sample complexity readily follows from our choice of $\delta = \Theta(L/\varepsilon)$ and the $O(1/\delta^2)$ sample complexity of learning an arbitrary real-valued distribution to Kolmogorov distance δ. ◀

▶ Remark 7 (On the tolerance of the tester). We note that the above analysis establishes a slightly stronger statement; namely, that the testing algorithm allows for some small tolerance, accepting distributions that are $O(\varepsilon/\log n)$-close to $\mathbf{q}$, and rejecting those that are ε-far. As we will see later, this $\Omega(\log n)$ factor in the amount of tolerance is essentially optimal, as by Theorem 17 reducing it to $o(\log n)$ would require sample complexity $n^{1/2-o(1)}$.

3.2 Lower bound

In this section, we show that the $O(\log^2 n)$ upper bound from the previous section is tight, by proving a matching lower bound on the sample complexity of identity testing under promise of permutation.

▶ **Theorem 8.** *Any algorithm which, given a reference distribution* $\mathbf{q}$ *over* $[n]$, $0 < \varepsilon \leq 1$ *such that* $\varepsilon = \tilde{\Omega}\big(1/n^{1/4}\big)$, *and sample access to an unknown distribution* $\mathbf{p} \in \Pi_n(\mathbf{q})$, *distinguishes with probability at least* 2/3 *between (i)* $\mathbf{p} = \mathbf{q}$ *and (ii)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, *must have sample complexity* $\Omega\Big(\frac{\log^2 n}{\varepsilon^2}\Big)$.

Proof. We first describe a construction with constant distance $\varepsilon = 1/9$, leading to an $\Omega\big(\log^2 n\big)$ lower bound; before explaining how to obtain the claimed $\Omega\big(\frac{1}{\varepsilon^2}\log^2 n\big)$ lower bound from it. Our lower bound will rely on a reference distribution $\mathbf{q}$ piecewise-constant

on $L = \Theta(\log n)$ buckets, where bucket ℓ has a number of elements proportional to 2^ℓ. The first and last buckets (that is, the smallest and largest) will each have total probability mass $1/3$ under $\mathbf{q}$, and be uniform. The remaining "middle" $L-2$ buckets all have $1/(3(L-2))$ total probability mass, and are uniform as well. We then build a family of perturbations $\{\mathbf{p}_\pi = \mathbf{q} \circ \pi\}_\pi \subseteq \Pi_n(\mathbf{q})$, such that under each perturbation $\mathbf{p}_\pi$ the middle buckets keep the exact same total probability mass $1/(3(L-2))$, by "cascading" mass from one bucket to the next. Details follow.

Set $L := \Theta(\log n)$ to be the largest integer such that $L2^L \leq \sqrt{n}$, and assume for convenience that $\lceil \sqrt{n} \rceil$ is a multiple of 3. The ℓth bucket B_ℓ, for $0 \leq \ell \leq L-2$, has size

$$|B_\ell| = \lceil \sqrt{n} \rceil \cdot 2^\ell$$

and $|B_{L-1}| = 2(L-2)|B_{L-2}|$, so that $\frac{n}{8} \leq \sum_{\ell=0}^{L-1} |B_\ell| = \lceil \sqrt{n} \rceil \cdot 2^{L-1}(L-1) \leq n$. (We hereafter focus on the first part of the domain, and will ignore the last $n - \sum_{\ell=0}^{L-1} |B_\ell|$ elements.) Note that each bucket contains at least $\sqrt{n}$ elements by construction, and has a size which is a multiple of 3. The reference distribution $\mathbf{q}$ is then uniform inside each bucket, where

- $\mathbf{q}(B_0) = \mathbf{q}(B_{L-1}) = \frac{1}{3}$, and
- $\mathbf{q}(B_\ell) = \frac{1}{3(L-2)}$ for all $0 < \ell < L-1$.

In particular, our choice of $|B_{L-1}|$ ensures that each element of the last bucket, under $\mathbf{q}$, will have probability mass

$$\frac{1}{3|B_{L-1}|} = \frac{1}{2} \cdot \frac{1}{3(L-2)|B_{L-2}|}$$

that is, half the probability mass of elements of the $(L-2)$th bucket.

Each perturbation will then have the same distribution over buckets:

- each of the $L-2$ middle buckets B_ℓ is (independently) partitioned uniformly at random into 3 sets $S_{\ell,1}, S_{\ell,2}, S_{\ell,3}$ of equal size. The permutation then swaps $S_{\ell,2} \cup S_{\ell,3}$ and $S_{\ell+1,1}$, for $1 \leq \ell \leq L-3$ (note that indeed $|S_{\ell,2} \cup S_{\ell,3}| = |S_{\ell+1,1}|$, but $\mathbf{q}(S_{\ell,2} \cup S_{\ell,3}) = 2\mathbf{q}(S_{\ell+1,1}) = \frac{2}{9(L-2)}$).
- a uniformly random subset $S_0 \subseteq B_0$ of size $\frac{|B_1|}{3(2L-5)} = O(|S_{1,1}|/L)$ is selected, and the permutation swaps it with a uniformly random subset $T_1 \subseteq S_{1,1}$ of equal size. By choice of the size, we had $\mathbf{q}(S_0) = \frac{2}{9(2L-5)}$ and $\mathbf{q}(T_1) = \frac{1}{9(2L-5)(L-2)}$, so that $\mathbf{q}(S_0) - \mathbf{q}(T_1) = \frac{1}{9(L-2)}$.
- similarly, the subset $S_{L-2,2} \cup S_{L-2,3}$ of size $\frac{2}{3}|B_{L-2}| = \frac{|B_{L-1}|}{3(L-2)}$ is swapped with a uniformly random subset $T_{L-1} \subseteq B_{L-1}$ of equal size. By choice of the size, we had $\mathbf{q}(S_{L-2,2} \cup S_{L-2,3}) = \frac{2}{9(L-2)}$ and $\mathbf{q}(T_{L-1}) = \frac{1}{9(L-2)}$, so that again $\mathbf{q}(S_{L-2,2} \cup S_{L-2,3}) - \mathbf{q}(T_{L-1}) = \frac{1}{9(L-2)}$.

As a result, we get that for each such perturbation $\mathbf{p} = \mathbf{q} \circ \pi$, $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq \frac{1}{9}$. The construction is illustrated in Figure 1.

By a birthday paradox-type argument, no element will be sampled twice unless the number of samples is at least $\Omega(1/\sqrt{\sum_{i=1}^n \mathbf{p}(i)^2}) = \Omega(1/\sqrt{n \max_{i \in [n]} \mathbf{p}(i)}) = \Omega(n^{1/4})$, which is far beyond the polylogarithmic regime we are working in. By construction, under each $\mathbf{p}$, all $L-2$ middle buckets have exactly the same probability mass $\frac{1}{3(L-2)}$, and elements inside are perturbed randomly, either having probability (compared to $\mathbf{q}$) multiplied by 2 with probability $1/3$ or divided by 2 with probability $1/3$. Because of the uniformly random choice of the 3-way partition inside each bucket and the fact that each of all those inner

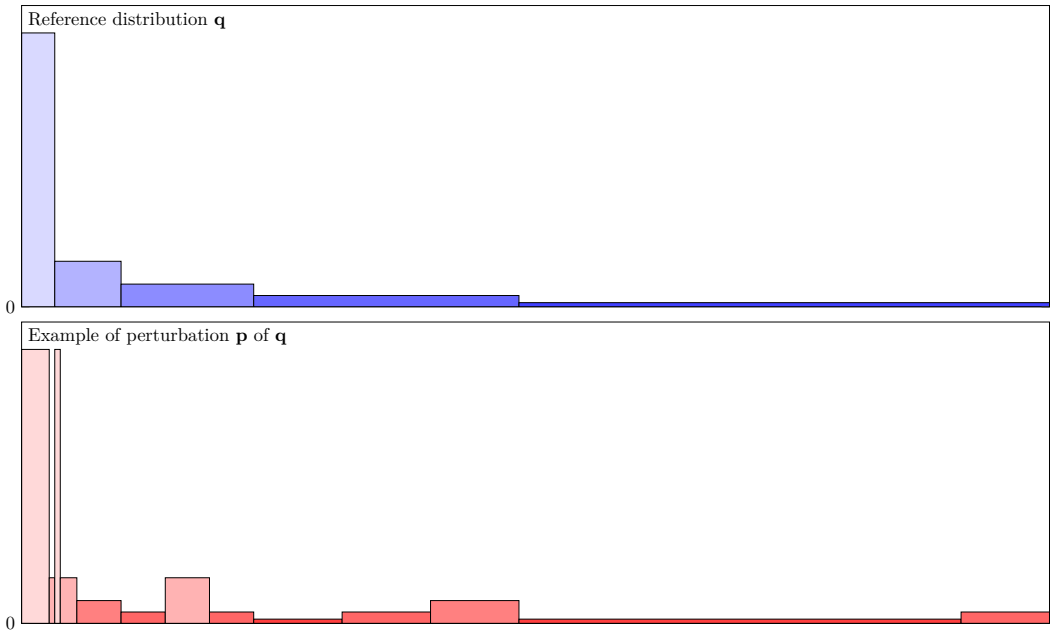

Figure 1 Reference distribution $\mathbf{q}$ and example of perturbation $\mathbf{p}$, for $L = 5$. Note that the total probability mass of each bucket of $\mathbf{q}$ is preserved under $\mathbf{p}$, except for the first and last one whose mass decreases and increases by $\Theta(1/L)$, respectively.

partitions are chosen independently across buckets, the information from those $L-2$ buckets does not provide any advantage in distinguishing them from $\mathbf{p}$ unless the same element is hit twice.[2]

This addresses the case of the middle $L-2$ buckets. Turning to the remaining two, the probability mass of both end buckets, under any perturbation $\mathbf{p}$, deviates from what it is under $\mathbf{q}$ by an additive $\delta := \frac{1}{9(L-2)}$. Since those buckets each have total probability mass $1/3$ under $\mathbf{p}$ and $1/3 \pm \delta$ under each $\mathbf{q}$ and we do not see any collisions with high probability, detecting this requires $\Omega(1/\delta^2) = \Omega(\log^2 n)$ samples, giving the lower bound for constant $\varepsilon = 1/9$.

To obtain the inverse quadratic dependence on the distance parameter, one can then simply repeat the above argument for any $0 < \varepsilon < 1/9$ by replacing our reference distribution $\mathbf{q}$ and all the perturbations $\mathbf{p}_\pi = \mathbf{q} \circ \pi$ by the mixtures

$$\mathbf{q}_\varepsilon := (1-9\varepsilon)\mathbf{u} + 9\varepsilon\mathbf{q}, \qquad \mathbf{p}_{\varepsilon,\pi} := (1-9\varepsilon)\mathbf{u} + 9\varepsilon\mathbf{p}_\pi = \mathbf{q}_\varepsilon \circ \pi$$

the last equality crucially using the fact that the uniform distribution $\mathbf{u}$ (over the domain) is invariant by permutation. Note that every such $\mathbf{p}_{\varepsilon,\pi}$ then does belong to $\Pi_n(\mathbf{q}_\varepsilon)$, and is at total variation distance exactly ε from $\mathbf{q}_\varepsilon$. Moreover, we can repeat the previous argument *mutatis mutandis*: (i) the middle buckets provide no information whatsoever unless an element is seen twice, which requires $\Omega\big(n^{1/4}/\varepsilon\big)$ samples (the extra $1/\varepsilon$ due to our mixture with weight 9ε); while the two outer buckets have a discrepancy only $\delta := \frac{\varepsilon}{L-2}$, which to be detected requires at least $\Omega(1/\delta^2) = \Omega((\log^2 n)/\varepsilon^2)$ samples overall. The minimum of these two quantities gives the claimed lower bound, as long as $n^{1/4}/\varepsilon = \Omega((\log^2 n)/\varepsilon^2)$, that is, $\varepsilon = \Omega\big((\log^2 n)/n^{1/4}\big)$. ◀

[2] That is, conditioned on seeing each element of those $L-2$ buckets at most once, the conditional distribution over those $L-2$ buckets under (i) $\mathbf{q}$ and (ii) the uniform mixture of all perturbations $\mathbf{p}$ are indistinguishable.

4 Tolerant testing

We now turn to the task of *tolerant* testing. As mentioned in the introduction, tolerant testing is well known to be harder than standard (non-tolerant) testing, with a nearly quadratic gap for the standard identity testing problem ($\sqrt{n}$ vs. $\frac{n}{\log n}$ sample complexity). Surprisingly, we are able to show that under the promise of permutation, the task does not suffer a merely polynomial blowup – the sample complexity of tolerant identity testing becomes *exponentially* harder than that of standard testing, jumping from $\log^2 n$ to $n^{1-o(1)}$.

The first component, an $O(n/\log n)$ upper bound for tolerant testing under promise of permutation (Theorem 9), is straightforward, and simply follows from the corresponding upper bound absent this promise. A much more challenging task is in establishing the lower bound. We actually provide two lower bounds: the first, an $\Omega(n^{1-o(1)})$ lower bound (Theorem 10), applies for the usual setting of tolerant testing with an additive gap δ between ε' and ε). The second (Theorem 17) is an $\Omega\left(\sqrt{n/2^{O(C)}}\right)$ sample complexity lower bound for any C-factor approximation of the distance, that is to distinguish between ε-close and $C\varepsilon$-far.

4.1 Upper bound

The claimed upper bound readily follows from the analogous upper bound on tolerant testing *without* the promise of permutation, due to Valiant and Valiant [22, Theorem 4] (see, also, [16]). Indeed, any such estimator can be used for our problem, ignoring the additional promise of identity up to permutation.

▶ **Theorem 9.** *There exists an algorithm which, for any reference distribution* $\mathbf{q}$ *over* $[n]$ *and any* $0 \leq \varepsilon, \delta \leq 1$ *such that* $\delta = \Omega\big(1/\sqrt{\log n}\big)$*, and given* $O\left(\frac{n}{\delta^2 \log n}\right)$ *samples from an unknown distribution* $\mathbf{p} \in \Pi_n(\mathbf{q})$*, distinguishes with probability at least* $2/3$ *between (i)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon$ *and (ii)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon + \delta$*.*

We note that the requirement $\delta = \Omega\big(1/\sqrt{\log n}\big)$ has been relaxed in [16].

4.2 Lower bound

In this section, we prove the theorem below, our lower bound on the sample complexity of tolerant testing under promise of permutation. Before doing so, we emphasize that the known $\Omega\left(\frac{n}{\delta^2 \log n}\right)$ sample complexity lower bound for tolerant testing *absent* this promise does not apply to our setting, as the promise of permutation makes the testing problem easier. In particular, the hard instances used to prove the aforementioned $\Omega\left(\frac{n}{\delta^2 \log n}\right)$ lower bound do not satisfy this promise.[3]

▶ **Theorem 10.** *Any algorithm which, given a reference distribution* $\mathbf{q}$ *over* $[n]$*,* $0 < \varepsilon, \delta \leq 1$*, and sample access to an unknown distribution* $\mathbf{p} \in \Pi_n(\mathbf{q})$*, distinguishes with probability at least* $2/3$ *between (i)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon$ *and (ii)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon + \delta$*, must have sample complexity* $\Omega\big(\delta^2 n^{1-O(1/\log(1/\delta))}\big)$*.*

[3] One can also note that the lower bound for "standard" tolerant testing is obtained by choosing the reference distribution to be uniform over $[n]$. Under promise of permutation, this particular instance of the problem is trivial, as any permutation of the uniform distribution is still the uniform distribution.

Proof. In what follows, we assume that $\delta = \Omega(1/\sqrt{n})$, as otherwise there is nothing to prove. Let $k \geq 1$ be an integer to be chosen during the course of the analysis (we will set $k = \Theta(1/\delta)$), and write $n = 2mk^2$ for some integer $m \geq 1$ (this can be done without loss of generality, as our assumption on δ ensures that $n \geq 2mk^2$). For $1 \leq \ell \leq 2k$, we define the integer interval $I_{k,\ell} := [k] + (\ell - 1)k$, so that $[2k^2] = \bigcup_{\ell=1}^{2k} I_{k,\ell}$.

Given two distributions $\mathbf{p}, \mathbf{q}$ over $[k]$, we define families of distributions $\mathcal{C}_{\mathbf{p},\mathbf{q}}$ and $\mathcal{F}_{\mathbf{p},\mathbf{q}}$ over $[n]$ as follows: first, we consider the distributions $\mathbf{c}, \mathbf{f}$, each over $[2k^2]$, obtained by "repeating and alternating" $\mathbf{p}$ and $\mathbf{q}$ as follows:

- For $1 \leq \ell \leq k$ and $j \in I_{k,\ell}$, $\mathbf{c}(j) = \frac{1}{2k}\mathbf{p}(j)$.
- For $1 \leq \ell \leq k$ and $j \in I_{k,k+\ell}$, $\mathbf{c}(j) = \frac{1}{2k}\mathbf{q}(j)$.

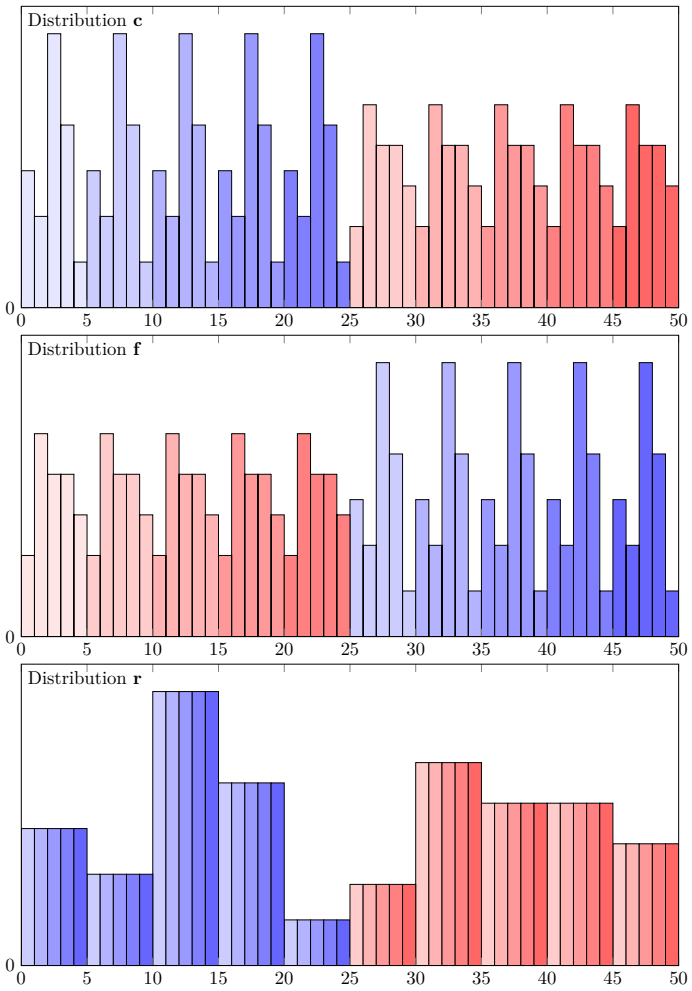

Figure 2 An example of $\mathbf{c}$ (top), $\mathbf{f}$ (middle), and $\mathbf{r}$ (bottom) over $[2k^2]$, for $k = 5$; here, we took $\mathbf{p} = \frac{1}{16}(3, 2, 6, 4, 1)$ and $\mathbf{q} = \frac{1}{18}(2, 5, 4, 4, 3)$.

We obtain $\mathbf{f}$ over $[2k^2]$ in a similar fashion, but swapping $I_{k,\ell}$ and $I_{k,k+\ell}$:

- For $1 \leq \ell \leq k$ and $j \in I_{k,\ell}$, $\mathbf{f}(j) = \frac{1}{2k}\mathbf{q}(j)$.
- For $1 \leq \ell \leq k$ and $j \in I_{k,k+\ell}$, $\mathbf{f}(j) = \frac{1}{2k}\mathbf{p}(j)$.

Further, we define our "reference" distribution $\mathbf{r}$ over $[2k^2]$ as

- For $1 \leq \ell \leq k$ and $j \in I_{k,\ell}$, $\mathbf{r}(j) = \frac{1}{2k}\mathbf{p}(\ell)$.
- For $k + 1 \leq \ell \leq 2k$ and $j \in I_{k,\ell}$, $\mathbf{r}(j) = \frac{1}{2k}\mathbf{q}(\ell)$.

We also define the reference distribution $\mathbf{r}^*_{\mathbf{p},\mathbf{q}}$ over $[n] = [2k^2m]$ by concatenating m copies of $\mathbf{r}$ and normalizing the result; that is,

$$\mathbf{r}^*_{\mathbf{p},\mathbf{q}} := \frac{1}{2m}(\mathbf{r} \sqcup \mathbf{r} \sqcup \cdots \sqcup \mathbf{r}),$$

where $\sqcup$ denotes the vector concatenation. Note that both $\mathbf{c}$ and $\mathbf{f}$ are permutations of $\mathbf{r}$, and that $\|\mathbf{r}\|_1 = \|\mathbf{c}\|_1 = \|\mathbf{f}\|_1 = 1$. Next, we bound the gap between $\mathrm{d}_{\mathrm{TV}}(\mathbf{f},\mathbf{r})$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{c},\mathbf{r})$, relating it to the distance between $\mathbf{p}$ and $\mathbf{q}$.

▷ Claim 11. $\mathrm{d}_{\mathrm{TV}}(\mathbf{f},\mathbf{r}) \geq \mathrm{d}_{\mathrm{TV}}(\mathbf{c},\mathbf{r}) + \frac{1}{k}\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q})$

Proof. We will analyze the contributions to $\mathrm{d}_{\mathrm{TV}}(\mathbf{c},\mathbf{r})$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{c},\mathbf{f})$ on $I_{k,\ell}$ and $I_{k,k+\ell}$ for $1 \leq \ell \leq k$. Without loss of generality, we can assume that $\mathbf{p},\mathbf{q}$ are non-decreasing. Then, from our definition of $\mathbf{c}$, $\mathbf{r}$, and $\mathbf{f}$, we have

$$\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(\mathbf{f},\mathbf{r}) &= \frac{1}{4k}\sum_{i=1}^{k}\sum_{j=1}^{k}(|\mathbf{p}(i)-\mathbf{q}(j)| + |\mathbf{q}(i)-\mathbf{p}(j)|) = \frac{1}{2k}\left(\sum_{i=1}^{k}\sum_{j=1}^{k}|\mathbf{p}(i)-\mathbf{q}(j)|\right) \\
&= \frac{1}{2k}\left(\sum_{i=1}^{k}|\mathbf{p}(i)-\mathbf{q}(i)| + \sum_{i=1}^{k}\sum_{j=1}^{i-1}(|\mathbf{p}(i)-\mathbf{q}(j)| + |\mathbf{p}(j)-\mathbf{q}(i)|)\right) \\
\mathrm{d}_{\mathrm{TV}}(\mathbf{c},\mathbf{r}) &= \frac{1}{4k}\sum_{i=1}^{k}\sum_{j=1}^{k}(|\mathbf{p}(i)-\mathbf{p}(j)| + |\mathbf{q}(i)-\mathbf{q}(j)|) \\
&= \frac{1}{2k}\sum_{i=1}^{k}\sum_{j=1}^{i-1}((\mathbf{p}(i)-\mathbf{p}(j)) + (\mathbf{q}(i)-\mathbf{q}(j)))
\end{aligned}$$

where for the last equality we used the assumption that $\mathbf{p},\mathbf{q}$ were non-decreasing to write

$$\sum_{i=1}^{k}\sum_{j=1}^{k}|\mathbf{p}(i)-\mathbf{p}(j)| = \sum_{i=1}^{k}\sum_{j=1}^{i-1}(\mathbf{p}(i)-\mathbf{p}(j)) + \sum_{i=1}^{k}\sum_{j=i+1}^{k}(\mathbf{p}(j)-\mathbf{p}(i)) = 2\sum_{i=1}^{k}\sum_{j=1}^{i-1}(\mathbf{p}(i)-\mathbf{p}(j)),$$

The conclusion then follows from recalling that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_{i=1}^{k}|\mathbf{p}(i)-\mathbf{q}(i)|$, and observing that $(\mathbf{p}(i)-\mathbf{p}(j)) + (\mathbf{q}(i)-\mathbf{q}(j)) = (\mathbf{p}(i)-\mathbf{q}(j)) + (\mathbf{q}(i)-\mathbf{p}(j)) \leq |\mathbf{p}(i)-\mathbf{q}(j)| + |\mathbf{p}(j)-\mathbf{q}(i)|$. ◁

To define $\mathcal{C}_{\mathbf{p},\mathbf{q}}$ and $\mathcal{F}_{\mathbf{p},\mathbf{q}}$, we will need one further piece of notation. We denote by $\mathcal{B}_k \subseteq \mathcal{S}_{2k^2}$ the set of all permutations of $[2k^2]$ "respecting the buckets," that is,

$$\mathcal{B}_k := \{\, \pi \in \mathcal{S}_{2k^2} \,:\, \pi(I_{k,\ell}) = I_{k,\ell} \forall \ell \in [2k] \,\}$$

We then let

$$\mathcal{C}_{\mathbf{p},\mathbf{q}} = \left\{ \frac{1}{2mk}(\mathbf{c}\circ\pi_1 \sqcup \mathbf{c}\circ\pi_2 \sqcup \cdots \sqcup \mathbf{c}\circ\pi_m) \,:\, \pi_1,\ldots,\pi_m \in \mathcal{B}_k \right\}$$

and

$$\mathcal{F}_{\mathbf{p},\mathbf{q}} = \left\{ \frac{1}{2mk}(\mathbf{f}\circ\pi_1 \sqcup \mathbf{f}\circ\pi_2 \sqcup \cdots \sqcup \mathbf{f}\circ\pi_m) \,:\, \pi_1,\ldots,\pi_m \in \mathcal{B}_k \right\}$$

where as before $\sqcup$ denotes the vector concatenation; that is, we stitch together m blocks, each consisting on a permuted version of either $\mathbf{c}$ or $\mathbf{f}$. Note that since $n = m \cdot 2k^2$ and each $\mathbf{c}$ (resp. $\mathbf{f}$) is a $(2k^2)$-dimensional vector, $\mathcal{C}_{\mathbf{p},\mathbf{q}}$ and $\mathcal{F}_{\mathbf{p},\mathbf{q}}$ are indeed families of probability distributions over $[n]$, and $\mathcal{C}_{\mathbf{p},\mathbf{q}}, \mathcal{F}_{\mathbf{p},\mathbf{q}} \subseteq \Pi_n(\mathbf{r}^*_{\mathbf{p},\mathbf{q}})$.

The construction above allows us to convert any two distributions $\mathbf{p},\mathbf{q}$ with sufficiently many matching moments to families of distributions (whose elements are all permutations of a single reference one) hard to distinguish:

▷ Claim 12. There exists some absolute constant $c > 0$ such that, if $\mathbf{p}, \mathbf{q}$ have matching first r-way moments, it is impossible to distinguish a uniformly random element of $\mathcal{C}_{\mathbf{p},\mathbf{q}}$ from a uniformly random element of $\mathcal{F}_{\mathbf{p},\mathbf{q}}$ given fewer than $cm^{1-\frac{1}{r+1}}$ samples.

Proof. By assumption on $\mathbf{p}, \mathbf{q}$ and out construction of $\mathbf{c}, \mathbf{f}$ from them, for every of the m contiguous blocks of $2k^2$ elements, the r-way moments of the corresponding conditional distributions exactly match. Given that a uniformly element drawn of $\mathbf{p}'$ from $\mathcal{C}_{\mathbf{p},\mathbf{q}}$ and $\mathbf{q}'$ from $\mathcal{F}_{\mathbf{p},\mathbf{q}}$ corresponds to independent permutations inside each block, any block in which fewer than $r+1$ samples falls brings exactly zero information about whether it comes from $\mathbf{p}'$ or $\mathbf{q}'$ (specifically, one could simulate the distribution of those $s < r+1$ samples without getting any sample from the real distribution). Since each of these m blocks has total probability $1/m$ under both $\mathbf{p}'$ and $\mathbf{q}'$, by a generalized birthday paradox (see, e.g., [19]), with probability at least $9/10$ no block will receive more than r samples unless the total number of samples is at least $cm^{1-\frac{1}{r+1}}$, for some absolute constant $c > 0$. ◁

It remains to specify *which* pair of distributions with "sufficiently many matching moments" we will use. While we could argue directly about the existence of such a pair of distributions with desirable properties, it is simpler to leverage a construction due to Valiant and Valiant [22], which exhibits the desired properties.

▷ Claim 13. There exists some $\varepsilon_0 > 0$ such that the following holds. For every sufficiently large r, there exists a pair of distributions (without loss of generality, non-decreasing) $\mathbf{p}_{\mathrm{VV}}, \mathbf{q}_{\mathrm{VV}}$ over $k = O(r2^r)$ elements with matching first r-way moments, but $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_{\mathrm{VV}}, \mathbf{q}_{\mathrm{VV}}) \geq \varepsilon_0$.

Proof. This follows from the lower bound construction of [22]. ◁

We will rely on this pair of distributions $\mathbf{p}_{\mathrm{VV}}, \mathbf{q}_{\mathrm{VV}}$, and hereafter write $\mathcal{C}, \mathcal{F}$, and $\mathbf{r}^*$ for $\mathcal{C}_{\mathbf{p}_{\mathrm{VV}},\mathbf{q}_{\mathrm{VV}}}, \mathcal{F}_{\mathbf{p}_{\mathrm{VV}},\mathbf{q}_{\mathrm{VV}}}$, and $\mathbf{r}^*_{\mathbf{p}_{\mathrm{VV}},\mathbf{q}_{\mathrm{VV}}}$, respectively.

▷ Claim 14. For every $\mathbf{p}' \in \mathcal{C}$ and $\mathbf{q}' \in \mathcal{F}$, we have $\mathrm{d}_{\mathrm{TV}}(\mathbf{q}', \mathbf{r}^*) > \mathrm{d}_{\mathrm{TV}}(\mathbf{p}', \mathbf{r}^*) + \frac{\varepsilon_0}{k}$.

Proof. Due to the definition of $\mathcal{C}$, $\mathcal{F}$, and $\mathbf{r}^*$ as m-fold concatenations, and since $\mathbf{r}$ is invariant by permutations from $\mathcal{B}_k$, it is sufficient to prove the claim for $\mathbf{p}_{\mathrm{VV}}$, $\mathbf{q}_{\mathrm{VV}}$, and $\mathbf{r}$ (over $[2k^2]$). The claimed bound then immediately follows from Claim 11. ◁

To finish the argument, it only remains to combine the various claims. We choose $k \geq \frac{\varepsilon_0}{\delta}$ and $m = n/(2k^2) \geq 1$ (since $\delta = \Omega(1/\sqrt{n})$. By Claim 13, we can then set $r := \Omega(\log k)$ and obtain, from Claim 12, a sample complexity lower bound of

$$\Omega\Big(m^{1-\frac{1}{r+1}}\Big) = \Omega\left(\delta^2 n^{1-O\left(\frac{1}{\log(1/\delta)}\right)}\right)$$

as desired. ◀

The theorem immediately implies the following two corollaries.

▶ **Corollary 15.** *For every $c > 0$, there exists some $\delta > 0$ such that the following holds. Any algorithm which, given a reference distribution $\mathbf{q}$ over $[n]$, $\varepsilon \in (0,1)$, and sample access to an unknown distribution $\mathbf{p} \in \Pi_n(\mathbf{q})$, distinguishes with probability at least $2/3$ between (i) $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon$ and (ii) $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon + \delta$, must have sample complexity $\Omega\big(n^{1-c}\big)$.*

▶ **Corollary 16.** *Any algorithm which, given a reference distribution $\mathbf{q}$ over $[n]$, $\varepsilon \in (0,1)$, and sample access to an unknown distribution $\mathbf{p} \in \Pi_n(\mathbf{q})$, distinguishes with probability at least $2/3$ between (i) $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon$ and (ii) $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon + 1/2^{\sqrt{\log n}}$, must have sample complexity $\frac{n}{2^{O(\sqrt{\log n})}}$.*

Tolerant testing C-approximation

We now turn to our second tolerant testing lower bound, which applies to algorithms providing a C-factor approximation of the distance to the reference distribution.

▶ **Theorem 17.** *Any algorithm which, given a reference distribution* $\mathbf{q}$ *over* $[n]$, $C \geq 2$, *and sample access to an unknown distribution* $\mathbf{p} \in \Pi_n(\mathbf{q})$, *distinguishes with probability at least* $2/3$ *between (i)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \leq \frac{1}{4C-1}$ *and (ii)* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \geq \frac{C}{4C-1}$, *must have sample complexity* $\Omega\big(\sqrt{\frac{n}{4^C}}\big)$.

▶ Remark 18. As discussed in Remark 7, Theorem 17 is essentially optimal, as it matches (up to polylogarithmic factors in the sample complexity) the upper bound from Theorem 5 when $C = \Theta(\log n)$.

Proof of Theorem 17. We will prove the theorem via a sequence of lemmas. We will assume that $C \geq 2$ is an integer, and we define $m = 2^C - 1$. Our proof will proceed similarly to the proof of Theorem 10. We will begin by working over $[m(2^{C+1} + 2^{C-1} - 3)]$. Throughout this section, we partition $[m(2^{C+1} + 2^{C-1} - 3)]$ into $C+1$ buckets, which we will denote $B_0, B_1, \ldots, B_C$, such that each B_i is a set of consecutive integers, $|B_C| = m2^{C-1}$, $|B_0| = m$, and $|B_i| = m2^{i+1}$ for $1 \leq i \leq C-1$. For convenience, we define $s := m(4C-1)2^{C-1}$.
We define a distribution $\mathbf{r}$ in the following way:

- For each $j \in B_0$, $\mathbf{r}(j) = \frac{2^C}{s}$.
- For each $1 \leq i \leq C-1$ and $j \in B_i$, $\mathbf{r}(j) = \frac{2^{C-i}}{s}$.
- For each $j \in B_C$, $\mathbf{r}(j) = \frac{1}{s}$.

We define two distributions $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{p}$ and $\mathbf{q}$ are hard to distinguish with few samples, such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{p})$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{q})$ are far apart. We define $\mathbf{q}$ in the following way:

- For each $j \in B_0$, $\mathbf{q}(j) = \frac{2^{C-1}}{s}$.
- For each $1 \leq i \leq C-1$,
  - For j in the first $m2^i$ elements of B_i, $\mathbf{q}(j) = \frac{2^{C-i-1}}{s}$.
  - For j in the next $m2^{i-1}$ elements of B_i, $\mathbf{q}(j) = \frac{2^{C-i}}{s}$.
  - For j in the last $m2^{i-1}$ elements of B_i, $\mathbf{q}(j) = \frac{2^{C-i+1}}{s}$.
- For each $j \in B_C$, $\mathbf{q}(j) = \frac{2}{s}$.

We define $\mathbf{p}$ as follows:

- For each $j \in B_0$,
  - If j is in the first 2^{C-1} elements of B_0, then $\mathbf{p}(j) = \frac{1}{s}$.
  - If j is in the last $m - 2^{C-1} = 2^{C-1} - 1$ elements of B_0, then $\mathbf{p}(j) = \frac{2^C}{s}$.
- For each $1 \leq i \leq C-1$ and $j \in B_i$, $\mathbf{p}(j) = \mathbf{r}(j) = \frac{2^{C-i}}{s}$.
- For each $j \in B_C$,
  - If j is in the first $(m-1)2^{C-1}$ elements of B_C, then $\mathbf{p}(j) = \frac{1}{s}$.
  - If j is in the last 2^{C-1} elements of B_j, then $\mathbf{p}(j) = \frac{2^C}{s}$.

▶ **Lemma 19.** *For* $0 \leq i \leq C$, $\sum_{j \in B_i} \mathbf{p}(j) = \sum_{j \in B_i} \mathbf{q}(j)$.

Proof. The proof is simply direct calculation. Observe that in bucket C,

$$s \sum_{j \in B_C} \mathbf{q}(j) = m2^{C-1} \cdot 2 = (m-1)2^{C-1} + (m+1)2^{C-1}$$
$$= (m-1)2^{C-1} \cdot 1 + 2^C \cdot 2^{C-1} = s \sum_{j \in B_C} \mathbf{p}(j).$$

In bucket 0, we have

$$s\sum_{j\in B_0}\mathbf{q}(j) = m\cdot 2^{C-1} = (m-1)2^{C-1}+2^{C-1} = (2^C-2)2^{C-1}+2^{C-1}$$
$$= (2^{C-1}-1)\cdot 2^C + 2^{C-1}\cdot 1 = s\sum_{j\in B_0}\mathbf{p}(j).$$

For $1 \le i \le C-1$, we have

$$s\sum_{j\in B_i}\mathbf{p}(j) = m2^{i+1}\cdot 2^{C-i}$$
$$= m(2^i + 2(2^{i-1}) + 2^{i+1})2^{C-1-i}$$
$$= m2^i\cdot 2^{C-i-1} + m2^{i-1}\cdot 2^{C-i} + m2^{i-1}\cdot 2^{C-i+1}$$
$$= s\sum_{j\in B_i}\mathbf{q}(j).$$

The claim follows by dividing the equalities by s. ◀

▶ **Lemma 20.** $\mathrm{d_{TV}}(\mathbf{r},\mathbf{q}) = \frac{C}{4C-1}$

Proof. By direct calculation,

$$2s\mathrm{d_{TV}}(\mathbf{r},\mathbf{q}) = s\sum_{j=1}^{s}|\mathbf{r}(j)-\mathbf{q}(j)|$$
$$= m2^{C-1}(2-1) + m(2^C - 2^{C-1})$$
$$+ \frac{1}{2}\sum_{i=1}^{C-1}\left(m2^i(2^{C-i}-2^{C-i-1}) + m2^{i-1}(2^{C-i+1}-2^{C-i})\right)$$
$$= m2^C + \sum_{i=1}^{C-1}(2^{i-1}m2^{C-i} + m2^{C-1-i}2^i)$$
$$= m2^C + \sum_{i=1}^{C-1}(m2^{C-1} + m2^{C-1})$$
$$= Cm2^C.$$

Dividing both sides by $2s$ yields the lemma. ◀

▶ **Lemma 21.** *For every* $0 \le i \le C$, $\mathbf{p}(B_i) \le \frac{2}{C+1}$ *(and similarly for* $\mathbf{q}(B_i)$*).*

Proof. We apply Lemma 19 and directly calculate. For bucket C, we get

$$\mathbf{p}(B_C) = \mathbf{q}(B_C) = \frac{2}{s}\cdot m2^C - 1 = \frac{2}{4C-1}.$$

For bucket 0, we get

$$\mathbf{p}(B_0) = \mathbf{q}(B_0) = \frac{2^{C-1}}{s}\cdot m = \frac{1}{4C-1}.$$

For $1 \le i \le C-1$, we get

$$\mathbf{q}(B_i) = \mathbf{p}(B_i) = \frac{2^i}{s}\cdot m(2^{C+1-i}) = \frac{4}{4C-1}.$$

The claim follows by observing that $\frac{4}{4C-1} \le \frac{2}{C+1}$ when $C \ge \frac{3}{2}$. ◀

▶ **Lemma 22.** $\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{p}) = \frac{1}{4C-1}$

Proof. By direct calculation,

$$2s\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{p}) = 2^{C-1}\cdot(2^C-1) + 2^{C-1}\cdot(2^C-1) = 2^C(2^{C-1}) = m2^C.$$

Dividing both sides by $2s$ yields the lemma. ◀

Let $w = m(2^{C+1}+2^{C-1}-3)$. We assume that n is a multiple of w, and define $t := \frac{n}{w}$. To define $\mathcal{C}$ and $\mathcal{F}$ over $[n]$, we will need one further piece of notation. We denote by $\mathcal{B}'_w \subseteq \mathcal{S}_w$ the set of all permutations of $[w]$ "respecting the buckets," that is, for every $0 \le i \le C$,

$$\mathcal{B}'_w = \{\pi \in \mathcal{S}_w : \pi(B_i) = B_i \forall i \in \{0,1,\dots,C\}\}$$

We then let $\mathbf{r}^* := \frac{1}{t}(\mathbf{r} \sqcup \mathbf{r} \sqcup \cdots \sqcup \mathbf{r})$ as well as

$$\mathcal{C} = \left\{ \frac{1}{t}(\mathbf{c}\circ\pi_1 \sqcup \mathbf{c}\circ\pi_2 \sqcup \cdots \sqcup \mathbf{c}\circ\pi_t) \ : \ \pi_1,\dots,\pi_t \in \mathcal{B}'_w \right\}$$
$$\mathcal{F} = \left\{ \frac{1}{t}(\mathbf{f}\circ\pi_1 \sqcup \mathbf{f}\circ\pi_2 \sqcup \cdots \sqcup \mathbf{f}\circ\pi_t) \ : \ \pi_1,\dots,\pi_t \in \mathcal{B}'_w \right\}$$

where as before $\sqcup$ denotes vector concatenation. Since $\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{c}\circ\pi) = \mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{c})$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{f}\circ\pi) = \mathrm{d}_{\mathrm{TV}}(\mathbf{r},\mathbf{f})$ for all $\pi \in \mathcal{B}'_s$, we have that $\mathrm{d}_{\mathrm{TV}}(\mathbf{r}^*,\mathbf{p}) = \frac{1}{4C-1}$ for every distribution $\mathbf{p} \in \mathcal{C}$, and $\mathrm{d}_{\mathrm{TV}}(\mathbf{r}^*,\mathbf{q}) = \frac{C}{4C-1}$ for every distribution $\mathbf{q} \in \mathcal{F}$. Further, repeating the same partitioning of each interval of s elements of $[n]$ into buckets $B_0, B_1, \dots, B_C$, we have $t(C+1)$ buckets, such that distinguishing a distribution in $\mathcal{C}$ from a distribution in $\mathcal{F}$ requires seeing at 2 samples in at least one of these buckets. Since the probability mass on each of the buckets at most $\frac{2}{t(C+1)}$ by Lemma 21, at least $\Omega(\sqrt{t(C+1)}) = \Omega(\sqrt{n(C+1)/w})$ queries to distinguish in $\mathcal{C}$ from a distribution in $\mathcal{F}$, completing the proof of Theorem 17. ◀

References

1. Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *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. URL: https://proceedings.neurips.cc/paper/2015/hash/1f36c15d6a3d18d52e8d493bc8187cb9-Abstract.html.
2. Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000)*, pages 259–269. IEEE Comput. Soc. Press, Los Alamitos, CA, 2000. doi:10.1109/SFCS.2000.892113.
3. Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Symposium on Theory of Computing Conference, STOC'04*, pages 381–390, New York, NY, USA, 2004. ACM. doi:10.1145/1007352.1007414.
4. Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Trans. Comput. Theory*, 11(2):Art. 6, 37, 2019. doi:10.1145/3305270.
5. Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi:10.4086/toc.gs.2020.009.

6 Clément L. Canonne and Karl Wimmer. Testing data binnings. In Jaroslaw Byrka and Raghu Meka, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2020, August 17-19, 2020, Virtual Conference*, volume 176 of *LIPIcs*, pages 24:1–24:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.APPROX/RANDOM.2020.24.

7 Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In Chandra Chekuri, editor, *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1193–1203. SIAM, 2014. doi:10.1137/1.9781611973402.88.

8 Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of SODA*, pages 1833–1852. Society for Industrial and Applied Mathematics (SIAM), 2013. URL: http://dl.acm.org/citation.cfm?id=2627817.2627948.

9 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 41, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

10 Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1841–1854. SIAM, Philadelphia, PA, 2015. doi:10.1137/1.9781611973730.123.

11 Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956. doi:10.1214/aoms/1177728174.

12 Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. doi:10.1080/14786440009463897.

13 Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Computational Complexity and Property Testing*, volume 12050 of *Lecture Notes in Computer Science*, pages 152–172. Springer, 2020.

14 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, July 1998.

15 Dayu Huang and Sean Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Trans. Inform. Theory*, 59(12):8157–8181, 2013. doi:10.1109/TIT.2013.2283266.

16 Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. *IEEE Trans. Inf. Theory*, 64(10):6672–6706, 2018.

17 Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

18 Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008. doi:10.1109/TIT.2008.928987.

19 Kazuhiro Suzuki, Dongvu Tonien, Kaoru Kurosawa, and Koji Toyota. Birthday paradox for multi-collisions. In *Information security and cryptology – ICISC 2006*, volume 4296 of *Lecture Notes in Comput. Sci.*, pages 29–40. Springer, Berlin, 2006. doi:10.1007/11927587_5.

20 Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010. URL: http://eccc.hpi-web.de/report/2010/179.

21 Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:180, 2010. URL: http://eccc.hpi-web.de/report/2010/180.

22 Gregory Valiant and Paul Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science – FOCS 2011*, pages 403–412. IEEE Computer Soc., Los Alamitos, CA, 2011. doi:10.1109/FOCS.2011.81.

23 Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. doi:10.1137/151002526.