Testing Data Binnings

Clément L. Canonne*   Karl Wimmer†

Abstract

Motivated by the question of data quantization and "binning," we revisit the problem of identity testing of discrete probability distributions. Identity testing (a.k.a. one-sample testing), a fundamental and by now well-understood problem in distribution testing, asks, given a reference distribution (model) $\mathbf{q}$ and samples from an unknown distribution $\mathbf{p}$, both over $[n] = \{1, 2, \ldots, n\}$, whether $\mathbf{p}$ equals $\mathbf{q}$, or is significantly different from it.

In this paper, we introduce the related question of *identity up to binning*, where the reference distribution $\mathbf{q}$ is over $k \ll n$ elements: the question is then whether there exists a suitable binning of the domain $[n]$ into k intervals such that, once "binned," $\mathbf{p}$ is equal to $\mathbf{q}$. We provide nearly tight upper and lower bounds on the sample complexity of this new question, showing both a quantitative and qualitative difference with the vanilla identity testing one, and answering an open question of Canonne [Can19]. Finally, we discuss several extensions and related research directions.

1 Introduction

Distribution testing [BFR⁺00], an area at the intersection of computer science, data science, and statistics which emerged as an offspring of the field of property testing [RS96, GGR98], concerns itself with the following type of questions: "upon observing independent data points originating from some unknown process or probability distribution $\mathbf{p}$, can we quickly and efficiently decide whether $\mathbf{p}$ satisfies some desirable property, or significantly violates this property?" One of the prototypical instances of this is the question of *identity testing* (also commonly known as one-sample testing, or goodness-of-fit), where one is given a reference distribution $\mathbf{q}$ and aims to test whether the unknown $\mathbf{p}$ is equal to this purported model $\mathbf{q}$, or far from it in statistical distance.

The sample and time complexity of identity testing have been thoroughly studied, and this question is now well-understood with regard to all parameters at play (see [GR00, BFF⁺01, Pan08, DKN15b, DK16, Gol20, DGPP18, VV17, BCG19], or the surveys [Can15, BW18]). However, at the very heart of the question's formulation lies a significant assumption: namely, that the *domain* of the distributions, generally taken to be the set $[n] := \{1, 2, \ldots, n\}$, is the "right" representation of the data. In many situations, this is not the case: the observations are made with a given (often arbitrary) level of granularity, e.g., imposed by the accuracy of the measuring equipment; this may lead to falsely (over)accurate measurements, with non-significant precision in the observations. In such cases, the domain $[n]$ of both the model $\mathbf{q}$ and the distribution of the measurements $\mathbf{p}$ are somewhat of a red herring, and relying on it to perform identity testing may lead to a wrong answer, by introducing discrepancies where there is none. Instead, a more robust approach is to decide if $\mathbf{p}$

---

*IBM Research, Almaden. Email: `ccanonne@cs.columbia.edu`

†Duquesne University. Email: `wimmerk@duq.edu`

conforms to $\mathbf{q}$ on some suitable quantization of the data: which leads to the question, first suggested as an open question in [Can19] and introduced in this work, of *testing identity up to binning*.

In more detail, the question we consider is the following:

Identity-Up-To-Binning$(n, \mathbf{q}, k, \varepsilon)$: Given a target number of bins k and a reference distribution $\mathbf{q}$ over $[k]$, a distance parameter ε, and i.i.d. samples from an unknown distribution $\mathbf{p}$ over $[n]$, is there a partition[1] of the domain in k intervals $I_1, \ldots, I_k$ such that $\mathbf{p}(I_j) = \mathbf{q}(j)$ for all $1 \leq j \leq k$, or is $\mathbf{p}$ at total variation distance at least ε from all distributions for which such a binning exists?

Before proceeding further, let us discuss some aspects of this formulation (which is illustrated in Fig. 1). First of all, this formulation is intrinsically tied to discrete distributions (which aligns with the motivations laid out earlier, relating to discretization of observations): indeed, if $\mathbf{p}$ is continuous, then for all choices of $\mathbf{q}$ there always exists a suitable partition of the domain; so that $\mathbf{p}$ trivially satisfies the property. This formulation also allows us to test for some underlying structure, without caring about other irrelevant details of the distribution. For instance, setting $\mathbf{q} = (1/3, 1/3, 1/3)$ captures all distributions $\mathbf{p}$ with three disjoint groups of elements, separated by an arbitrary number of zero-probability elements, each with total probability mass $1/3$.

In terms of the range of parameters, we encourage the reader to keep in mind the setting where $k \ll n$ (or even where k is a relatively small constant), which captures fitting an over-accurate set of measurements to a simple model. Yet, we emphasize that even the case $k = n$ is of interest and does not collapse to identity testing. Indeed, due to our allowing empty intervals, the problem is both qualitatively and quantitatively different from identity testing, and can be interpreted as identity testing up to merging some clusters of adjacent domain elements and having zero-probability elements in the reference distributions for data points considered irrelevant (see Fig. 2 for a simple illustration).
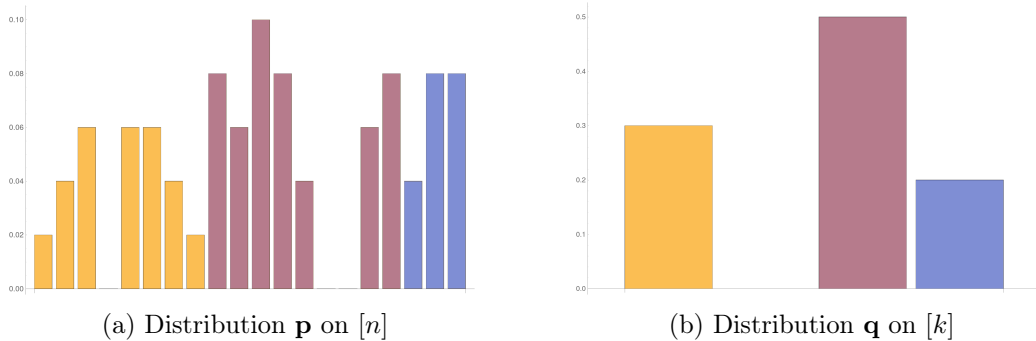

(a) Distribution $\mathbf{p}$ on $[n]$ (b) Distribution $\mathbf{q}$ on $[k]$

Figure 1: An example, for $n = 20$ and $k = 4$. Here, $\mathbf{q} = (3/10, 0, 1/2, 1/5)$, and $\mathbf{p} = \frac{1}{50}(1, 2, 3, 0, 3, 3, 2, 1, 4, 3, 5, 4, 2, 0, 0, 3, 4, 2, 4, 4)$. A possible partitioning is $I_1 = \{1, 2, \ldots, 8\}$, $I_2 = \emptyset$, $I_3 = \{9, \ldots, 17\}$, and $I_4 = \{18, 19, 20\}$.

We further note that one could consider variants of this problem, each with a slightly different focus. The first variant, *simultaneous binning*, would feature a reference distribution $\mathbf{q}$ over $[n]$

---

[1]Throughout this paper, by *partitions* we refer to ordered partitions with possibly empty subsets: i.e., a partition of $[n]$ in k sets is a sequence of pairwise disjoint sets $(I_1, \ldots, I_k)$ such that $\cup_{j=1}^{k} I_j = [n]$, with $|I_j| \geq 0$ for all j.
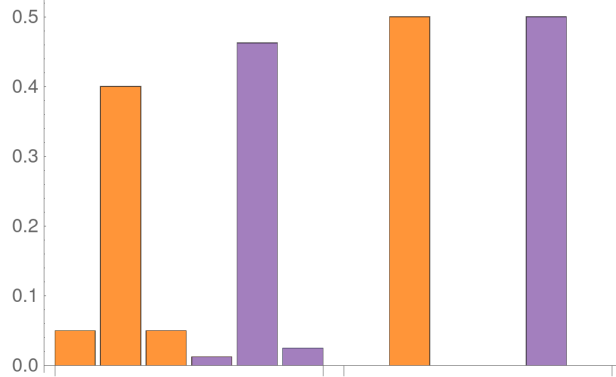
Figure 2: An example for $n = k = 6$. Here, $\mathbf{q} = (0, 1/2, 0, 0, 1/2, 0)$ (on the right), and $\mathbf{p} = (1/20, 2/5, 1/20, 1/80, 37/80, 1/40)$ (on the left). The reference distribution $\mathbf{q}$ focuses on the two modes of the distribution (writing off the other elements as "noise" or spurious), while the sampled distribution $\mathbf{p}$ also features those elements.

(instead of a set of bin probabilities $\mathbf{q}$ on $[k]$), and ask about the existence of an interval partition $I_1, \ldots, I_k$ such that $\mathbf{p}(I_j) = \mathbf{q}(I_j)$ for all j. One interesting aspect of this variant is that it generalizes non-trivially to continuous distributions (or mixtures of continuous and discrete). However, even insisting for non-empty intervals I_j's, this formulation is only interesting for distributions putting significant probability mass on the first $k - 1$ elements.[2] For $k = o(n)$, and in particular constant n, this is a significant restriction. A second variant, which suffers the same drawback but allows more flexibility, also provides a reference distribution $\mathbf{q}$ over $[n]$, but only asks about the existence of *independent binnings*: two interval partitions $I_1, \ldots, I_k$ and $I'_1, \ldots, I'_k$ such that $\mathbf{p}(I_j) = \mathbf{q}(I'_j)$ for all j.

We note that our upper bound (Theorem 1.1) easily extends to these two different variants as well; and our lower bound (Theorem 1.2) applies to the last independent binnings variant as well.

1.1 Our results and techniques

In this paper, we establish nearly matching upper and lower bounds on the **Identity-Up-To-Binning** problem. Our first result is an efficient algorithm for testing the property:

Theorem 1.1. *For every domain size n, number of bins $1 \le k \le n$, and fixed reference distribution $\mathbf{q}$ over $[k]$, there exists a (computationally efficient) algorithm for testing identity-up-to-binning to $\mathbf{q}$ with sample complexity $O\big(k/\varepsilon^2\big)$, where $\varepsilon \in (0, 1]$ is the distance parameter.*

Crucially, this upper bound is *independent* of the underlying domain size n of the unknown distribution, and only depends, linearly, on the target number of bins k. This is to be compared to the "standard" identity testing problem, which is known to have sample complexity $\Theta(\sqrt{n})$. One may wonder if the dependence on the number of bins in the above theorem is necessary, or if one could hope to achieve a *sublinear* dependence on k. Our next (and main) result shows that a near-linear dependence is unavoidable, ruling out any strongly sublinear dependence on k:

---

[2]Indeed, unless both the reference $\mathbf{q}$ and the unknown $\mathbf{p}$ put $\Omega(\varepsilon)$ mass on $\{1, 2, \ldots, k - 1\}$, a trivial partitioning with $I_j = \{j\}$ for $1 \le j \le k - 1$ and $I_k = \{k, \ldots, n\}$ provides an immediate answer to the problem.

Theorem 1.2 (Informal version of Theorem 4.1). *For every $k \geq 1$, there exists $n = n(k)$ and an absolute constant $\varepsilon_0 > 0$ such that the following holds. There exists a reference distribution $\mathbf{q}$ over $[k]$ such that any algorithm for testing identity-up-to-binning to $\mathbf{q}$ to distance ε_0 must have sample complexity $\Omega(k^{1-o(1)})$.*

Note that by standard techniques, this readily imply a lower bound of $\Omega(k^{1-o(1)}/\varepsilon)$, for all $\varepsilon \in (0, \varepsilon_0]$. We further conjecture the tight bound to be $\Omega(k/\varepsilon^2)$ (i.e., matching our upper bound). It is worth noting that the obvious and natural approach, a reduction from (standard) identity testing, would fall short of this goal, as it would only lead to a much weaker $\Omega(\sqrt{k}/\varepsilon^2)$ lower bound.

Our techniques. Our upper bound proceeds via the "testing-by-learning" paradigm. Specifically, given a reference distribution $\mathbf{q}$ over $[k]$, and sample access to an unknown distribution $\mathbf{p}$ over $[n]$, we learn a hypothesis distribution $\hat{\mathbf{p}}$ such that, informally, for every binning of $[n]$ into k intervals, the distributions $\mathbf{p}$ and $\hat{\mathbf{p}}$ are close. (More formally, the distribution $\hat{\mathbf{p}}$ to close to $\mathbf{p}$ in $\mathcal{A}_k$-distance, which we define later.) Since we have exact access to $\hat{\mathbf{p}}$, we can enumeratively test every potential binning. Although this does not save on the query complexity, we are able to use dynamic programming to find the best binning in time polynomial in n and k.

The proof of our lower bound is significantly more involved and proceeds in three stages, which we outline here. We first show that, if n is sufficiently large as a function of k, then we may assume that the algorithm only looks at the ordering of the samples received from $[n]$, instead of at their actual "names" (we refer to the ordering as the *ordered fingerprint* of the sample). Our bound uses a Ramsey theory argument very similar to that of [DKN15a].

We then show, for every integer m, the existence of two different distributions $\mathbf{p}'$ and $\mathbf{q}'$ over $[\text{poly}(m) \cdot 2^m]$ such that (i) $\mathbf{p}'$ and $\mathbf{q}'$ are far in total variation distance, and (ii) given the ordered fingerprint of m samples generated by one of $\mathbf{p}'$ and $\mathbf{q}'$, it is information-theoretically impossible to determine which of $\mathbf{p}'$ and $\mathbf{q}'$ generated the ordered fingerprint. There are 2^m such ordered fingerprints, one for every composition of m, so it is unclear if the support size of the distributions in this claim can be significantly reduced. While this is an independently interesting problem, any progress on this problem would only directly result in a decrease of the $o(1)$ exponent in our main theorem.

Finally, using the above pair of distributions as "buckets" of two distributions $\mathbf{p}$ and $\mathbf{q}$, respectively, we show that any algorithm distinguishes $\mathbf{p}$ and $\mathbf{q}$ and only considers ordered fingerprints must have access to at least $m + 1$ samples from at least one of the (assuming m is independent of k) $\Omega(k)$ buckets. By collision bounds for the generalized birthday problem (see also [STKT06] for more formal statements), with high probability we need a sample size of $\Omega(k^{1-1/(m+1)})$ to ensure that the algorithm succeeds.

1.2 Future directions

We conclude with a few research directions we deem particularly promising. A first natural question is to understand the analogue of the **Identity-Up-To-Binning** problem for *closeness* testing, that is, when both $\mathbf{p}$ (over $[n]$) and $\mathbf{q}$ (over $[k]$) are unknown and available only through i.i.d. samples. This would in particular capture situations where data is collected from two different sources (or sensors), each using a different discretization scheme or with different precision, and one aims at deciding whether the underlying distributon is the same.

Another avenue is to consider the *tolerant* testing version of the problem, that is, to decide whether **p** is *close* to a binning of **q** (versus far from any such binning). It is known that tolerant identity testing has significantly larger (namely, nearly linear in the domain size) sample complexity than identity testing; whether this is still the case for their identity-up-to-binning analogues is an intriguing question.

Lastly, it would be interesting to generalize the question to other partially-ordered domains (e.g., the hypergrid $[n]^d$, or the Boolean hypercube $\{0,1\}^n$), for the suitable notion of "interval" in these posets.

2 Preliminaries

All throughout the paper, we write log for the binary logarithm. Given integers n and k, we write $\mathcal{J}_{n,k}$ for the set of all $\binom{n+k-1}{n}$ partitions of $[n]$ in k consecutive (and possibly empty) pairwise disjoint intervals. Recall that the total variation distance between two distributions $\mathbf{p}, \mathbf{p}'$ over $[n]$ is $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2) = \sup_{S \subseteq [n]}(\mathbf{p}(S) - \mathbf{p}'(S))$, and is equal to $\frac{1}{2}\sum_{i=1}^n |\mathbf{p}(i) - \mathbf{p}'(i)|$, half the ℓ_1 distance between their probability mass functions.

The **Identity-Up-To-Binning**$(\mathbf{q}, k, \varepsilon)$ problem we consider in this paper is then formally defined as follows. Given a reference distribution $\mathbf{q}$ on $[k]$, we consider the property

$$\mathcal{P}_{\mathbf{q}} := \left\{ \mathbf{p}' \in \Delta_n \; : \; \exists (I_1, \ldots, I_k) \in \mathcal{J}_{n,k}, \; \max_{j \in [k]} \left| \mathbf{p}'(I_j) - \mathbf{q}(j) \right| = 0 \right\} \tag{1}$$

The question is then, given a distance parameter $\varepsilon \in (0,1]$, to distinguish (i) $\mathbf{p} \in \mathcal{P}_{\mathbf{q}}$ from (ii) $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}}) > \varepsilon$, where $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}}) = \min_{\mathbf{p}' \in \mathcal{P}_{\mathbf{q}}} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}')$. We will rely extensively, for our lower bounds, on the below fact, which allows us to lower bound $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}})$ by a more actionable quantity.

Fact 2.1. *For any distributions* $\mathbf{p}$ *and* $\mathbf{q}$ *(over* $[n]$ *and* $[k]$, *respectively) we have*

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}}) \geq \frac{1}{2} \, \mathrm{dist}(\mathbf{p}, \mathbf{q}) := \min_f \sum_{j=1}^k \left| \mathbf{q}(j) - \sum_{t : f(t)=j} \mathbf{p}(t) \right|,$$

where the minimum is taken over all nondecreasing functions $f \colon [n] \to [k]$.

Proof. Fix $\mathbf{p}, \mathbf{q}$ as in the statement. One can think of such a nondecreasing f as defining a decomposition of $[n]$ into k disjoint (and possible empty) intervals $I_1, \ldots, I_k$, where $I_j = f^{-1}(j)$. In particular, let $\mathbf{p}^* \in \mathcal{P}_{\mathbf{q}}$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}^*) = \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}})$, and $I_1^*, \ldots, I_k^*$ the corresponding partition for $\mathbf{p}^*$. Define $f^* \colon [n] \to [k]$ from this partition by setting $f(t) := \sum_{j=1}^k j \mathbb{1}_{I_j^*}(t)$: f^* is then non-decreasing, and

$$\sum_{j=1}^k \left| \mathbf{q}(j) - \sum_{t : f^*(t)=j} \mathbf{p}(t) \right| = \sum_{j=1}^k \left| \mathbf{q}(j) - \mathbf{p}(I_j^*) \right| = \sum_{j=1}^k \left| \mathbf{p}^*(I_j^*) - \mathbf{p}(I_j^*) \right| \leq \sum_{i=1}^n \left| \mathbf{p}^*(i) - \mathbf{p}(i) \right|$$

where the last equality is due to the fact that $\mathbf{p}^*(I_j^*) = \mathbf{q}(j)$ for all $j \in [k]$ (since $\mathbf{p}^* \in \mathcal{P}_{\mathbf{q}}$), and last inequality is the triangle inequality. Since the RHS equals $2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}^*)$, we get $\min_f \sum_{j=1}^k |\mathbf{q}(j) - \sum_{t : f(t)=j} \mathbf{p}(t)| \leq 2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}})$. $\square$

3 Algorithms

In this section, we prove Theorem 1.1, restated below:

Theorem 3.1. *For any fixed reference distribution* $\mathbf{q}$ *over* $[n]$ *and number of bins* $1 \leq k \leq n$, *there exists a (computationally efficient) algorithm for* **Identity-Up-To-Binning**$(n, \mathbf{q}, k, \varepsilon)$ *with sample complexity* $O\big(k/\varepsilon^2\big)$.

Proof. We will use the following notion of $\mathcal{A}_\ell$-*distance*, which interpolates between Kolmogorov distance ($\ell = 2$) and (twice the) total variation ($\ell = n$) [DL01, CDSS12, DKN15b, DKN15a]: for any two distributions $\mathbf{p}', \mathbf{q}'$ over $[n]$,

$$\|\mathbf{p}' - \mathbf{q}'\|_{\mathcal{A}_\ell} := \max_{(I_1, \ldots, I_\ell) \in \mathcal{J}_{n,\ell}} \sum_{j=1}^{\ell} \big|\mathbf{p}'(I_j) - \mathbf{q}'(I_j)\big| \tag{2}$$

One can check that this defines a *bona fide* norm. Further, it is known, from the VC inequality, that $O\big(\ell/\varepsilon^2\big)$ samples are sufficient to learn any distribution in $\mathcal{A}_\ell$-distance ε (and with failure probability $1/6$) [DL01, CDSS12], and, further, that the empirical estimator achieves this bound. Using this, the algorithm is as follows:

1. Learn the unknown $\mathbf{p}$ to $\mathcal{A}_k$-distance ε, with failure probability $1/6$. Call the result (the empirical estimator, which is a distribution over $[n]$) $\hat{\mathbf{p}}$.

2. Compute the minimum $\Delta_{\mathbf{q}}(\mathcal{I}, \hat{\mathbf{p}}) := \sum_{j=1}^{k} |\hat{\mathbf{p}}(I_j) - \mathbf{q}(j)|$ over all partitions $(I_1, \ldots, I_k) \in \mathcal{J}_{n,k}$ in intervals such that $|I_j| > 0$ whenever $\mathbf{q}(j) > 0$.

3. If there exists $\mathcal{I}$ such that $\Delta_{\mathbf{q}}(\mathcal{I}, \hat{\mathbf{p}}) \leq \varepsilon$, return accept; otherwise, return reject.

Note that only the first step requires samples from $\mathbf{p}$, so that the sample complexity is indeed $O\big(k/\varepsilon^2\big)$. The second step is purely computational, and can be implemented in time $\mathrm{poly}(n, k)$ via a simple dynamic programming approach.

We now argue correctness. With probability at least $5/6$, the first step produced a correct $\hat{\mathbf{p}}$, i.e., one that is $(\varepsilon/4)$-close to $\mathbf{p}$ in $\mathcal{A}_k$-distance; we hereafter assume this holds.

Completeness. Suppose $\mathbf{p} \in \mathcal{P}_{\mathbf{q}}$, and let $\mathcal{I}^* = (I_1^*, \ldots, I_k^*) \in \mathcal{J}_{n,k}$ be any partition witnessing it. Note that this partition then satisfies $|I_j^*| > 0$ for all j such that $\mathbf{q}(j) > 0$, as otherwise the distance after binning is positive. Then,

$$\Delta_{\mathbf{q}}(\mathcal{I}^*, \hat{\mathbf{p}}) = \sum_{j=1}^{k} \big|\hat{\mathbf{p}}(I_j^*) - \mathbf{q}(j)\big| \leq \sum_{j=1}^{k} \big|\hat{\mathbf{p}}(I_j^*) - \mathbf{p}(I_j^*)\big| + \sum_{j=1}^{k} \big|\mathbf{p}(I_j^*) - \mathbf{q}(j)\big| \leq \|\hat{\mathbf{p}} - \mathbf{p}\|_{\mathcal{A}_k} \leq \varepsilon$$

the second-to-last inequality by definition of $\mathcal{A}_k$-distance and the fact that $\sum_{j=1}^{k} |\mathbf{p}(I_j^*) - \mathbf{q}(j)| = 0$. Therefore, the algorithm will find a good partition and output accept.

Soundness. Suppose now by contrapositive that the algorithm outputs accept. This means it found, in step 2, some partition $\mathcal{I}^* = (I_1^*, \ldots, I_k^*) \in \mathcal{J}_{n,k}$ such that $\Delta_{\mathbf{q}}(\mathcal{I}^*, \hat{\mathbf{p}}) \leq \varepsilon/4$. But then,

$$\sum_{j=1}^{k} \big|\mathbf{p}(I_j^*) - \mathbf{q}(j)\big| = \sum_{j=1}^{k} \big|\mathbf{p}(I_j^*) - \hat{\mathbf{p}}(I_j^*)\big| + \sum_{j=1}^{k} \big|\hat{\mathbf{p}}(I_j^*) - \mathbf{q}(j)\big| \leq \|\hat{\mathbf{p}} - \mathbf{p}\|_{\mathcal{A}_k} + \Delta_{\mathbf{q}}(\mathcal{I}^*, \hat{\mathbf{p}}) \leq 2\varepsilon$$

the second-to-last inequality again by definition of $\mathcal{A}_k$-distance. We claim that this implies that $\mathbf{p}$ is ε-close (in total variation distance) to some $\mathbf{p}^* \in \mathcal{P}_{\mathbf{q}}$. Indeed, we can build $\mathbf{p}^*$ in a greedy fashion from $\mathbf{p}$: as long as there exist $I_{j_1}^*$, $I_{j_2}^*$ such that $\mathbf{p}(I_{j_1}^*) > \mathbf{q}(j_1)$ but $\mathbf{p}(I_{j_2}^*) < \mathbf{q}(j_2)$, we move $\delta := \min(|\mathbf{p}(I_{j_1}^*) - \mathbf{q}(j_1)|, |\mathbf{p}(I_{j_2}^*) - \mathbf{q}(j_2)|)$ probability mass from (arbitrary) elements of $I_{j_1}^*$ to an arbitrary element of $I_{j_2}^*$. Here, we used our condition on the partition found in step 2, which ensures $I_{j_2}^*$ is non-empty. This incurs total variation δ (from $\mathbf{p}$), and reduces $\sum_{j=1}^{k} |\mathbf{p}(I_j^*) - \mathbf{q}(j)|$ by 2δ. Repeating until it is no longer possible, we obtain the claimed $\mathbf{p}^*$, and therefore that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_{\mathbf{q}}) \leq \varepsilon$. $\qquad\square$

Remark 3.2. As mentioned in the introduction, it is straightforward to adapt the algorithm and argument to either the "simultaneous binning" and "independent binnings" variants of the problem, leading to the same sample complexity upper bound (and time complexity) for these related questions.

4   Lower Bounds

In this section, we prove our main theorem (Theorem 1.2), restated below:

Theorem 4.1. *For every $k \geq 1$, there exists $n = n(k)$ such that the following holds. There exists a reference distribution $\mathbf{q}$ over $[k]$ such that any algorithm for **Identity-Up-To-Binning**$(n, \mathbf{q}, k, \varepsilon)$ must have sample complexity $\frac{k}{2^{O(\sqrt{\log k})}}$.*

At a high level, the proof of Theorem 4.1 consists of three parts:

(1) First, we show that, for every n, there is an $N(n)$ such that, if there is a algorithm that succeeds in the data binning problem over $[N(n)]$ with s samples, then there is an algorithm that succeeds in the data binning problem over $[n]$ with s samples, such that this latter algorithm only considers the "ordered fingerprint" of the sample. This approach, analogous to that of [DKN15a], uses Ramsey theory, and enables us to restrict ourselves to (simpler to analyze) order-based algorithms. (Proposition 4.4)

(2) Second, we show for every constant m, there exist $b(m)$ and two distributions $\mathbf{p}'$ and $\mathbf{q}'$ on $[b]$ that (i) are "far from being cyclic shifts" of one another (which will allow us to argue about distance to being a binning later), and (ii) cannot be distinguished from ordered fingerprints from s samples. The function $b(m)$ is exponential in m. (Lemma 4.7)

(3) Third, we consider the distributions $\mathbf{p}$ and $\mathbf{q}$ on $[k]$ constructed such that the distribution restricted to each block of $b(m)$ elements is $\mathbf{p}'$ and $\mathbf{q}'$, respectively, in the natural way. We show that (i) $\mathbf{p}$ is "far" from any binning of $\mathbf{q}$, and (ii) distinguishing these two distributions with ordered fingerprints requires seeing $m + 1$ samples from one particular block. By folklore collision bounds, this implies a $\Omega(k^{1-1/(m+1)})$ sample lower bound for distinguishing $\mathbf{p}$ and $\mathbf{q}$ from ordered fingerprints. (Corollary 4.11)

Combining the three (applying (1) with n set to k, and m in (3) set to $\sqrt{\log k}$) then yields the theorem.

4.1 From samples to ordered fingerprints

The first step of our reduction will consist in "hiding" information from the algorithms. Roughly speaking, for the sake of our lower bound analysis we would like to argue that, without loss of generality, any testing algorithm can be assumed to only be given the *order relations* between the samples from the unknown distribution, instead of the samples themselves. For instance, instead of seeing four samples $12, 7, 98, 7$, we wish to restrict our analysis to algorithms which only see that $x_{(1)} = x_{(2)} < x_{(3)} < x_{(4)}$. Since we are drawing i.i.d. samples from the distribution, the order that the samples come in does not matter.

However, this simplifying assumption is not actually without loss of generality, and does not quite hold as stated: an algorithm *can* sometimes infer more information from the values of the samples from $\mathbf{p}$ than from their ordered relations alone. We will prove a weaker statement, sufficient for our purposes; in order to do so, we start by introducing some notions formalizing the aforementioned "order relations."

Definition 4.2 (Ordered fingerprints). The *ordered fingerprint* of a sequence of s values in $[n]$ is an ordered frequency vector of the elements occurring in the sequence, with labels removed. Formally, for each element $i \in [n]$, let $\boldsymbol{F}_i$ be the number of times element i appears, and let $j_1 < j_2 < \ldots < j_t$ be the indices such that $\boldsymbol{F}_{j_i}$ is positive (so that t is the number of distinct elements in the sequence). The ordered fingerprint is then the ordered collection of positive integers $(\boldsymbol{F}_{j_1}, \boldsymbol{F}_{j_2}, \ldots, \boldsymbol{F}_{j_t})$ such that $\sum_{i=1}^{t} \boldsymbol{F}_{j_i} = s$.

To later argue about indistinguishability of our instances, we will also rely on the below notion of s-way moments (induced by an ordered fingerprint):

Definition 4.3 (s-way moments). Given a distribution $\mathbf{p}$ over $[n]$ and an ordered tuple of positive integers $F = (F_1, F_2, \ldots, F_t)$ such that $\sum_{i=1}^{t} F_i = s$, the probability that the ordered fingerprint on s samples is $(F_1, F_2, \ldots, F_t)$ is given by

$$\mathbf{p}^F := \binom{s}{F_1, F_2, \ldots, F_t} \sum_{1 \le i_1 < i_2 < \cdots < i_t \le n} \prod_{j=1}^{t} \mathbf{p}(i_j)^{F_j} \tag{3}$$

We call such an expression an *s-way moment of* $\mathbf{p}$. Note that the s-way moments of $\mathbf{p}$ completely determine the distribution of the ordered fingerprint of s random samples drawn from $\mathbf{p}$. Thus, for two distributions $\mathbf{p}$ and $\mathbf{q}$, if $\mathbf{p}^F = \mathbf{q}^F$ for all F such that $|F| = s$, then $\mathbf{p}$ and $\mathbf{q}$ cannot be distinguished from ordered fingerprints on s samples.[3]

We now state and prove the following lemma, which captures the intuition discussed above and will be the first component of our lower bound:

Proposition 4.4. *For every $n \ge 1$ and $s \ge 1$, there exists $N \ge 1$ such that the following holds for every $k \ge 1$, reference distribution $\mathbf{q}$ over $[k]$, and $\varepsilon \in (0, 1]$. If there exists an algorithm $\mathfrak{A}_N$ that, for every distribution $\mathbf{p}'$ over $[N]$, solves the problem* **Identity-Up-To-Binning**$(N, \mathbf{q}, k, \varepsilon)$ *with s samples from $\mathbf{p}'$, then there exists an algorithm $\mathfrak{A}_n$ that, for every distribution $\mathbf{p}$ over $[n]$, solves the problem* **Identity-Up-To-Binning**$(n, \mathbf{q}, k, \varepsilon)$ *with s samples from $\mathbf{p}$. Moreover, $\mathfrak{A}_n$ only considers the ordered fingerprint of the s samples.*

---

[3]Indeed, for every s, the $(s-1)$-way moments are linear combinations of s-way moments; we omit the details.

Proof. The argument is similar to that of Diakonikolas, Kane, and Nikishkin [DKN15a], and relies on a result of Conlon, Fox, and Sudakov:

Lemma 4.5 ([CFS10]). *Given a set S and an integer t, let $\binom{S}{t}$ denote the set of subsets of cardinality t. For all positive integers, a, b, and c, there exists a positive integer N so that for any function $\chi\colon \binom{[N]}{a} \to [b]$, there exists $S \subseteq [N]$ with $|S| = c$ such that χ is constant on $\binom{S}{a}$.*

As in the proof of [DKN15a, Theorem 13], given n, s, we will invoke Lemma 4.5 to obtain a new domain size N, in a way detailed below: loosely speaking, a corresponds to the number of samples s, b to the number of possible decision (binary) functions from a given set of s samples which only depend on the ordered fingerprint of those samples, and c to the target domain size s. We will use this to, given $\mathfrak{A}_N$ (which induces a mapping χ from sets of s samples to such order-based decision functions), find a set $S \subseteq [N]$ of size n, and a monotone function $f\colon [n] \to [N]$ such that $f([n]) = S$. The algorithm $\mathfrak{A}_s$ then runs the promised algorithm $\mathfrak{A}_N$ on the distribution $f(\mathbf{p})$ obtained by applying f to the s samples. Since f is order-preserving, the distances will not change. Since we can assume that $\mathfrak{A}_N$ is deterministic once we know which distinct samples we get from $[n]$, the fact that the induced χ is constant on $\binom{f([n])}{s}$ guarantees the output will be a function of the ordered fingerprint only.

However, we have an issue that [DKN15a] does not. When we draw a total of s samples, we likely will not get s distinct samples. [DKN15a] gets around this issue by "dividing" all elements of $[n]$ into sub-elements, and upon seeing an element in the sample, assign it to a uniformly random one of its sub-elements. However, our main parameter of interest k differs from theirs: if we applied this procedure to our problem, the bounds obtained would deteriorate. That is, while the $\mathcal{A}_k$-distance (the focus of [DKN15a]) would not change, the coarsening distance we consider here would.

We handle this issue as follows. Given an integer s, we will color all nonempty subsets of $[N]$ with *at most* s elements. This color associates to a set of s samples the function that $\mathfrak{A}_N$ uses to accept or reject given the ordered fingerprint that accompanies these samples. Now, we apply Lemma 4.5 repeatedly. In our first application, we find a subset $S_s \subseteq [N]$ such that, conditioning on distinct samples, $\mathfrak{A}_N$ has consistent behavior.[4] Now we remove all subsets that contain elements outside of S_s, and we apply Lemma 4.5 again to find a set $S_{s-1} \subseteq S_s$ such that $\mathfrak{A}_N$'s behavior only depends on the ordered fingerprint given that in the s samples, there are at least $s - 1$ distinct values seen. Continuing in this fashion, after applying Lemma 4.5 a total of s times, we arrive at a set S_1 (with $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_s$) such that, conditioned on samples coming only from S_1, $\mathfrak{A}_N$ depends only on the ordered fingerprint of the samples seen. In every application of the lemma, we have $a \leq s$, and since there are 2^s possible ordered fingerprints given s samples, we have $b \leq 2^{2^s}$. We set $c = n$ in our final application. $\qquad\square$

4.2 Sequences far from being shifts of each other

As a building block from our lower bound construction, we will require the existence of (non-negative) sequences which are "far" from each other, in that any circular shift of the first remains far, in Hamming distance, from the second.

Definition 4.6 (Partial shifts). Given an alphabet Σ and two strings x and y in Σ^n, we say that y is an *r-partial cyclic shift of x* if there is an integer ℓ and a nondecreasing function $f\colon [n] \to [n]$ such

---

[4]In this case, the ordered fingerprint is the all ones vector of length s, so the functions are not very interesting.

that, for at least r indices i of $[n]$, $x_i = y_{(f(i)+\ell) \bmod n}$. That is, thinking of the symbols of y placed in a circle, y contains a substring r of x (without wrapping around more than once).

For two distributions $\mathbf{p}$ and $\mathbf{q}$ over $[n]$, we then say that $\mathbf{q}$ *is an r-partial cyclic shift of* $\mathbf{p}$ if the string $\mathbf{q}(1)\mathbf{q}(2)\cdots\mathbf{q}(n)$ is an r-partial cyclic shift of $\mathbf{p}(1)\mathbf{p}(2)\cdots\mathbf{p}(n)$.

We observe that finding a witness to x and y being r-partial cyclic shifts of each other is equivalent to finding the *cyclic longest common subsequence* between x and y, which is a problem with applications in DNA sequencing [Ngu12]. With these notions in hand, we are ready to prove the key technical lemma underlying part (2) of our argument:

Lemma 4.7. *Let $m > 4$ be a positive integer, and $b = 5m^2 2^m$. Then there exist two distributions $\mathbf{p}'$ and $\mathbf{q}'$ over $[b]$ such that (i) $\mathbf{p}'$ and $\mathbf{q}'$ are not $(99b/100)$-partial cyclic shifts of each other, and (ii) $\mathbf{p}'$ and $\mathbf{q}'$ agree on all m-way moments.*

Proof. Let $\mathbf{p}$ be a probability distribution over $[b]$ such that its probability mass function takes the values $2/(\frac{5}{2}b)$ and $3/(\frac{5}{2}b)$ each on $b/2$ of the elements of $[b]$. Each m-way moment $\mathbf{p}^F$ can be expressed as a homogeneous polynomial in the b values $\mathbf{p}(i)$ of total degree m, such that each coefficient is 0 or $K_F := \binom{s}{F_1, F_2, \ldots, F_t}$; note that K_F is independent of $\mathbf{p}$. The number of terms of such a polynomial is at most $\binom{b+m-1}{m} \le (b+m-1)^m \le (2b)^m$. Thus, every m-way moment of $\mathbf{p}$ evaluates to an expression of the form $qK_F/(\frac{5}{2}b)^m$, where q is a integer such that $0 \le q \le (6b)^m$. Since there are 2^m m-way moments of $\mathbf{p}$, there are at most $(6b)^{m2^m}$ possible values of the m-way moments of $\mathbf{p}$.

For the claim about cyclic shifts, we use a counting argument. We will show that for any string $x \in \{2,3\}^b$, there are at most $2^{b/5}$ other strings that are $(99b/100)$-partial cyclic shifts of x. Indeed, any $(99b/100)$-partial cyclic shift y of x can be constructed by the following process: first, select $99b/100$ positions of x that will be present in y–these positions will form a witness. Next, select $99b/100$ positions of y that these bits of x will appear in, and a cyclic shift for these bits. Finally, select any values in $\{2,3\}$ for the remaining $b/100$ positions of y. Thus, the number of possible y's is at most

$$\binom{b}{b/100}^2 (99b/100)2^{b/100} \le 2^{2h(1/100)}2^{b/50} \le 2^{b/5}$$

where we have used the bound $\binom{n}{\alpha n} \le 2^{nh(\alpha)}$, where $h(\alpha) = \alpha \log \alpha + (1-\alpha)\log(1-\alpha)$ is the binary entropy. Since there are a total of $\binom{b}{b/2} \ge 2^b/b$ strings of length b with an equal number of 2's and 3's, then for every b, we can find a set S of $2^{4b/5}/b$ strings of length b such that no two of them are $(99b/100)$-partial cyclic shifts of each other. We map strings over $\{2,3\}^b$ to distributions over $[b]$ in a natural way: given $x \in S$, we define the corresponding distribution $\mathbf{p}$ such that $\mathbf{p}(i) = x_i/(\frac{5}{2}b)$. Thus, we have $2^{4b/5}/b$ distributions over $[b]$, no two of which are $(99b/100)$-partial cyclic shifts of each other. Setting $b = 5m^2 2^m$, we then have

$$\frac{1}{b}2^{4b/5} = \frac{2^{4m^2 2^m}}{5m^2 2^m} \ge 2^{3m^2 2^m} = (2^{3m})^{m2^m} > (30m^2 2^m)^{m2^m} = (6b)^{m2^m}.$$

Thus, by the pigeonhole principle there exist two distributions $\mathbf{p}'$ and $\mathbf{q}'$ over $[b]$ that are not $(99b/100)$-partial cyclic shifts of each other, yet such that $\mathbf{p}'$ and $\mathbf{q}'$ agree on all m-way moments. $\square$

4.3 Constructing the hard-to-distinguish instances over $[n]$

For a fixed integer $m > 1$ (to be determined later) and sufficiently large k, let $\mathbf{p}'$ and $\mathbf{q}'$ be the two distributions over $[b]$ promised by Lemma 4.7, where $b = 5m^2 2^m$; and let $k' := k/b$. We will define $\mathbf{p}$ and $\mathbf{q}$ to be distributions over $[bk']$ in the following way: We partition $[bk']$ into blocks $B_1, B_2, \ldots, B_{k'}$, where $B_j = \{b(j-1)+1, b(j-1)+2, \ldots, bj\}$. Now $\mathbf{p}$ (resp. $\mathbf{q}$) is the distribution resulting from the following process:

- Pick a uniformly random element $i \in [k']$ and a random sample j from $\mathbf{p}'$ (resp. $\mathbf{q}'$).

- Output, as resulting sample from $\mathbf{p}$ (resp. $\mathbf{q}$), the element $(i-1)k' + j$.

Since $\mathbf{p}'(i)$ and $\mathbf{q}'(i)$ are in $\{4/(5b), 6/(5b)\}$ for $i \in [b]$, we have that $\mathbf{p}(i)$ and $\mathbf{q}(i)$ are always in $\{4/(5bk'), 6/(5bk')\}$ for all $i \in [bk']$.

We will prove that $\mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}_\mathbf{q})$ is large. To facilitate this, recalling Fact 2.1 we will lower bound the distance of $\mathbf{p}$ to a binning of $\mathbf{q}$ in the following way:

$$\mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}_\mathbf{q}) \geq \frac{1}{2}\,\mathrm{dist}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\min_f \sum_{i=1}^{bk'} \Big| \mathbf{q}(i) - \sum_{t:f(t)=i} \mathbf{p}(t) \Big|, \tag{4}$$

where the minimum is taken over all nondecreasing functions $f \colon [bk'] \to [bk']$. To use the language of a partition[5], the interval I_i is simply $f^{-1}(i)$. We show the following:

Claim 4.8. *There exists $\varepsilon_0 > 0$ such that, for $\mathbf{p}$ and $\mathbf{q}$ as described above, $\mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}_\mathbf{q}) \geq \varepsilon_0$. (Moreover, one can take $\varepsilon_0 = 10^{-7}$.)*

Proof. We will show that $\mathrm{dist}(\mathbf{p}, \mathbf{q}) \geq 1/(5 \cdot 10^6)$. We first rewrite the distance as $\mathrm{dist}(\mathbf{p}, \mathbf{q}) = \min_f \sum_{j=1}^{k'} d(B_j)$, where, using the same blocks as described earlier, we define

$$d(B_j) := \sum_{i \in B_j} \Big| \mathbf{q}(i) - \sum_{t:f(t)=i} \mathbf{p}(t) \Big|$$

(suppressing its dependence on $\mathbf{p}, \mathbf{q}$, and f). We will call a block B_j *good* if $d(B_j) \leq 1/(2500k')$. Note that f need not be a function that maps elements of B_j to other elements of B_j. Since $\mathbf{p}(i)$ and $\mathbf{q}(i)$ are always in $\left\{ \frac{4}{5bk'}, \frac{6}{5bk'} \right\}$, any element $i \in B_j$ such that $|f^{-1}(i)| \neq 1$ will contribute at least $\frac{2}{5bk'}$ to $d(B_j)$. Thus, we call an element i *good* if there is a unique t such that $f(t) = i$ and $\mathbf{q}(i) = \mathbf{p}(t)$. Since each element that is not good contributes $\frac{2}{5bk'}$ to $d(B_j)$, it follows that, in a good block, at most $1/(2500k')/(\frac{2}{5bk'}) = b/1000$ elements are not good, and therefore a good block must contain at least $999b/1000$ good elements.

Let B_j be a good block: it must be the case that for at least $999b/1000$ values of $g \in B_j$, there exists a unique t such that $f(t) = g$. Let $g_1 < g_2 < \ldots < g_v$ be the good elements in B_j (where $v \geq 999b/1000$), and let $t_1, t_2, \ldots, t_v$ be such that $f(t_i) = g_i$ for $i = 1, 2, \ldots, v$. By assumption, the distributions $\mathbf{p}$ and $\mathbf{q}$ are not $(99b/100)$-partial cyclic shifts of each other. Thus, for every i' and i'' where $i' < i''$, if $t_{i''} - t_{i'} < b$, then $i'' - i' \leq 99b/100$, or else the good elements of B_j and their preimages under f would be a witness to $\mathbf{p}'$ and $\mathbf{q}'$ being $(99b/100)$-partial cyclic shifts of one another. It follows that

$$t_v - t_1 \geq b + (999b/1000 - 99b/100) \geq (1 + 1/1000)b.$$

---

[5]Recall that the sets in the partition are allowed to be empty.

That is, every good block "uses up" a bit more than b elements of the domain of f. Since $f\colon [bk'] \to [bk']$ must be a nondecreasing function, we can have at most $\dfrac{bk'}{(1+1/1000)b} \leq (1-1/2000)k'$ good blocks. Thus, at least $k'/2000$ of the blocks are not good. It follows that

$$\mathrm{dist}(\mathbf{p}, \mathbf{q}) = \min_f \sum_{j=1}^{k'} d(B_j) \geq (k'/2000)(1/2500k') = 1/(5 \cdot 10^6),$$

as claimed. $\qquad\square$

We are now in position to conclude the step (3) of our overall proof, establishing the following lower bound:

Theorem 4.9. *There exists an absolute constant $\varepsilon_0 > 0$ such that, for every sufficiently large integers $k, m \geq 1$, there exists $\mathbf{q}$ over $[k$ for which testing **Identity-Up-To-Binning**$(k, \mathbf{q}, k, \varepsilon_0)$ from ordered fingerprints only requires $\Omega\big((k/m^2 2^m)^{1-1/m}\big)$ samples. (Moreover, one can take $\varepsilon_0 = 10^{-7}$.)*

Proof. Let $\mathbf{p}$ and $\mathbf{q}$ be the distributions described above, with $k = bk'$ and $\varepsilon_0 := 10^{-7}$ as in Claim 4.8. We consider the problem **Identity-Up-To-Binning**$(k, \mathbf{q}, k, \varepsilon_0)$. We will show that any algorithm that takes too few samples, and only considers their ordered fingerprint, can distinguish between $\mathbf{p}$ and $\mathbf{q}$ with probability at most $1/3$. Since $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon_0 := 10^{-7}$ (and, of course, $d_{\mathrm{TV}}(\mathbf{q}, \mathbf{q}) = 0$), this will establish our lower bound.

By Lemma 4.7, an ordered fingerprint from m samples from one of the k' blocks of size b gives no information in distinguishing $\mathbf{p}$ and $\mathbf{q}$. Further, it is immediate to see, from the definition of ordered fingerprints, that the following holds as well:

Fact 4.10. *Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be two distributions over $[n]$, such that the distribution of fingerprints for $\mathbf{p}_b$ is given by $\{\boldsymbol{F}_j^b\}_{j \in [n]}$. Suppose there exists a partition of $[n]$ into $(I_1, I_2, \ldots, I_m)$ such that the sequences $\{\boldsymbol{F}_j^1\}_{j \in I_i}$ and $\{\boldsymbol{F}_j^2\}_{j \in I_i}$ are identically distributed for all i in $[m]$. Then $\mathbf{p}_1$ and $\mathbf{p}_2$ cannot be distinguished by ordered fingerprints.*

This implies that if an algorithm sees at most m samples from every block, then it cannot distinguish between $\mathbf{p}$ and $\mathbf{q}$ with ordered fingerprints from this sample. It follows that any algorithm that distinguishes between $\mathbf{p}$ and $\mathbf{q}$ using ordered fingerprints must receive at least $m + 1$ samples from at least one of the k' many blocks.

To conclude, note that the distribution over the blocks is uniform; thus, by folklore collision estimates (see also Suzuki et al. [STKT06]), the probability that one of the k' blocks contains $m + 1$ samples is at most $1/3$, given $O(k'^{1-1/(m+1)})$ samples. Since $k' = k/b = k/(5m^2 2^m)$, we get the claimed lower bound. $\qquad\square$

Setting $m = \sqrt{\log k}$ in Theorem 4.9, we get:

Corollary 4.11. *There exists an absolute constant $\varepsilon_0 > 0$ such that, for every sufficiently large integers $k \geq 1$, there exists $\mathbf{q}$ over $[k$ for which testing **Identity-Up-To-Binning**$(k, \mathbf{q}, k, \varepsilon_0)$ from ordered fingerprints only requires $\dfrac{k}{2^{O(\sqrt{\log k})}}$ samples.*

References

[BCG17]   Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *32nd Computational Complexity Conference*, volume 79 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 28, 40. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2017. 4.3

[BCG19]   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. Journal version of [BCG17]. 1

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 442–451. IEEE Computer Soc., Los Alamitos, CA, 2001. 1

[BFR+00] 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. 1

[BW18]   Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: a selective review. *Ann. Appl. Stat.*, 12(2):727–749, 2018. 1

[Can15]   Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015. To appear as a graduate student survey in Theory of Computing. 1

[Can19]   Clément L. Canonne. Problem 96: Identity testing up to coarsenings, 2019. Asked at the 3rd Workshop on Local Algorithms (WOLA 2019). (document), 1

[CDSS12] Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1380–1394. SIAM, Philadelphia, PA, 2012. 3, 3

[CFS10]   David Conlon, Jacob Fox, and Benny Sudakov. Hypergraph Ramsey numbers. *J. Amer. Math. Soc.*, 23(1):247–266, 2010. 4.5

[DGPP18] 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. 1

[DK16]   Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 685–694. IEEE Computer Soc., Los Alamitos, CA, 2016. 1

[DKN15a] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 1183–1202. IEEE Computer Soc., Los Alamitos, CA, 2015. 1.1, 3, (1), 4.1, 4.1

[DKN15b] 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. 1, 3

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York, 2001. 3, 3

[GGR98] 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. 1

[Gol20] 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. 1

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000. 1

[Ngu12] Andy Nguyen. Solving cyclic longest common subsequence in quadratic time. *CoRR*, abs/1208.0396, 2012. 4.2

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008. 1

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996. 1

[STKT06] 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. 1.1, 4.3

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014*, pages 51–60. IEEE Computer Soc., Los Alamitos, CA, 2014. 4.3

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. Journal version of [VV14]. 1