2016 IEEE 57th Annual Symposium on Foundations of Computer Science

A New Approach for Testing Properties of Discrete Distributions

Ilias Diakonikolas
CS
University of Southern California
Los Angeles, CA, USA
diakonik@usc.edu

Daniel M. Kane
CSE & Math
University of California San Diego
La Jolla, CA, USA
dakane@cs.ucsd.edu

***Abstract*—We study problems in distribution property testing: Given sample access to one or more unknown discrete distributions, we want to determine whether they have some global property or are epsilon-far from having the property in L1 distance (equivalently, total variation distance, or "statistical distance"). In this work, we give a novel general approach for distribution testing. We describe two techniques: our first technique gives sample–optimal testers, while our second technique gives matching sample lower bounds. As a consequence, we resolve the sample complexity of a wide variety of testing problems.**

Our upper bounds are obtained via a modular reduction-based approach. Our approach yields optimal testers for numerous problems by using a standard L2-identity tester as a black-box. Using this recipe, we obtain simple estimators for a wide range of problems, encompassing many problems previously studied in the TCS literature, namely: (1) identity testing to a fixed distribution, (2) closeness testing between two unknown distributions (with equal/unequal sample sizes), (3) independence testing (in any number of dimensions), (4) closeness testing for collections of distributions, and (5) testing histograms. For all of these problems, our testers are sample-optimal, up to constant factors. With the exception of (1), ours are the *first sample-optimal testers for the corresponding problems.* Moreover, our estimators are significantly simpler to state and analyze compared to previous results.

As an important application of our reduction-based technique, we obtain the first *adaptive* algorithm for testing equivalence between two *unknown* distributions. The sample complexity of our algorithm depends on the *structure of the unknown distributions* – as opposed to merely their domain size – and is significantly better compared to the worst-case optimal L1-tester in many natural instances. Moreover, our technique naturally generalizes to other metrics beyond the L1-distance. As an illustration of its flexibility, we use it to obtain the first near-optimal equivalence tester under the Hellinger distance.

Our lower bounds are obtained via a direct information-theoretic approach: Given a candidate hard instance, our proof proceeds by bounding the mutual information between appropriate random variables. While this is a classical method in information theory, prior to our work, it had not been used in this context. Previous lower bounds relied either on the birthday paradox, or on moment-matching and were thus restricted to symmetric properties. Our lower bound approach does not suffer from any such restrictions and gives tight sample lower bounds for the aforementioned problems.

***Keywords*-distribution testing, property testing, hypothesis testing**

I. Introduction

A. Background

The problem of determining whether an unknown object fits a model based on observed data is of fundamental scientific importance. We study the following formalization of this problem: Given samples from a collection of probability distributions, can we determine whether the distributions in question satisfy a certain property? This is the prototypical question in *statistical hypothesis testing* [1], [2]. During the past two decades, this question has received considerable attention by the TCS community in the framework of *property testing* [3], [4], with a focus on discrete probability distributions.

The area of distribution property testing [5], [6] has developed into a mature research field with connections to information theory, learning and statistics. The generic inference problem in this field is the following: given sample access to one or more unknown distributions, determine whether they have some global property or are "far" (in statistical distance or, equivalently, ℓ_1 norm) from having the property. The goal is to obtain statistically and computationally efficient testing algorithms, i.e., algorithms that use the information-theoretically minimum sample size and run in polynomial time. See [7], [5], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] for a sample of works and [24], [25] for two recent surveys.

In this work, we give a new general approach for distribution testing. We describe two novel techniques: our first technique yields sample–optimal testers, while our second technique gives matching sample lower bounds. As a consequence, we resolve the sample complexity of a wide variety of testing problems.

All our upper bounds are obtained via a collection of modular *reductions*. Our reduction-based method provides a simple recipe to obtain *optimal* testers under the ℓ_1-norm (and other metrics), by applying a randomized transformation to a basic ℓ_2-identity tester. While the ℓ_2-norm has been used before as a tool in distribution testing [5], our reduction-based approach is conceptually and technically different than previous approaches. We elaborate on this point in Section I-C. We use our reduction-based approach

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.78

to resolve a number of open problems in the literature (see Section I-B). In addition to pinning–down the sample complexity of a wide range of problems, a key contribution of our algorithmic approach is methodological. *In particular, the main conceptual message is that one does not need an inherently different statistic for each testing problem.* In contrast, all our testing algorithms follow the same pattern: They are obtained by applying a simple transformation to a basic statistic – one that tests the identity between two distributions in ℓ_2-norm – in a black-box manner. Following this scheme, we obtain the first sample-optimal testers for many properties. Importantly, our testers are simple and, in most cases, their analysis fits in a paragraph.

As our second main contribution, we provide a direct, elementary approach to prove sample complexity lower bounds for distribution testing problems. Given a candidate hard instance, our proof proceeds by bounding the mutual information between appropriate random variables. Our analysis leads to new, optimal lower bounds for several problems, including testing closeness (under various metrics), testing independence (in any dimension), and testing histograms. Notably, proving sample complexity lower bounds by bounding the mutual information is a classical approach in information theory. Perhaps surprisingly, prior to our work, this method had not been used in distribution testing. Previous techniques were either based on the birthday paradox or on moment-matching [26], [13], and were thus restricted to testing symmetric properties. Our technique circumvents the moment-matching approach, and is not restricted to symmetric properties.

B. Our Contributions

The main contribution of this paper is a reduction–based framework to obtain testing algorithms, and a direct approach to prove lower bounds. We do not aim to exhaustively cover all possible applications of our techniques, but rather to give some selected results that are indicative of the generality and power of our methods. More specifically, we obtain the following results:

1) We give an alternative optimal ℓ_1-identity tester against a fixed distribution, with sample complexity $O(\sqrt{n}/\epsilon^2)$, matching the recently obtained tight bound [19], [20]. The main advantage of our tester is its simplicity: Our reduction and its analysis are remarkably short and simple in this case. Our tester straightforwardly implies the "χ^2 versus ℓ_1" guarantee recently used as the main statistical test in [22].

2) We design an optimal tester for ℓ_1-closeness between two unknown distributions in the standard and the (more general) unequal-sized sample regimes. For the standard regime (i.e., when we draw the same number of samples from each distribution), we recover the tight sample complexity of $O(\max(n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2))$, matching [18]. Importantly, our tester straightforwardly extends to unequal-sized samples, giving the first optimal tester in this setting. Closeness testing with unequal sized samples was considered in [27] that gives sample upper and lower bounds with a polynomial gap between them. Our tester uses $m_1 = \Omega(\max(n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2))$ samples from one distribution and $m_2 = O(\max(nm_1^{-1/2}/\epsilon^2, \sqrt{n}/\epsilon^2))$ from the other. This tradeoff is sample-optimal (up to a constant factor) for all settings, and improves on the recent work [28] that obtains the same tradeoff under the additional assumption that $\epsilon > n^{-1/12}$. In sharp contrast to [28], our algorithm is extremely simple and its analysis fits in a few lines.

3) We study the problem of ℓ_1-testing closeness between two *unknown* distributions in an *adaptive* setting, where the goal is to design estimators whose sample complexity depends on the *(unknown) structure of the sampled distributions* – as opposed to merely their domain size. We obtain the first algorithm for this problem: Our tester uses $O(\mathrm{polylog}(n/\epsilon) \cdot \min_{m>0}(m + \|q^{<1/m}\|_0 \cdot \|q^{<1/m}\|_2/\epsilon^2 + \|q\|_{2/3}/\epsilon^2))$ samples from each of the distributions p, q on $[n]$. Here, $q^{<1/m}$ denotes the pseudo-distribution obtained from q by removing the domain elements with mass $\geq 1/m$, and $\|q^{<1/m}\|_0$ is the number of elements with mass $< 1/m$. There are a few remarks to be made about the sample complexity of this algorithm. To begin with, note that since $\|q^{<1/m}\|_2 \leq 1/\sqrt{m}$, taking $m = m_0 := \min(n, n^{2/3}/\epsilon^{4/3})$ attains the sample complexity of the standard ℓ_1-closeness testing algorithm to within logarithmic factors. However, unlike the standard ℓ_1-closeness testing algorithm, our algorithm will *only* have this kind of complexity, if q has approximately m_0 bins of mass approximately $1/m_0$ and approximately n smaller bins, a situation which seems unlikely to occur in natural settings when $m_0 \ll n$. In fact, the $\|q\|_{2/3}/\epsilon^2$ term in the sample complexity makes our tester comparable to the instance-optimal tester from [19]. In particular, [19] give an identity tester against an explicit distribution q that has essentially the best possible sample complexity of any tester for that q. This sample complexity (for a broad range of q and ϵ) is proportional to $\|q\|_{2/3}/\epsilon^2$. Our tester achieves this term in its sample complexity without knowing q ahead of time.

4) We show that our framework easily generalizes to give near-optimal algorithms and lower bounds for other metrics as well, beyond the ℓ_1-norm. As an illustration of this fact, we describe an algorithm and a nearly-matching lower bound for testing closeness under Hellinger distance, $H^2(p, q) = (1/2)\|\sqrt{p} - \sqrt{q}\|_2^2$, one of the most powerful f-divergences. This question has been studied before: [29] gave a tester for this problem with sample complexity $\tilde{O}(n^{2/3}/\epsilon^4)$. The sample complexity of our algorithm is $\tilde{O}(\min(n^{2/3}/\epsilon^{4/3}, n^{3/4}/\epsilon))$, and we prove a lower bound of $\Omega(\min(n^{2/3}/\epsilon^{4/3}, n^{3/4}/\epsilon))$. Note that the second term of $n^{3/4}/\epsilon$ in the sample complexity differs from the corresponding ℓ_1 term of $n^{1/2}/\epsilon^2$.

5) We obtain the first sample-optimal algorithm and

matching lower bound for testing independence over $\times_{i=1}^{d}[n_i]$. Prior to our work, the sample complexity of this problem remained open, even for the two-dimensional case. We prove that the *optimal* sample complexity of independence testing (upper and lower bound) is $\Theta(\max_j((\prod_{i=1}^{d} n_i)^{1/2}/\epsilon^2, n_j^{1/3}(\prod_{i=1}^{d} n_i)^{1/3}/\epsilon^{4/3}))$. Previous testers for independence were suboptimal up to polynomial factors in n and $1/\epsilon$, even for $d = 2$. Specifically, Batu *et al.* [8] gave an independence tester over $[n] \times [m]$ with sample complexity $\widetilde{O}(n^{2/3}m^{1/3}) \cdot \mathrm{poly}(1/\epsilon)$, for $n \geq m$. On the lower bound side, Levi, Ron, and Rubinfeld [16] showed a sample complexity lower bound of $\Omega(\sqrt{nm})$ (for all $n \geq m$), and $\Omega(n^{2/3}m^{1/3})$ (for $n = \Omega(m \log m)$). More recently, Acharya *et al.* [22] gave an upper bound of $O(((\prod_{i=1}^{d} n_i)^{1/2} + \sum_{i=1}^{d} n_i)/\epsilon^2)$, which is optimal up to constant factors for the very special case that all the n_i's are the same. In summary, we resolve the sample complexity of this problem in any dimension d, up to a constant factor, as a function of all relevant parameters.

6) We obtain the first sample-optimal algorithms for testing equivalence for collections of distributions [16] in the sampling and the oracle model, improving on [16] by polynomial factors. In the sampling model, we observe that the problem is equivalent to (a variant of) two-dimensional independence testing. In fact, in the unknown-weights case, the problem is identical. In the known-weights case, the problem is equivalent to two-dimensional independence testing, where the algorithm is given explicit access to one of the marginals (say, the marginal on $[m]$). For this setting, we give a sample-optimal tester with sample size $O(\max(\sqrt{nm}/\epsilon^2, n^{2/3}m^{1/3}/\epsilon^{4/3}))$[1]. In the query model, we give a sample-optimal closeness tester for m distributions over $[n]$ with sample complexity $O(\max(\sqrt{n}/\epsilon^2, n^{2/3}/\epsilon^{4/3}))$. This bound is independent of m and matches the worst-case optimal bound for testing closeness between two unknown distributions.

7) As a final application of our techniques, we study the problem of testing whether a distribution belongs in a given "structured family" [22], [23], [30]. We focus on the property of being a k-histogram over $[n]$, i.e., that the probability mass function is piecewise constant with at most k *known* interval pieces. This is a natural problem of particular interest in model selection. For $k = 1$, the problem is tantamount to uniformity testing, while for $k = \Omega(n)$ it can be seen to be equivalent to testing closeness between two unknown distributions over a domain of size $\Omega(n)$. We design a tester for the property of being a k-histogram (with respect to a given set of intervals) with sample complexity $O(\max(\sqrt{n}/\epsilon^2, n^{1/3}k^{1/3}/\epsilon^{4/3}))$ samples. We also prove that this bound is information-theoretically optimal, up to constant factors. In concurrent work, Canonne [30] obtained a nearly-optimal tester for the harder setting where the k intervals are unknown.

[1]It should be noted that, while this is the same form as the sample complexity for independence testing in two dimensions, there is a crucial difference. In this setting, the parameter m represents the support size of the marginal that is explicitly given to us, rather than the marginal with smaller support size.

C. Prior Techniques and Overview of our Approach

In this section, we provide a detailed intuitive explanation of our two techniques, in tandem with a comparison to previous approaches. We start with our upper bound approach. It is reasonable to expect that the ℓ_2-norm is useful as a tool in distribution property testing. Indeed, for elements with "small" probability mass, estimating second moments is a natural choice in the sublinear regime. Alas, a direct ℓ_2-tester will often not work for the following reason: The error coming from the "heavy" elements will force the estimator to draw too many samples.

In their seminal paper, Batu *et al.* [5], [6] gave an ℓ_2-closeness tester and used it to obtain an ℓ_1-closeness tester. To circumvent the aforementioned issue, their ℓ_1-tester has two stages: It first explicitly learns the pseudo-distribution supported on the heavy elements, and then it applies the ℓ_2-tester on the pseudo-distribution over the light elements. This approach of combining learning (for the heavy elements) and ℓ_2-closeness testing (for the light elements) is later refined by Chan *et al.* [18], where it is shown that it *inherently* leads to a suboptimal sample complexity for the testing closeness problem. Motivated by this shortcoming, it was suggested in [18] that the use of the ℓ_2-norm may be insufficient, and that a more direct approach may be needed to achieve sample-optimal ℓ_1-testers. This suggestion led researchers to consider different approaches to ℓ_1-testing, in particular appropriately rescaled versions of the chi-squared test [31], [18], [19], [28], [22]. This line of work has led to sample-optimal testers for closeness testing [18] and identity testing [19], [22]. A major difference between the explicit rescaling performed by chi-squared testers and our reduction-based framework is that the former approach seems to require a new algorithm with a highly-nontrivial analysis for each particular testing problem.

Our upper bound approach postulates that the inefficiency of [5], [6] is due to the explicit learning of the heavy elements and not to the use of the ℓ_2-norm. Our approach provides a simple and general way to essentially remove this learning step. We achieve this via a collection of simple *reductions*: Starting from a given instance of an ℓ_1-testing problem $\mathcal{A}$, we construct a new instance of an appropriate ℓ_2-testing problem $\mathcal{B}$, so that the answers to the two problems for these instances are identical. Here, problem $\mathcal{A}$ can be *any* of the testing problems discussed in Section I-B, while problem $\mathcal{B}$ is always the same. Namely, we define $\mathcal{B}$ to be the problem of ℓ_2-testing closeness between two unknown distributions, under the promise that *at least one* of the distributions in question has small ℓ_2-norm. Our reductions have the property that a sample-optimal algorithm for problem

$\mathcal{B}$ implies a (nearly) sample-optimal algorithm for $\mathcal{A}$. An important conceptual consequence of our direct reduction-based approach is that problem $\mathcal{B}$ is of central importance in distribution testing, since a wide range of problems can be reduced to it with optimal sample guarantees. We remark that sample-optimal algorithms for problem $\mathcal{B}$ are known in the literature: a natural estimator from [18], as well as a similar estimator from [5] achieve optimal bounds.

The precise form of our reductions naturally depends on the problem $\mathcal{A}$ that we start from. While the details differ based on the problem, all our reductions rely on a common recipe: We randomly transform the initial distributions in question (i.e., the distributions we are given sample access to) to new distributions (over a potentially larger domain) such that at least one of the *new* distributions has appropriately small ℓ_2-norm. Our transformation preserves the ℓ_1-norm, and is such that we can easily simulate samples from the new distributions. More specifically, our transformation is obtained by drawing random samples from one of the distributions in question to discover its heavy bins. We then artificially subdivide each heavy bin into multiple bins, so that the resulting distribution becomes approximately flat. This procedure decreases the ℓ_2-norm while increasing the domain size. By balancing these two quantities, we obtain sample-optimal testers for a wide variety of properties.

We note that our technique is at a high level similar to the approaches employed in [20], [21]. Both techniques relate an "ℓ_1-type" testing problem to an ℓ_2-testing problem (or a collection of ℓ_2-testing problems) on a suitably transformed domain. However, this is where the similarities end. In [20] and [21], the primary obstacle is the potentially huge domain size, and the approach was to take advantage of structure in the underlying distributions in order to reduce the ℓ_1-testing problem to an ℓ_2-testing problem on a notably *smaller* domain. In this work, our transformations are primarily designed to deal with the problem that standard ℓ_2-testers perform poorly if the distributions involved have large ℓ_2 norm. Thus, our transformations produce distributions on a *larger* domain in order to appropriately flatten them.

In summary, our upper bound approach provides reductions of numerous distribution testing problems to a specific ℓ_2-testing problem $\mathcal{B}$ that yield sample-optimal algorithms. It is tempting to conjecture that optimal reductions in the opposite direction exist, which would allow translating lower bounds for problem $\mathcal{B}$ to tight lower bounds for other problems. We do not expect optimal reductions in the opposite direction, roughly because the hard instances for many of our problems are substantially different from the hard instances for problem $\mathcal{B}$. This naturally brings us to our lower bound approach, explained below.

Our lower bounds proceed by constructing explicit distributions $\mathcal{D}$ and $\mathcal{D}'$ over (sets of) distributions, so that a random distribution p drawn from $\mathcal{D}$ satisfies the property, a random distribution p from $\mathcal{D}'$ is far from satisfying the property (with high probability), and it is hard to distinguish between the two cases given a small number of samples. Our analysis is based on classical information-theoretic notions and is significantly different from previous approaches in this context. Instead of using techniques involving matching moments [26], [13], we are able to directly prove that the mutual information between the set of samples drawn and the distribution that p was drawn from is small. Appropriately bounding the mutual information is perhaps a technical exercise, but remains quite manageable only requiring elementary approximation arguments. We believe that this technique is more flexible than the techniques of [26], [13] (e.g., it is not restricted to symmetric properties), and may prove useful in future testing problems.

Remark I.1. Our approach provides a unifying framework to obtain tight bounds for distribution testing problems. Since the dissemination of an earlier version of our paper, Oded Goldreich gave an excellent exposition of our approach in his upcoming book [32].

D. Organization

In Section II, we describe our reduction-based approach and exploit it to obtain optimal testers for a variety of problems. In Section III, we describe our lower bound approach and apply it to prove tight lower bounds for various problems. Due to space constraints, many proofs are deferred to the full version.

II. Our Reduction and its Algorithmic Applications

In Section II-A, we describe our basic reduction from ℓ_1 to ℓ_2 testing. In Section II-B, we apply our reduction to a variety of concrete distribution testing problems.

A. Reduction of ℓ_1-testing to ℓ_2-testing

The starting point of our reduction-based approach is a "basic tester" for the identity between two unknown distributions with respect to the ℓ_2-norm. We emphasize that a simple and natural tester turns out to be optimal in this setting. More specifically, we will use the following simple lemma (that follows, e.g., from Proposition 3.1 in [18]):

Lemma II.1. *Let p and q be two unknown distributions on $[n]$. There exists an algorithm that on input n, $\epsilon > 0$, and $b \geq \max\{\|p\|_2, \|q\|_2\}$ draws $O(bn/\epsilon^2)$ samples from each of p and q, and with probability at least $2/3$ distinguishes between the cases that $p = q$ and $\|p - q\|_1 > \epsilon$.*

Remark II.2. We remark that Proposition 3.1 of [18] provides a somewhat stronger guarantee than the one of Lemma II.1. Specifically, it yields a *tolerant* ℓ_2-closeness tester with the following performance guarantee: Given $O(bn/\epsilon^2)$ samples from distributions p, q over $[n]$, where $b \geq \max\{\|p\|_2, \|q\|_2\}$, the algorithm distinguishes (with probability at least $2/3$) between the cases that $\|p - q\|_2 \leq$

$\epsilon/(2\sqrt{n})$ and $\|p-q\|_2 \geq \epsilon/\sqrt{n}$. The soundness guarantee of Lemma II.1 follows from the Cauchy-Schwarz inequality.

Observe that if $\|p\|_2$ and $\|q\|_2$ are both small, the algorithm of Lemma II.1 is in fact sample-efficient. For example, if both are $O(1/\sqrt{n})$, its sample complexity is an optimal $O(\sqrt{n}/\epsilon^2)$. On the other hand, the performance of this algorithm degrades as $\|p\|_2$ or $\|q\|_2$ increases. Fortunately, there are some simple reductions to circumvent this issue. To begin with, we note that it suffices that only one of $\|p\|_2$ and $\|q\|_2$ is small. This is essentially because if there is a large difference between the two, this is easy to detect.

Lemma II.3. *Let p and q be two unknown distributions on $[n]$. There exists an algorithm that on input n, $\epsilon > 0$, and $b \geq \min\{\|p\|_2, \|q\|_2\}$ draws $O(bn/\epsilon^2)$ samples from each of p and q and, with probability at least $2/3$, distinguishes between the cases that $p = q$ and $\|p-q\|_1 > \epsilon$.*

Proof: The basic idea is to first test if $\|p\|_2 = \Theta(\|q\|_2)$, and if so to run the tester of Lemma II.1. To test whether $\|p\|_2 = \Theta(\|q\|_2)$, we estimate $\|p\|_2$ and $\|q\|_2$ up to a multiplicative constant factor. It is known [7], [8] that this can be done with $O(\sqrt{n}) = O(\min(\|p\|_2, \|q\|_2)n)$ samples. If $\|p\|_2$ and $\|q\|_2$ do not agree to within a constant factor, we can conclude that $p \neq q$. Otherwise, we use the tester from Lemma II.1, and note that the number of required samples is $O(\|p\|_2 n/\epsilon^2)$. ∎

In our applications of Lemma II.3, we take the parameter b to be equal to our upper bound on $\min\{\|p\|_2, \|q\|_2\}$. In all our algorithms in Section II-B this upper bound will be clear from the context. If both our initial distributions have large ℓ_2-norm, we describe a new way to reduce the ℓ_2-norm of at least one of them by splitting the large weight bins (domain elements) into pieces. The following key definition is the basis for our reduction:

Definition II.4. Given a distribution p on $[n]$ and a multiset S of elements of $[n]$, define the *split distribution* p_S on $[n+|S|]$ as follows: For $1 \leq i \leq n$, let a_i equal 1 plus the number of elements of S that are equal to i. Thus, $\sum_{i=1}^n a_i = n + |S|$. We can therefore associate the elements of $[n+|S|]$ to elements of the set $B = \{(i,j) : i \in [n], 1 \leq j \leq a_i\}$. We now define a distribution p_S with support B, by letting a random sample from p_S be given by (i,j), where i is drawn randomly from p and j is drawn randomly from $[a_i]$.

We now point out two basic facts about split distributions:

Fact II.5. *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then: (i) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively. (ii) It holds $\|p_S - q_S\|_1 = \|p-q\|_1$.*

Fact II.5 implies that it suffices to be able to test the closeness of p_S and q_S, for some S. In particular, we want to find an S so that $\|p_S\|_2$ and $\|q_S\|_2$ are small. The following lemma shows how to achieve this:

Lemma II.6. *Let p be a distribution on $[n]$. Then: (i) For any multisets $S \subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \leq \|p_S\|_2$, and (ii) If S is obtained by taking $\mathrm{Poi}(m)$ samples from p, then $\mathbb{E}[\|p_S\|_2^2] \leq 1/m$.*

Proof: Let a_i equal one plus the number of copies of i in S, and a'_i equal one plus the number of copies of i in S'. We note that $p_S = (i,j)$ with probability p_i/a_i. Therefore, for (i) we have that

$$\|p_S\|_2^2 = \sum_{i=1}^n \sum_{j=1}^{a_i} (p_i/a_i)^2 = \sum_{i=1}^n p_i^2/a_i \geq \sum_{i=1}^n p_i^2/a'_i = \|p_{S'}\|_2^2.$$

For claim (ii), we note that the expected squared ℓ_2-norm of p_S is $\sum_{i=1}^n p_i^2 \mathbb{E}[a_i^{-1}]$. We note that a_i is distributed as $1+X$ where X is a $\mathrm{Poi}(mp_i)$ random variable. Recall that if Y is a random variable distributed as $\mathrm{Poi}(\lambda)$, then $\mathbb{E}[z^Y] = e^{\lambda(z-1)}$. Taking an integral we find that

$$\begin{aligned} \mathbb{E}\left[1/(1+X)\right] &= \mathbb{E}\left[\int_0^1 z^X dz\right] = \int_0^1 \mathbb{E}[z^X] dz = \int_0^1 e^{\lambda(z-1)} dz \\ &= (1-e^{-\lambda})/\lambda \leq 1/\lambda. \end{aligned}$$

Therefore, we have that $\mathbb{E}[\|p_S\|_2^2] \leq \sum_{i=1}^n p_i^2/(mp_i) = (1/m)\sum_{i=1}^n p_i = 1/m$. This completes the proof. ∎

B. Algorithmic Applications

1) Testing Identity to a Known Distribution: We start by applying our framework to give a simple alternate optimal identity tester to a fixed distribution in the minimax sense. In this case, our algorithm is extremely easy, and provides a much simpler proof of the known optimal bound [19], [20]:

Proposition II.7. *There exists an algorithm that given an explicit distribution q supported on $[n]$ and $O(\sqrt{n}/\epsilon^2)$ independent samples from a distribution p over $[n]$ distinguishes with probability at least $2/3$ between the cases where $p = q$ and $\|p-q\|_1 \geq \epsilon$.*

Proof: Let S be the multiset where S contains $\lfloor nq_i \rfloor$ copies of i. Note that $|S| \leq \sum_{i=1}^n nq_i = n$. Note also that q_S assigns probability mass at most $1/n$ to each bin. Therefore, we have that $\|q_S\|_2 = O(1/\sqrt{n})$. It now suffices to distinguish between the cases that $p_S = q_S$ and the case that $\|p_S - q_S\|_1 \geq \epsilon$. Using the basic tester from Lemma II.3 for $b = O(1/\sqrt{n})$, we can do this using $O(2nb/\epsilon^2) = O(\sqrt{n}/\epsilon^2)$ samples from p_S. This can be simulated using $O(\sqrt{n}/\epsilon^2)$ samples from p, which completes the proof. ∎

Remark II.8. It is easy to see that the identity tester of Proposition II.7 satisfies a stronger guarantee: More specifically, it distinguishes between the cases that $\chi^2(p,q) := \sum_{i=1}^n (p_i - q_i)^2/q_i \leq \epsilon^2/10$ versus $\|p-q\|_1 \geq \epsilon$. Hence, it implies Theorem 1 of [22]. See the full version for an explanation.

Remark II.9. After the dissemination of an earlier version of this paper, inspired by our work, Goldreich [33] reduced testing identity to a fixed distribution to its special case of uniformity testing, via a refinement of the above idea. This elegant idea does not seem to generalize to other problems considered here.

2) Testing Closeness: We now turn to the problem of testing closeness between two unknown distributions p, q. The difficulty of this case lies in the fact that, not knowing q, we cannot subdivide into bins in such a way as to guarantee that $\|q_S\|_2 = O(1/\sqrt{n})$. However, we can do nearly as well by first drawing an appropriate number of samples from q, and then using them to provide our subdivisions. In particular, we want to divide heavier bins more times, so we will split a bin a number of times given by the number of samples drawn from it in an initial step.

Proposition II.10. *There exists an algorithm that given sample access to two distributions p and q over $[n]$ distinguishes with probability $2/3$ between the cases $p = q$ and $\|p - q\|_1 > \epsilon$ using $O(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2))$ samples from each of p and q.*

Proof: The algorithm is as follows:

Algorithm Test-Closeness

Input: Sample access to distributions p and q supported on $[n]$ and $\epsilon > 0$.
Output: "YES" with probability at least $2/3$ if $p = q$, "NO" with probability at least $2/3$ if $\|p - q\|_1 \geq \epsilon$.

1) Let $k = \min(n, n^{2/3}\epsilon^{-4/3})$.
2) Define a multiset S by taking $\mathrm{Poi}(k)$ samples from q.
3) Run the tester from Lemma II.3 to distinguish between $p_S = q_S$ and $\|p_S - q_S\|_1 \geq \epsilon$.

To show correctness, we first note that with high probability we have $|S| = O(n)$. Furthermore, by Lemma II.6 it follows that the expected squared ℓ_2 norm of q_S is at most $1/k$. Therefore, with probability at least $9/10$, we have that $|S| = O(n)$ and $\|q_S\|_2 = O(1/\sqrt{k})$.

The tester from Lemma II.3 distinguishes between $p_S = q_S$ and $\|p_S - q_S\|_1 \geq \epsilon$ with $O(nk^{-1/2}/\epsilon^2)$ samples. By Fact II.5, this is equivalent to distinguishing between $p = q$ and $\|p - q\|_1 \geq \epsilon$. Thus, the total number of samples taken by the algorithm is $O(k + nk^{-1/2}/\epsilon^2) = O(\max(n^{2/3}\epsilon^{-4/3}, \sqrt{n}/\epsilon^2))$. ∎

We consider a generalization of testing closeness where we have access to different size samples from the two distributions, and use our technique to provide the first sample-optimal algorithm for the entire range of parameters:

Proposition II.11. *There exists an algorithm that given sample access to two distributions, p and q over $[n]$ distinguishes with probability $2/3$ between the cases $p = q$ and $\|p - q\|_1 > \epsilon$ given m_1 samples from q and an additional $m_2 = O(\max(nm_1^{-1/2}/\epsilon^2, \sqrt{n}/\epsilon^2))$ samples from each of p and q.*

The basic idea of this algorithm is the same as above, except that if $m_1 \gg m_2$, we can use m_1 samples from q to flatten it more efficiently. See the full version for the details.

3) Adaptive Testing: In this subsection, we provide near-optimal testers for identity and closeness in the adaptive setting. We start with the simpler case of testing identity to a fixed distribution. This serves as a warm-up for the more challenging case of two unknown distributions.

Note that the identity tester of Proposition II.7 is sample-optimal only for a worst-case choice of the explicit distribution q. (It turns out that the worst case corresponds to q being the uniform distribution over $[n]$.) Intuitively, for most choices of q, one can actually do substantially better. This fact was first formalized and shown in [19], where it is shown that $\Theta(\|q\|_{2/3}/\epsilon^2)$ samples are optimal in most cases.

In the following proposition, we give a very simple tester with a compact analysis whose sample complexity is essentially optimal as a function of q. The basic idea of our tester is the following: First, we partition the domain into categories based on the approximate mass of the elements of q, and then we run an ℓ_2-tester independently on each category. See the full version for the details.

Proposition II.12. *There exists an algorithm that on input an explicit distribution q over $[n]$, a parameter $\epsilon > 0$, and $O(\mathrm{polylog}(n/\epsilon)\|q\|_{2/3}/\epsilon^2)$ samples from a distribution p over $[n]$ distinguishes with probability at least $2/3$ between the cases where $p = q$ and $\|p - q\|_1 \geq \epsilon$.*

We now show how to use our reduction-based approach to obtain the first nearly adaptive algorithm for testing closeness between two unknown distributions. Note that the algorithm of Proposition II.12 crucially exploits the a priori knowledge of the explicit distribution. In the setting where both distributions are unknown, this is no longer possible. At a high-level, our adaptive closeness testing algorithm is similar to that of Proposition II.12: We start by partitioning $[n]$ into categories based on the approximate mass of one of the two unknown distributions, say q, and then we run an ℓ_2-tester independently on each category. A fundamental difficulty in our setting is that q is unknown. Hence, to achieve this, we will need to take samples from q and create categories based on the number of samples coming from each bin. To state our result, we need the following notation:

Definition II.13. Let q be a discrete distribution and $x > 0$. We denote by $q^{<x}$ the pseudo-distribution obtained from q by setting the probabilities of all domain elements with probability at least x to 0.

The main result of this subsection is the following:

Proposition II.14. *Given sample access to two unknown*

distributions, p, q over $[n]$ and $\epsilon > 0$, there exists a computationally efficient algorithm that draws an expected

$$O(\mathrm{polylog}(n/\epsilon) \min_{m>0}(m+\|q^{<1/m}\|_0\|q^{<1/m}\|_2/\epsilon^2+\|q\|_{2/3}/\epsilon^2))$$

samples from each of p and q, and distinguishes with probability $2/3$ between $p = q$ and $\|p - q\|_1 \geq \epsilon$.

The proof of Proposition II.14 is deferred to the full version. Note that since $\|q^{<1/m}\|_2 \leq 1/\sqrt{m}$, taking $m = \min(n, n^{2/3}/\epsilon^{4/3})$ attains the complexity of the standard ℓ_1-closeness testing algorithm to within logarithmic factors.

We now illustrate with a number of examples that the algorithm of Proposition II.14 performs substantially better than the worst-case optimal ℓ_1-closeness tester in a number of interesting cases. First, consider the case that the distribution q is essentially supported on relatively heavy bins. It is easy to see that the sample complexity of our algorithm will then be roughly proportional to $\|q\|_{2/3}/\epsilon^2$. We remark that this bound is essentially optimal, even for the easier setting that q had been given to us explicitly. As a second example, consider the case that q is roughly uniform. In this case, we have that $\|q\|_2$ will be small, and our algorithm will have sample complexity $\tilde{O}(\sqrt{n}/\epsilon^2)$.

Finally, consider the case that the bins of the distribution q can be partitioned into two classes: they have mass either approximately $1/n$ or approximately $x > 1/n$. For this case, our above algorithm will need $\tilde{O}(\min(x^{-1} + \sqrt{n}/\epsilon^2, nx^{-1/2}/\epsilon^2))$ samples. We remark that this sample bound can be shown to be optimal for such distributions (up to the logarithmic factor in the $\tilde{O}$). Also note that the aforementioned sample upper bound is strictly better than the worst-case bound of $n^{2/3}/\epsilon^{4/3}$, unless x equals $n^{-2/3}\epsilon^{4/3}$.

Using ideas similar to those in our adaptive closeness tester, our reduction-based approach yields a nearly sample-optimal algorithm for testing closeness of two unknown distributions with respect to the Hellinger distance. See the full version for the details.

4) Independence Testing: In this subsection we study the problem of testing independence of a d-dimensional discrete distribution p. We start by giving an optimal independence tester for the two-dimensional case, and then handle the case of arbitrary dimension.

The basic idea of our algorithm is as follows: Let q be the product of the marginal distributions of p. We want to test whether or not $p = q$ or $\|p - q\|_1 > \epsilon$. We do this in our standard way, by first flattening q and then using an appropriate ℓ_2 tester. However, because q is a *product* distribution, we can flatten it more efficiently by flattening its marginal distributions.

Our algorithm for testing independence in two dimensions is as follows:

Algorithm Test-Independence-2D

Input: Sample access to a distribution p on $[n] \times [m]$ with $n \geq m$ and $\epsilon > 0$.

Output:"YES" with probability at least $2/3$ if the coordinates of p are independent, "NO" with probability at least $2/3$ if p is ϵ-far from any product distribution on $[n] \times [m]$.

1) Let $k = \min(n, n^{2/3}m^{1/3}\epsilon^{-4/3})$.
2) Let S_1 be a multiset in $[n]$ obtained by taking $\mathrm{Poi}(k)$ samples from $p_1 = \pi_1(p)$. Let S_2 be a multiset in $[m]$ obtained by taking $\mathrm{Poi}(m)$ samples from $p_2 = \pi_2(p)$. Let S be the multiset of elements of $[n]\times[m]$ so that
$$1 + \{\text{Number of copies of } (a,b) \text{ in } S\} = (1 + \{\text{Number of copies of } a \text{ in } S_1\})(1+ \{\text{Number of copies of } b \text{ in } S_2\}).$$
3) Let q be the distribution on $[n] \times [m]$ obtained by taking $(x_1, y_1), (x_2, y_2)$ independent samples from p and returning (x_1, y_2). Run the tester from Lemma II.3 to distinguish between the cases $p_S = q_S$ and $\|p_S - q_S\|_1 \geq \epsilon$.

For correctness, we note that by Lemma II.6, with probability at least $9/10$ over our samples from S_1 and S_2, all of the above hold: (i) $|S_1| = O(n)$ and $|S_2| = O(m)$, and (ii) $\|(p_1)_{S_1}\|_2^2 = O(1/k)$, $\|(p_2)_{S_2}\|_2^2 = O(1/m)$. We henceforth condition on this event. We note that the distribution q is exactly $p_1 \times p_2$. Therefore, if the coordinates of p are independent, then $p = q$. On the other hand, since q has independent coordinates, if p is ϵ-far from any product distribution, $\|p - q\|_1 \geq \epsilon$. Therefore, it suffices to distinguish between $p = q$ and $\|p - q\|_1 \geq \epsilon$. By Fact II.5, this is equivalent to distinguishing between $p_S = q_S$ and $\|p_S - q_S\|_1 \geq \epsilon$. This completes correctness.

We now analyze the sample complexity. We first draw samples when picking S_1 and S_2. With high probability, the corresponding number of samples is $O(m + k) = O(\max(n^{2/3}m^{1/3}\epsilon^{-4/3}, \sqrt{nm}/\epsilon^2))$. Next, we note that $q_S = (p_1)_{S_1} \times (p_2)_{S_2}$. Therefore, by Lemma II.3, the number of samples drawn in the last step of the algorithm is at most

$$\begin{aligned} O(nm\|q_S\|_2/\epsilon^2) &= O(nm\|(p_1)_{S_1} \times (p_2)_{S_2}\|_2/\epsilon^2) \\ &= O(nm\|(p_1)_{S_1}\|_2\|(p_2)_{S_2}\|_2/\epsilon^2) \\ &= O(nmk^{-1/2}m^{-1/2}/\epsilon^2) \\ &= O(\max(n^{2/3}m^{1/3}\epsilon^{-4/3}, \sqrt{nm}/\epsilon^2)). \end{aligned}$$

Drawing a sample from q requires taking only two samples from p, which completes the analysis.

In the following proposition, we generalize the two-dimensional algorithm to optimally test independence in any number of dimensions.

Proposition II.15. *Let p be a distribution on $\times_{i=1}^d [n_i]$.*

There is an algorithm that draws

$$O(\max_j((\prod_{i=1}^{d} n_i)^{1/2}/\epsilon^2, n_j^{1/3}(\prod_{i=1}^{d} n_i)^{1/3}/\epsilon^{4/3}))$$

samples form p and with probability at least $2/3$ distinguishes between the coordinates of p being independent and p being ϵ-far from any such distribution.

Roughly speaking, our independence tester in general dimension uses recursion to reduce to the 2-dimensional case, in which case we may apply Test-Independence-2D. The details are given in the full version.

5) Testing Properties of Collections of Distributions: In this subsection, we consider the model of testing properties of collections of distributions [16] in both the sampling and query models.

We begin by considering the sampling model, as this is closely related to independence testing. In fact, in the unknown-weights case, the problem is identical. In the known-weights case, the problem is equivalent to independence testing, where the algorithm is given explicit access to one of the marginals (say, the distribution on $[m]$). For this setting, we give a tester with sample complexity $O(\max(\sqrt{nm}/\epsilon^2, n^{2/3}m^{1/3}/\epsilon^{4/3}))$. We also note that this bound can be shown the be optimal. Formally, we prove the following:

Proposition II.16. *There is an algorithm that given sample access to a distribution p on $[n] \times [m]$ and an explicit description of the marginal of p on $[m]$ distinguishes between the cases that the coordinates of p are independent and the case where p is ϵ-far from any product distribution on $[n] \times [m]$ with probability at least $2/3$ using $O(\max(\sqrt{nm}/\epsilon^2, n^{2/3}m^{1/3}/\epsilon^{4/3}))$ samples.*

Next, we consider the query model. In this model, we are essentially guaranteed that the distribution on $[m]$ is uniform, but are allowed to extract samples conditioned on a particular value of the second coordinate. Equivalently, there are m distributions $q_1, \ldots, q_m$ on $[n]$. We wish to distinguish between the cases that the q_i's are identical and the case where there is no distribution q so that $\frac{1}{m}\sum_{i=1}^{m} \|q - q_i\|_1 \leq \epsilon$. We show that we can solve this problem with $O(\max(\sqrt{n}/\epsilon^2, n^{2/3}/\epsilon^{4/3}))$ samples for any m. This is optimal for all $m \geq 2$, even if we are guaranteed that $q_1 = q_2 = \ldots = q_{\lfloor m/2 \rfloor}$ and $q_{\lfloor m/2+1 \rfloor} = \ldots = q_m$.

Proposition II.17. *There is an algorithm that given sample access to distributions $q_1, \ldots, q_m$ on $[n]$ distinguishes between the cases that the q_i's are identical and the case where there is no distribution q so that $\frac{1}{m}\sum_{i=1}^{m} \|q - q_i\|_1 \leq \epsilon$ with probability at least $2/3$ using $O(\max(\sqrt{n}/\epsilon^2, n^{2/3}/\epsilon^{4/3}))$ samples.*

The basic idea of the algorithm is as follows. Firstly, we let q_* denote the average of the distributions q_i. We note that it suffices to distinguish between the case where $q_i = q_*$ for all i and the case where $\sum_{i=1}^{m} \|q_i - q_*\|_1 \gg m\epsilon$. There are many ways this could happen. For example, if the *average* size of $\|q_i - q_*\|_1$ is on the order of ϵ, we could test for this by testing a few q_i against q. Alternatively, it could instead be the case that most q_i are close, but a small number (say m/a of them) have distance on the order of $a\epsilon$. However, this is even easier to test for. We would merely need to check a random sample of $O(a)$ i's and test for $a\epsilon$-closeness. Fortunately, the decrease in sample complexity from having a larger ϵ will more than compensate for the increase in the number of times we must run the test. In order to get the algorithm to work, we merely need to carefully balance this sort of test for different values of a and deal appropriately with the error probabilities. See the full version for the proof.

6) Testing k-Histograms: Finally, in this subsection we use our framework to design a sample-optimal algorithm for the property of being a k-histogram with known intervals.

Let $\mathcal{I}$ be a partition of $[n]$ into k intervals. We wish to be able to distinguish between the cases where a distribution p has constant density on each interval versus the case where it is ϵ-far from any such distribution. We show the following:

Proposition II.18. *Let $\mathcal{I}$ be a partition of $[n]$ into k intervals. Let p be a distribution on $[n]$. There exists an algorithm which draws $O(\max(\sqrt{n}/\epsilon^2, n^{1/3}k^{1/3}/\epsilon^{4/3}))$ independent samples from p and distinguishes between the cases where p is uniform on each of the intervals in $\mathcal{I}$ from the case where p is ϵ-far from any such distribution with probability at least $2/3$.*

We provide a sketch of the algorithm deferring the details to the full version. First, we wish to guarantee that each of the intervals has reasonably large support. We can achieve this as follows: For each interval $I \in \mathcal{I}$ we divide each bin within I into $\lceil n/(k|I|) \rceil$ bins. Next, in order to use an ℓ_2-closeness tester, we want to further subdivide bins using our randomized transformation. To this end, we let $m = \min(k, n^{1/3}k^{1/3}/\epsilon^{4/3})$ and take $\mathrm{Poi}(m)$ samples from p. Then, for each interval $I_i \in \mathcal{I}$, we divide each bin in I_i into $\lfloor na_i/(k|I_i|) \rfloor + 1$ new bins, where a_i is the number of samples that were drawn from I_i. Let I_i' denote the new interval obtained from I_i. Let q' be the distribution obtained by sampling from p' and then returning a uniform random bin from the same interval I_i' as the sample. We claim that the ℓ_2-norm of q' is small. We can now apply the tester from Lemma II.3 to distinguish between the cases where $p' = q'$ and $\|p' - q'\|_1 > \epsilon$ with $O(n^{1/2}k^{1/2}m^{-1/2}/\epsilon^2) = O(\max(\sqrt{n}/\epsilon^2, n^{1/3}k^{1/3}/\epsilon^{4/3}))$ samples.

III. Sample Complexity Lower Bounds

We illustrate our lower bound technique by proving tight information-theoretic lower bounds for testing independence (in any dimension), testing closeness in Hellinger distance, and testing histograms. Due to space limitations, we present

our lower bound for 2-dimensional independence testing and defer the rest to the full version.

A. Lower Bound for Two-Dimensional Independence Testing

Theorem III.1. *Let $n \geq m \geq 2$ be integers and $\epsilon > 0$ a sufficiently small universal constant. Then, any algorithm that draws samples from a distribution p on $[n] \times [m]$ and, with probability at least $2/3$, distinguishes between the case that the coordinates of p are independent and the case where p is ϵ-far from any product distribution must use $\Omega(\max(\sqrt{nm}/\epsilon^2, n^{2/3}m^{1/3}/\epsilon^{4/3}))$ samples.*

We prove the easier lower bound of $\Omega(\sqrt{nm}\epsilon^{-2})$ and defer the $\Omega(n^{2/3}m^{1/3}/\epsilon^{4/3})$ lower bound to the full version. First, we note that it suffices to consider the case where n and m are each sufficiently large, since $\Omega(\epsilon^{-2})$ samples are required to distinguish the uniform distribution on $[2] \times [2]$ from the distribution which takes value (i, j) with probability $(1 + (2\delta_{i,j} - 1)\epsilon)/2$.

Our goal is to exhibit distributions $\mathcal{D}$ and $\mathcal{D}'$ over distributions on $[n] \times [m]$ so that all distributions in $\mathcal{D}$ have independent coordinates, and all distributions in $\mathcal{D}'$ are ϵ-far from product distributions, so that for any $k = o(\sqrt{nm}/\epsilon^2)$, no algorithm given k independent samples from a random element of either $\mathcal{D}$ or $\mathcal{D}'$ can determine which family the distribution came from with greater than 90% probability.

We will analyze the following generalization in order to simplify the argument. First, we use the standard Poissonization trick: instead of drawing k samples from the appropriate distribution, we will draw $\mathrm{Poi}(k)$ samples. This is acceptable because with 99% probability, this is at least $\Omega(k)$ samples. Next, we relax the condition that elements of $\mathcal{D}'$ be ϵ-far from product distributions, and simply require that they are $\Omega(\epsilon)$-far from product distributions with 99% probability. This is clearly equivalent upon accepting an additional 1% probability of failure, and altering ϵ by a constant factor.

Finally, we will relax the constraint that elements of $\mathcal{D}$ and $\mathcal{D}'$ are probability distributions. Instead, we will merely require that they are positive measures on $[n] \times [m]$, so that elements of $\mathcal{D}$ are product measures and elements of $\mathcal{D}'$ are $\Omega(\epsilon)$-far from being product measures with probability at least 99%. We will require that the selected measures have total mass $\Theta(1)$ with probability at least 99%, and instead of taking samples from these measures (as this is no longer as sensible concept), we will use the points obtained from a Poisson process of parameter k (so the number of samples in a given bin is a Poisson random variable with parameter k times the mass of the bin). This is sufficient, because the output of such a Poisson process for a measure μ is identical to the outcome of drawing $\mathrm{Poi}(\|\mu\|_1 k)$ samples from the distribution $\mu/\|\mu\|_1$. Moreover, the distance from μ to the nearest product distribution is $\|\mu\|_1$ times the distance from $\mu/\|\mu\|_1$ to the nearest product distribution.

We are now prepared to describe $\mathcal{D}$ and $\mathcal{D}'$ explicitly:

- We define $\mathcal{D}$ to deterministically return the uniform distribution μ with $\mu(i, j) = \frac{1}{nm}$ for all $(i, j) \in [n] \times [m]$.
- We define $\mathcal{D}'$ to return the positive measure ν so that for each $(i, j) \in [n] \times [m]$ the value $\nu(i, j)$ is either $\frac{1+\epsilon}{nm}$ or $\frac{1-\epsilon}{nm}$ each with probability $1/2$ and independently over different pairs (i, j).

It is clear that $\|\mu\|_1, \|\nu\|_1 = \Theta(1)$ deterministically. We need to show that the relevant Poisson processes return similar distributions. To do this, we consider the following procedure: Let X be a uniformly random bit. Let p be a measure on $[n] \times [m]$ drawn from either $\mathcal{D}$ if $X = 0$ or from $\mathcal{D}'$ if $X = 1$. We run a Poisson process with parameter k on p, and let $a_{i,j}$ be the number of samples drawn from bin (i, j). We wish to show that, given access to all $a_{i,j}$'s, one is not able to determine the value of X with probability more than 51%. To prove this, it suffices to bound from above the mutual information between X and the set of samples $(a_{i,j})_{(i,j)\in[n]\times[m]}$. In particular, this holds true because of the following simple fact:

Lemma III.2. *If X is a uniform random bit and A is a correlated random variable, then if f is any function so that $f(A) = X$ with at least 51% probability, then $I(X : A) \geq 2 \cdot 10^{-4}$.*

In order to bound $I(X : \{a_{i,j}\})$ from above, we note that the $a_{i,j}$'s are independent conditional on X, and therefore

$$I(X : (a_{i,j})_{(i,j)\in[n]\times[m]}) \leq \sum_{(i,j)\in[n]\times[m]} I(X : a_{i,j}). \quad (1)$$

By symmetry, it is clear that all of the $a_{i,j}$'s are the same, so it suffices to consider $I(X : a)$ for a being one of the $a_{i,j}$. We prove the following technical lemma:

Lemma III.3. *For all $(i, j) \in [n] \times [m]$, it holds $I(X : a_{i,j}) = O(k^2\epsilon^4/(m^2n^2))$.*

The proof of this lemma is technical and is deferred to the full version. The essential idea is that we condition on whether or not $\lambda := k/(nm) \geq 1$. If $\lambda < 1$, then the probabilities of seeing 0 or 1 samples are approximately the same, and most of the information comes from how often one sees exactly 2 samples. For $\lambda \geq 1$, we are comparing a Poisson distribution to a mixture of Poisson distributions with the same average mean, and we can deal with the information theory by making a Gaussian approximation.

By Lemma III.3, (1) yields that $I(X : (a_{i,j})_{(i,j)\in[n]\times[m]}) = O(k^2\epsilon^4/mn) = o(1)$. In conjunction with Lemma III.2, this implies that $o(\sqrt{mn}/\epsilon^2)$ samples are insufficient to reliably distinguish an element of $\mathcal{D}$ from an element of $\mathcal{D}'$. To complete the proof, it remains to show that elements of $\mathcal{D}$ are all product distributions, and that most elements of $\mathcal{D}'$ are far from product distributions. The former follows trivially, and the latter is not difficult.

ACKNOWLEDGMENT

We would like to thank Oded Goldreich for numerous useful comments and insightful conversations that helped us improve the presentation of this work. We are grateful to Oded for his excellent exposition of our technique in his lecture notes [32]. Part of this work was performed while ID was at the University of Edinburgh. ID supported by a Marie Curie Career Integration Grant, a SICSA grant, and a USC startup. DK supported in part by NSF Award CCF-1553288 (CAREER).

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London.*, vol. 231, no. 694-706, pp. 289–337, 1933.

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

[3] R. Rubinfeld and M. Sudan, "Robust characterizations of polynomials with applications to program testing," *SIAM J. on Comput.*, vol. 25, pp. 252–271, 1996.

[4] O. Goldreich, S. Goldwasser, and D. Ron, "Property testing and its connection to learning and approximation," *Journal of the ACM*, vol. 45, pp. 653–750, 1998.

[5] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *FOCS*, 2000, pp. 259–269.

[6] ——, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, p. 4, 2013.

[7] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," ECCC, Tech. Rep. TR00-020, 2000.

[8] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *FOCS*, 2001, pp. 442–451.

[9] T. Batu, "Testing properties of distributions," Ph.D. dissertation, Cornell University, 2001.

[10] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, "The complexity of approximating entropy," in *STOC*, 2002.

[11] T. Batu, R. Kumar, and R. Rubinfeld, "Sublinear algorithms for testing monotone and unimodal distributions," in *ACM Symposium on Theory of Computing*, 2004, pp. 381–390.

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

[13] P. Valiant, "Testing symmetric properties of distributions," *SIAM J. Comput.*, vol. 40, no. 6, pp. 1927–1968, 2011.

[14] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant, "Testing k-modal distributions: Optimal algorithms via reductions," in *SODA*, 2013, pp. 1833–1852.

[15] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," *Journal of Machine Learning Research - Proceedings Track*, vol. 19, pp. 47–68, 2011.

[16] R. Levi, D. Ron, and R. Rubinfeld, "Testing properties of collections of distributions," in *ICS*, 2011, pp. 179–194.

[17] P. Indyk, R. Levi, and R. Rubinfeld, "Approximating and Testing k-Histogram Distributions in Sub-linear Time," in *PODS*, 2012, pp. 15–22.

[18] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *SODA*, 2014, pp. 1193–1203.

[19] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," in *FOCS*, 2014.

[20] I. Diakonikolas, D. M. Kane, and V. Nikishkin, "Testing Identity of Structured Distributions," in *SODA*, 2015.

[21] ——, "Optimal algorithms and lower bounds for testing closeness of structured distributions," in *FOCS*, 2015.

[22] J. Acharya, C. Daskalakis, and G. Kamath, "Optimal testing for properties of distributions," in *NIPS*, 2015.

[23] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, "Testing shape restrictions of discrete distributions," in *STACS*, 2016, pp. 25:1–25:14.

[24] R. Rubinfeld, "Taming big probability distributions," *XRDS*, vol. 19, no. 1, pp. 24–28, 2012.

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

[26] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, "Strong lower bounds for approximating distribution support size and the distinct elements problem," *SIAM J. Comput.*, vol. 39, no. 3, pp. 813–842, 2009.

[27] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Sublinear algorithms for outlier detection and generalized closeness testing," in *2014 IEEE ISIT*, 2014, pp. 3200–3204.

[28] B. B. Bhattacharya and G. Valiant, "Testing closeness with unequal sized samples," *CoRR*, vol. abs/1504.04599, 2015.

[29] S. Guha, A. McGregor, and S. Venkatasubramanian, "Sublinear estimation of entropy and information distances," *ACM Trans. Algorithms*, vol. 5, no. 4, pp. 35:1–35:16, Nov. 2009.

[30] C. L. Canonne, "Are few bins enough: Testing histogram distributions," in *PODS*, 2016, pp. 455–463.

[31] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh, "Competitive classification and closeness testing," in *COLT*, 2012.

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

[33] ——, "The uniform distribution is complete with respect to testing identity to a fixed distribution," *ECCC*, vol. 23, 2016.