2014 IEEE Annual Symposium on Foundations of Computer Science

An Automatic Inequality Prover and Instance Optimal Identity Testing

Gregory Valiant*
Stanford University
valiant@stanford.edu

Paul Valiant†
Brown University
pvaliant@gmail.com

***Abstract*—We consider the problem of verifying the identity of a distribution: Given the description of a distribution over a discrete support $p = (p_1, p_2, \ldots, p_n)$, how many samples (independent draws) must one obtain from an unknown distribution, q, to distinguish, with high probability, the case that $p = q$ from the case that the total variation distance (L_1 distance) $||p - q||_1 \geq \epsilon$? We resolve this question, up to constant factors, on an *instance by instance* basis: there exist universal constants c, c' and a function $f(p, \epsilon)$ on distributions and error parameters, such that our tester distinguishes $p = q$ from $||p-q||_1 \geq \epsilon$ using $f(p, \epsilon)$ samples with success probability $> 2/3$, but no tester can distinguish $p = q$ from $||p-q||_1 \geq c \cdot \epsilon$ when given $c' \cdot f(p, \epsilon)$ samples. The function $f(p, \epsilon)$ is upper-bounded by a multiple of $\frac{||p||_{2/3}}{\epsilon^2}$, but is more complicated, and is significantly smaller in some cases when p has many small domain elements, or a single large one. This result significantly generalizes and tightens previous results: since distributions of support at most n have $L_{2/3}$ norm bounded by $\sqrt{n}$, this result immediately shows that for such distributions, $O(\sqrt{n}/\epsilon^2)$ samples suffice, tightening the previous bound of $O(\frac{\sqrt{n}\,\mathrm{polylog}\, n}{\epsilon^4})$ for this class of distributions, and matching the (tight) known results for the case that p is the uniform distribution over support n.**

The analysis of our very simple testing algorithm involves several hairy inequalities. To facilitate this analysis, we give a complete characterization of a general class of inequalities—generalizing Cauchy-Schwarz, Hölder's inequality, and the monotonicity of L_p norms. Specifically, we characterize the set of sequences $(a)_i = a_1, \ldots, a_r$, $(b)_i = b_1, \ldots, b_r$, $(c)_i = c_1, \ldots, c_r$, for which it holds that for all finite sequences of positive numbers $(x)_j = x_1, \ldots$ and $(y)_j = y_1, \ldots$,

$$\prod_{i=1}^{r} \left(\sum_j x_j^{a_i} y_j^{b_i} \right)^{c_i} \geq 1.$$

For example, the standard Cauchy-Schwarz inequality corresponds to the sequences $a = (1, 0, \frac{1}{2})$, $b = (0, 1, \frac{1}{2})$, $c = (\frac{1}{2}, \frac{1}{2}, -1)$. Our characterization is of a non-traditional nature in that it uses linear programming to compute a derivation that may otherwise have to be sought through trial and error, by hand. We do not believe such a characterization has appeared in the literature, and hope its computational nature will be useful to others, and facilitate analyses like the one here.

* Supported in part by NSF CAREER Award CCF-1351108. Portions of this work were done while the author was a postdoc at Microsoft Research, New England, and while the author was a visitor at the Simons Institute for the Theory of Computing, UC Berkeley.

†Part of this work was done while the author was a visitor at the Simons Institute for the Theory of Computing, UC Berkeley.

I. Introduction

Suppose you have a detailed record of the distribution of IP addresses that visit your website. You recently proved an amazing theorem, and are keen to determine whether this result has changed the distribution of visitors to your website (or is it simply that the usual crowd is visiting your website more often?). How many visitors must you observe to decide this, and, algorithmically, how do you decide this? Formally, given some known distribution p over a discrete (though possibly infinite) domain, a parameter $\epsilon > 0$, and an unknown distribution q from which we may draw independent samples, we would like an algorithm that will distinguish the case that $q = p$ from the case that the total variation distance, $d_{tv}(p, q) > \epsilon$. We consider this basic question of verifying the identity of a distribution, also known as the problem of "identity testing against a known distribution". This problem has been well studied, and yielded the punchline that it is sometimes possible to perform this task using far fewer samples than would be necessary to accurately learn the distribution from which the samples were drawn. Nevertheless, previous work on this problem either considered only the problem of verifying a uniform distribution (the case that $p = Unif[n]$), or was from the perspective of worst-case analysis—aiming to bound the number of samples required to verify a worst-case distribution of a given support size.

Here, we seek a deeper understanding of this problem. We resolve, up to constant factors, the sample complexity of this task on an *instance-by-instance* basis—determining the optimal number of samples required to verify the identity of a distribution, *as a function of the distribution in question*.

Throughout much of TCS, the main challenge and goal is to characterize problems from a worst-case standpoint, and the efforts to describe algorithms that perform well "in practice" is often relegated to the sphere of heuristics. Still, there is a developing understanding of domains and approaches for which one may provide analysis beyond the worst-case (e.g. random instances, smoothed analysis, competitive analysis, analysis with respect to various parameterizations of the problems, etc.). Against this backdrop, it seems especially exciting when a rich setting seems amenable to the development and analysis of *instance optimal* algorithms, not to mention that instance optimality gives a strong

0272-5428/14 $31.00 © 2014 IEEE
DOI 10.1109/FOCS.2014.14

recommendation for the practical viability of the proposed algorithms.

In the setting of this paper, having the distribution p explicitly provided to the tester enables our approach; nevertheless, it is tantalizing to ask whether this style of "instance-by-instance optimal" property testing/estimation or learning is possible in more general distributional settings. The authors are optimistic that such strong theoretical results are both within our reach, and that pursuing this line may yield practical algorithms suited to making the best use of available data.

To more cleanly present our results, we introduce the following notation.

Definition 1. *For a probability distribution p, let $p^{-\max}$ denote the vector of probabilities obtained by removing the entry corresponding to the element of largest probability. For $\epsilon > 0$, define $p_{-\epsilon}$ to be the vector obtained from p by iteratively removing the smallest domain elements and stopping before more than ϵ probability mass is removed.*

Hence $p_{-\epsilon}^{-\max}$ is the vector of probabilities corresponding to distribution p, after the largest domain element and the smallest domain elements have been removed. Our main result is the following:

Theorem 1. *There exist constants c_1, c_2 such that for any $\epsilon > 0$ and any known distribution p, for any unknown distribution q, our tester will distinguish $q = p$ from $||p - q||_1 \geq \epsilon$ with probability $2/3$ when run on a set of at least $c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon/16}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ samples drawn from q, and no tester can do this task with probability at least $2/3$ with a set of fewer than $c_2 \cdot \max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ samples.*

In short, over the entire range of potential distributions p, our tester is optimal, up to constant factors in ϵ and the number of samples. The distinction of "constant factors in ϵ" is needed, as $||p_{-\epsilon/16}||_{2/3}$ might *not* be within a constant factor of $||p_{-\epsilon}||_{2/3}$ if, for example, the vast majority of the $2/3$-norm of p comes from tiny domain elements that only comprise an ϵ fraction of the 1-norm (and hence would be absent from $p_{-\epsilon}$, though not from $p_{-\epsilon/16}$).

Because our tester is constant-factor tight, the subscript and superscript and the max in the sample complexity $\max\left\{\frac{1}{\epsilon}, \frac{||p_{-O(\epsilon)}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ all mark real phenomena, and are not just artifacts of the analysis. However, except for rather pathological distributions, the theorem says that $\Theta(\frac{||p||_{2/3}}{\epsilon^2})$ is the optimal number of samples. Additionally, note that the subscript and superscripts only reduce the value of the norm: $||p_{-\epsilon}^{-\max}||_{2/3} < ||p_{-\epsilon}||_{2/3} \leq ||p_{-\epsilon/16}||_{2/3} \leq ||p||_{2/3}$, and hence $O(||p||_{2/3}/\epsilon^2)$ is always an upper bound on the number of samples required. Since $x^{2/3}$ is concave, for distributions p of support size at most n the $L_{2/3}$ norm is maximized on the uniform distribution, yielding that $||p||_{2/3} \leq \sqrt{n}$, with equality if and only if p is the uniform distribution. This immediately yields a worst-case bound of $O(\sqrt{n}/\epsilon^2)$ on the number of samples required to test distributions supported on at most n elements, tightening the previous bound of $O(\frac{\sqrt{n}\,\mathrm{polylog}\, n}{\epsilon^4})$ from [5], and matching the tight bound on the number of samples required for testing the uniform distribution given in [13].

While the algorithm we propose is extremely simple, the analysis involves sorting through several messy inequalities. To facilitate this analysis, we give a complete characterization of a general class of inequalities. We characterize the set of sequences $a = a_1, \ldots, a_r$, $b = b_1, \ldots, b_r$, $c = c_1 \ldots, c_r$, for which it holds that for all finite sequences of positive numbers $(x)_j = x_1, \ldots$ and $(y)_j = y_1, \ldots,$

$$\prod_{i=1}^{r} \left(\sum_j x_j^{a_i} y_j^{b_i} \right)^{c_i} \geq 1. \tag{1}$$

This is an extremely frequently encountered class of inequalities, and contains the Cauchy-Schwarz inequality and its generalization, the Hölder inequality, in addition to inequalities representing the monotonicity of the L_p norm, and also clearly contains any finite product of such inequalities. Additionally, we note that the constant 1 on the right hand side cannot be made larger, for all such inequalities are false when the sequences x and y consist of a single 1; also, as we show, this class of inequality is unchanged if 1 is replaced by any other constant in the interval $(0, 1]$.

Example 1. *The classic Cauchy-Schwarz inequality can be expressed in the form of Equation 1 as $\left(\sum_j X_j\right)^{1/2} \left(\sum_j Y_j\right)^{1/2} \left(\sum_j \sqrt{X_j Y_j}\right)^{-1} \geq 1$, corresponding to the 3-term sequences $a = (1, 0, \frac{1}{2})$, $b = (0, 1, \frac{1}{2})$, and $c = (\frac{1}{2}, \frac{1}{2}, -1)$. This inequality is tight when the sequences X and Y are proportional to each other. The Hölder inequality generalizes Cauchy-Schwarz by replacing $\frac{1}{2}$ by $\lambda \in [0, 1]$, yielding the inequality defined by the triples $a = (1, 0, \lambda)$, $b = (0, 1, 1 - \lambda)$, and $c = (\lambda, 1 - \lambda, -1)$.*

Example 2. *A fundamentally different inequality that can also be expressed in the form of Equation 1 is the fact that the L_p norm is a non-increasing function of p. For $p \in [0, 1]$ we have the inequality $\left(\sum_j X_j^p\right) \left(\sum_j X_j\right)^{-p} \geq 1$, corresponding to the 2-term sequences $a = (p, 1)$, $b = (0, 0)$, and $c = (1, -p)$. This inequality is tight only when the sequence $(X)_j$ consists of a single nonzero term.*

We show that the cases where Equation 1 holds are exactly those cases expressible as a product of inequalities of the above two forms, where two arbitrary combinations of x and y are substituted for the sequence X and the sequence Y in the above examples:

Theorem 2. *For fixed sequences* $(a)_i = a_1, \ldots, a_r, (b)_i = b_1, \ldots, b_r$, *and* $\underset{c_i}{(c)_i} = c_1, \ldots, c_r$, *the inequality* $\prod_{i=1}^{r} \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i} \geq 1$ *holds for all finite sequences of positive numbers* $(x)_j, (y)_j$ *if and only if it can be expressed as a finite product of positive powers of the Hölder inequalities* $\left(\sum_j x_j^{a'} y_j^{b'}\right)^{\lambda} \left(\sum_j x_j^{a''} y_j^{b''}\right)^{1-\lambda} \geq \sum_j x_j^{\lambda a' + (1-\lambda)a''} y_j^{\lambda b' + (1-\lambda)b''}$, *and the* L_p *monotonicity inequalities* $\left(\sum_j x_j^a y_j^b\right)^{\lambda} \leq \sum_j x_j^{\lambda a} y_j^{\lambda b}$, *for* $\lambda \in [0, 1]$.

We state this theorem for pairs of sequences $(x)_j, (y)_j$, although an analogous statement (Theorem 3 stated in Section II) holds for any number of sequences and is yielded by a trivial extension of the proof of the above theorem. Most commonly encountered instances of inequalities of the above form, including those involved in our identity testing result, involve only pairs of sequences. Further, the result is nontrivial even for inequalities of the above form that only involve a single sequence—see Example 3 for a discussion of a single sequence inequality with surprising properties.

Our proof of Theorem 2 is algorithmic in nature; in fact, we describe an algorithm which, when given the sequences a, b and c, as input, will run in polynomial time, and either output a derivation of the desired inequality as a product of a polynomial number of Hölder and L_p monotonicity inequalities, or the algorithm will output a witness from which a pair of sequences $(x)_j, (y)_j$ that violate the inequality can be constructed. It is worth stressing that the algorithm is efficient despite the fact that the shortest counter-example sequences $(x)_j, (y)_j$ might require a doubly-exponential number of terms (doubly-exponential in the number of bits required to represent the sequences a, b, c—see Example 3).

The characterization of Theorem 2 seems to be a useful and general tool, and seems absent from the literature, perhaps because linear programming duality is an unexpected tool with which to analyze such inequalities. The ability to efficiently verify inequalities of the above form greatly simplified the tasks of proving our instance optimality results; we believe this tool will prove useful to others and have made a Matlab implementation of our inequality prover/refuter publicly available at http://theory.stanford.edu/~valiant/code.

A. Related Work

The general area of hypothesis testing was launched by Pearson in 1900, with the description of Pearson's chi-squared test. In this current setting of determining whether a set of k samples was drawn from distribution $p = p_1, p_2, \ldots$, that test would correspond to evaluating $\sum_i \frac{1}{p_i}(X_i - kp_i)^2$, where X_i denotes the number of occurrences of the ith domain element in the samples, and then outputting "yes" if the value of this statistic is sufficiently small. Traditionally, such tests are evaluated in the asymptotic regime, for a fixed distribution p as the number of samples tends to infinity. In the current setting of trying to verify the identity of a distribution, using this chi-squared statistic might require using many more samples than would be necessary even to accurately *learn* the distribution from which the samples were drawn (see, e.g. Example 6).

Over the past fifteen years, there has been a body of work exploring the general question of how to estimate or test properties of distributions *using fewer samples than would be necessary to actually learn the distribution in question*. Such properties include "symmetric" properties (properties whose value is invariant to relabeling domain elements) such as entropy, support size, and distance metrics between distributions (such as L_1 distance), with work on both the algorithmic side (e.g. [6], [4], [10], [11], [12], [3], [8]), and on establishing lower bounds [14], [18]. Such problems have been almost exclusively considered from a worst-case standpoint, with bounds on the sample complexity parameterized by an upper bound on the support size of the distribution. The recent work [16], [17] resolved the worst-case sample complexities of estimating many of these symmetric properties. Also see [15] for a recent survey.

The specific question of verifying the identity of a distribution was one of the first questions considered in this line of work. Motived by a connection to testing the expansion of graphs, Goldreich and Ron [9] first considered the problem of distinguishing whether a set of samples was drawn from the uniform distribution of support n versus from a distribution that is least ϵ far from the uniform distribution, with the tight bound of $\Theta(\frac{\sqrt{n}}{\epsilon^2})$ subsequently given by Paninski [13]. For the more general problem of verifying an arbitrary distribution, Batu et al. [5], showed that for worst-case distributions of support size n, $O(\frac{\sqrt{n}\,\mathrm{polylog}\, n}{\epsilon^4})$ samples are sufficient.

In a similar spirit to this current paper, motivated by a desire to go beyond worst-case analysis, Acharya et al. [1], [2] recently considered the question of identity testing with two unknown distributions (i.e. both distributions p and q are unknown, and one wishes to deduce if $p = q$ from samples) from the standpoint of *competitive analysis*. They asked how many samples are required as a function of the number of samples that would be required for the task of distinguishing whether samples were drawn from p versus q in the case where p and q were known to the algorithm. Their main results are an algorithm that performs the desired task using $m^{3/2}\,\mathrm{polylog}\, m$ samples, and a lower bound of $\Omega(m^{7/6})$, where m represents the number of samples required to determine whether a set of samples were drawn from p versus q in the setting where p and q are explicitly known. One of the main conceptual messages from Acharya et al.'s results is that knowledge of the underlying distributions is extremely helpful—without such knowledge one loses a polynomial factor in sample complexity. Our results build on this moral, in some sense describing the "right" way that knowledge of a distribution could be used to test identity.

The form of our tester may be seen as rather similar to those in [7], which considered testing whether two distributions were close or not when *both* distributions are unknown. The testers in that paper and the tester proposed here consist essentially of summing up carefully chosen expressions independently evaluated at the different domain elements and comparing this sum to a threshold, which is somewhat simpler what has been pursued by many other lines of research. The moral here is perhaps that combining all your data in "clean" ways can lead to "clean" and powerful results, at least provided one has the right analysis tools. The lower bounds from [7] rely on Theorem 4 from the present paper, which uses Hellinger distance to introduce a flexible class of lower bounds, needed for the tight results there and here.

B. Organization

We begin with our characterization of the class of inequalities, as we feel that this tool may be useful to the broad TCS community; this first section is entirely self-contained. Section III-A contains the definitions and terminology relevant to the distribution testing portion of the paper, and Section III-B describes our very simple instance-optimal distribution identity testing algorithm, and provides some context and motivation for the algorithm. Section IV discusses the lower-bounds, establishing the optimality of our tester.

II. A class of inequalities generalizing Cauchy-Schwarz and the monotonicity of L_p norms

In this section we characterize under what conditions a large class of inequalities holds, showing both how to derive these inequalities when they are true and how to refute them when they are false. We encounter such inequalities repeatedly in the analysis of our tester in Section III.

The basic question we resolve is: for what sequences $(a)_i, (b)_i, (c)_i$ is it true that for all sequences of positive numbers $(x)_j, (y)_j$ we have

$$\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i} \right)^{c_i} \geq 1 \tag{2}$$

We note that the constant 1 on the right hand side cannot be made larger, for all such inequalities are false when the sequences x and y consist of a single 1; also, as we will show later, if this inequality can be violated, it can be violated by an arbitrary amount, so if any right hand side constant works, for a given $(a)_i, (b)_i, (c)_i$, then 1 works, as stated above.

Such inequalities are typically proven by hand, via trial and error. One basic tool for this is the Cauchy-Schwarz inequality, $\left(\sum_j X_j\right)^{1/2} \left(\sum_j Y_j\right)^{1/2} \geq \sum_j \sqrt{X_j Y_j}$, or the slightly more general Hölder inequality, a weighted version of Cauchy-Schwarz, where for $\lambda \in (0,1)$ we have $\left(\sum_j X_j\right)^{\lambda} \left(\sum_j Y_j\right)^{1-\lambda} \geq \sum_j X_j^{\lambda} Y_j^{1-\lambda}$. Writing this in the form of Equation 2, and substituting arbitrary combinations of x and y for X and Y yields families of inequalities of the form:

$$\left(\sum_j x_j^{a_1} y_j^{b_1}\right)^{\lambda} \left(\sum_j x_j^{a_2} y_j^{b_2}\right)^{1-\lambda} \left(\sum_j x_j^{\lambda a_1 + (1-\lambda) a_2} y_j^{\lambda b_1 + (1-\lambda) b_2}\right)^{-1} \geq 1,$$

and we can multiply inequalities of this form together to get further cases of the inequality in Equation 2. This inequality is tight when the two sequences X and Y are proportional to each other.

A second and different basic inequality of our general form, for $\lambda \in [0,1)$, is: $\left(\sum_j X_j\right)^{\lambda} \leq \sum_j X_j^{\lambda}$, which is the fact that the L_p norm is a decreasing function of p. (Intuitively, this is a slight generalization of the trivial fact that $x^2 + y^2 \leq (x+y)^2$, and follows from the fact that the derivative of x^{λ} is a decreasing function of x, for positive x). As above, products of powers of x and y may be substituted for X to yield a more general class of inequalities: $\sum_j x_j^{\lambda a} y_j^{\lambda b} \left(\sum_j x_j^a y_j^b\right)^{-\lambda} \geq 1$, for $\lambda \in (0,1]$. Unlike the previous case, these inequalities are tight when there is only a single nonzero value of X, and the inequality may seem weak for nontrivial cases.

The main result of this section is that the cases where Equation 2 holds are *exactly* those cases expressible as a product of inequalities of the above two forms, and that such a representation can be efficiently found. While we have been discussing inequalities involving two sequences, these results apply to inequalities on d sequences, for any positive integer d. For completeness, we restate Theorem 2 in this more general form. The proof of this more general theorem is similar to that of its two-sequence analog, Theorem 2.

Theorem 3. *For $d+1$ fixed sequences $(a)_{1,i} = a_{1,1} \ldots, a_{1,r}, \ \ldots, \ (a)_{d,i} = a_{d,1}, \ldots, a_{d,r}$, and $(c)_i = c_1, \ldots, c_r$, the inequality $\prod_{i=1}^{r} \left(\sum_j \left(\prod_{k=1}^{d} x_{k,j}^{a_{k,i}}\right)\right)^{c_i} \geq 1$ holds for all sets of d finite sequences of positive numbers $(x)_{k,j}$ if and only if it can be expressed as a finite product of positive powers of the Hölder inequalities $\left(\sum_j \left(\prod_{k=1}^{d} x_{k,j}^{a'_k}\right)\right)^{\lambda} \left(\sum_j \left(\prod_{k=1}^{d} x_{k,j}^{a''_k}\right)\right)^{1-\lambda} \geq \sum_j \left(\prod_{k=1}^{d} x_{k,j}^{\lambda a'_k + (1-\lambda) a''_k}\right)$, and the L_p monotonicity inequalities $\left(\sum_j \left(\prod_{k=1}^{d} x_{k,j}^{a'_k}\right)\right)^{\lambda} \leq \sum_j \left(\prod_{k=1}^{d} x_{k,j}^{\lambda a'_k}\right)$, for $\lambda \in [0,1]$.*

Further, there exists an algorithm which, given $d+1$ sequences $(a)_{1,i} = a_{1,1} \ldots, a_{1,r}, \ \ldots, \ (a)_{d,i} = a_{d,1}, \ldots, a_{d,r}$, and $(c)_i = c_1, \ldots, c_r$ describing the inequality, runs in time polynomial in the input description, and either outputs a representation of the desired inequality

as a product of a polynomial number of positive powers of Hölder and L_p monotonicity inequalities, or yields a witness describing d finite sequences of positive numbers $(x)_{k,j}$ that violate the inequality.

The second portion of the theorem—the existence of an efficient algorithm that provides a derivation or refutation of the inequality—is surprising. As the following example demonstrates, it is possible that the shortest sequences x, y that violate the inequality have a number of terms that is *doubly exponential* in the description length of the sequences a, b, c (and exponential in the inverse of the accuracy of the sequences). Hence, in the case that the inequality does not hold, our algorithm cannot be expected to return a pair of counter-example sequences. Nevertheless, we show that it efficiently returns a witness describing such a construction. We observe that the existence of this example precludes any efficient algorithm that tries to approach this problem by solving some linear or convex program in which the variables correspond to the elements of the sequences x, y.

Example 3. *Consider for some $\epsilon \geq 0$ the single-sequence inequality*

$$\Big(\sum_j x_j^{-2}\Big)^{-1}\Big(\sum_j x_j^{-1}\Big)^{3}\Big(\sum_j x_j^{0}\Big)^{-2-\epsilon}\Big(\sum_j x_j^{1}\Big)^{3}\Big(\sum_j x_j^{2}\Big)^{-1} \geq 1,$$

which can be expressed in the form of Equation 1 via the sequences $a = (-2,-1,0,1,2), b = (0,0,0,0,0)$, and $c = (-1,3,-2-\epsilon,3,-1)$. This inequality is true for $\epsilon = 0$ but false for any positive ϵ. However, the shortest counterexample sequences have length that grows as $exp(\frac{1}{\epsilon})$ as ϵ approaches 0. Counterexamples are thus hard to write down, though possibly easy to express—for example, letting $n = 64^{1/\epsilon}$, the sequence x of length $2+n$ consisting of $n, \frac{1}{n}$, followed by n ones violates the inequality.[1]

In the following section we give an overview of the linear programming based proof of Theorem 2. In Section II-B we provide an intuitive interpretation of the computation being performed by the linear program.

A. Proof Overview of Theorem 2

Our proof is based on constructing and analyzing a certain linear program, whose variables ℓ_i represent $\log \sum_j x_j^{a_i} y_j^{b_i}$ for each i in the index set of the sequences $(a)_i, (b)_i, (c)_i$. Letting r denote the size of this index set, the linear program will have r variables, and $poly(r)$ constraints. We will show that if the linear program does *not* have objective value zero then we can construct a counterexample pair of sequences $(x)_j, (y)_j$ for which the inequality is contradicted. Otherwise, if the objective value is zero, then we will consider a solution to the *dual* of this linear program, and interpret this solution as an explicit (finite) combination of Hölder and L_p monotonicity inequalities whose product yields the desired inequality in question. Combined, these results imply that we can efficiently either derive or refute the inequality in all cases.

Given (finite) sequences $(x)_j, (y)_j$, consider the function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(a,b) = \log \sum_j x_j^a y_j^b$. We will call this the *norm graph* of the sequences, and will analyze this function for the remainder of this proof and show how to capture many of its properties via linear programming. The inequality in question, $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i} \geq 1$, is equivalent (taking logarithms) to the claim that for all valid norm graphs ℓ we have $\sum_i c_i \cdot \ell(a_i, b_i) \geq 0$.

The Hölder inequalities explicitly represent the fact that norm graphs ℓ must be convex, namely for each $\lambda \in (0,1)$ and each pair $(a',b'), (a'',b'')$ we have $\lambda\ell(a',b') + (1-\lambda)\ell(a'',b'') \geq \ell(\lambda a' + (1-\lambda)a'', \lambda b' + (1-\lambda)b'')$. The L_p monotonicity inequalities can correspondingly be expressed in terms of norm graphs ℓ, intuitively as "any secant of ℓ (interpreted as a line in 3 dimensions) that intersects the z-axis must intersect it at a nonnegative z-coordinate;" explicitly, for all (a',b') and all $\lambda \in (0,1)$ we have $\lambda\ell(a',b') \leq \ell(\lambda a', \lambda b')$.

Instead of modeling the class of norm graphs directly, we instead model the class of functions that are convex and satisfy the secant property, what we could call "linearized norm graphs": let $\mathfrak{L}$ represent this family of functions from $\mathbb{R}^2$ to $\mathbb{R}$, namely, those functions that are convex and whose secants through the z-axis pass through-or-above the origin. As we will show, this class $\mathfrak{L}$ essentially captures the class of functions $\ell : \mathbb{R}^2 \to \mathbb{R}$ that can be realised as $\ell(a,b) = \log \sum_j x_j^a y_j^b$ for some sequences $(x)_j, (y)_j$, provided we only care about the values of ℓ at a finite number of points (a_i, b_i), and provided we only care about the r-tuple $\ell(a_i, b_i)$ up to scaling by positive numbers. In other words, the inequality $\sum_i c_i \cdot \ell(a_i, b_i) \geq 0$ holds for all norm graphs if and only if it holds for all linearized norm graphs, showing that products of positive powers of Hölder and L_p monotonicity inequalities (used to define the class of linearized norm graphs) exactly capture all norm graph inequalities. In this manner we can reduce the very complicated combinatorial phenomena surrounding Equation 2 to a linear program.

The proof can be decomposed into four steps:

1) We construct a homogeneous linear program ("homogeneous" means the constraints have no additive constants) which we will analyze in the rest of the proof. The linear program has r variables $(\ell)_i$, where feasible points will represent valid r-tuples $\ell(a_i, b_i)$ for linearized norm graphs $\ell \in \mathfrak{L}$. As will become important later, we set the objective

[1] Showing that counterexample sequences must be essentially this long requires technical machinery from the proof of Theorem 2, however one can glean intuition by evaluating the inequality on the given sequence—n, $\frac{1}{n}$, followed by n ones.

function to minimize the expression corresponding to the logarithm of the desired inequality: $\min \sum_i c_i \cdot \ell_i$. Also, as will become important later, we will construct each of the constraints of the linear program so that they are positive linear combinations of logarithms of Hölder and L_p monotonicity inequalities when the $(\ell)_i$ are interpreted as the values of a norm graph at the points (a_i, b_i).

2) We show that for each feasible point, an r-tuple $(\ell)_i$, there is a *linearized* norm graph $\ell : \mathbb{R}^2 \to \mathbb{R}$ that extends $\ell_i = \ell(a_i, b_i)$ to the whole plane, where the function ℓ is the maximum of a finite number of affine functions (namely, of the form $\alpha a + \beta b + \gamma$).

3) For any desired accuracy $\epsilon > 0$, we show that for sufficiently small $\delta > 0$ there is a (regular, not linearized) norm graph ℓ' such that for any $(a, b) \in \mathbb{R}^2$ the scaled version $\delta \cdot \ell'(a, b)$ approximates the linearized norm graph constructed in the previous part, $\ell(a, b)$, to within error ϵ.

Namely, any feasible point of our linear program corresponds to a (possibly scaled) norm graph. Thus, if there exists a feasible point for which the objective function is negative, $\sum_i c_i \cdot \ell_i < 0$, then we can construct sequences $(x)_j, (y)_j$ and a corresponding norm graph $\ell'(a, b) = \log \sum_j x_j^a y_j^b$ for which (because ℓ' can be made to approximate ℓ arbitrarily well at the points (a_i, b_i), up to scaling) we have $\sum_i c_i \cdot \ell'(a_i, b_i) < 0$, meaning that the sequences $(x)_j, (y)_j$ violate the desired inequality. Thus we have constructed the desired counterexample

4) In the other case, where the objective function of the linear program cannot be negative, we note that because by construction we have a homogeneous linear program (each constraint has a right hand side of 0), the optimal objective value must be 0. The solution to the *dual* of our linear program gives a proof of optimality, in a particularly convenient form: the dual solution describes a nonnegative linear combination of the constraints that shows the objective function is always nonnegative, $\sum_i c_i \cdot \ell_i \geq 0$. Recall that, by construction, if each ℓ_i is interpreted as the value of a norm graph at point (a_i, b_i) then each of the linear program constraints is a positive linear combination of the logarithms of certain Hölder and L_p monotonicity inequalities expressed via values of the norm graph. Combining these two facts yields that the inequality $\sum_i c_i \cdot \ell(a_i, b_i) \geq 0$ can be derived as a positive linear combination of the logarithms of certain Hölder and L_p monotonicity inequalities. Exponentiating yields that the desired inequality can be derived as the product of positive powers of certain Hölder and L_p monotonicity inequalities, as desired.

See the full version of the paper for details of the proof.

B. Intuition behind the LP

We provide a pleasing and intuitive interpretation of the computation being performed by the linear program in the proof of Theorem 2. This interpretation is most easily illustrated via an example, and we use one of the inequalities that we encounter in Section III in the the analysis of our instance-optimal tester.

Example 4. *One of the components of the proof of Theorem 1 consists of showing the inequality*

$$\Big(\sum_j x_j^2 y_j^{-2/3}\Big)^2 \Big(\sum_j x_j^2 y_j^{-1/3}\Big)^{-1} \Big(\sum_j x_j\Big)^{-2} \Big(\sum_j y_j^{2/3}\Big)^{3/2} \geq 1.$$

In the notation of Theorem 2, this inequality corresponds to the sequence of four triples $(a_i, b_i, c_i) = (2, -\frac{2}{3}, 2), (2, -\frac{1}{3}, -1), (1, 0, -2), (0, \frac{2}{3}, \frac{3}{2})$. *How does Theorem 2 help us, even without going through the algorithmic machinery presented in the proof?*

Consider the task of proving this inequality via a combination of Hölder and L_p monotonicity inequalities as trying to win the following game. At any moment, the game board consists of some numbers written on the plane (with the convention that every point without a number is interpreted as having a 0), and you win if you can remove all the numbers from the board via a combination of moves of the following two types:

1) *Any two positive numbers can be moved to their weighted mean. (Namely, we can subtract 1 from one location in the plane, subtract 3 from a second location in the plane, and add 4 to a point $\frac{3}{4}$ of the way from the first location to the second location.)*
2) *Any negative number can be moved towards the origin by a factor $\lambda \in (0, 1)$ and scaled by $\frac{1}{\lambda}$. (Namely, we can add 1 to one location in the plane, and subtract 2 from a location half the way to the origin.)*

Thus our desired inequality corresponds to the "game board" having a "2" at location $(2, -\frac{2}{3})$, a "-1" at location $(2, -\frac{1}{3})$, a "-2" at location $(1, 0)$, and a "$\frac{3}{2}$" at location $(0, \frac{2}{3})$. And the rules of the game allow us to push positive numbers together, and push negative numbers towards the origin (scaling them). Our visual intuition is quite good at solving these types of puzzles. (Try it!)

The answer is to first realize that 3 of the points lie on a line, with the "-2" halfway between the "$\frac{3}{2}$" and the "2". Thus we take 1 unit from each of the endpoints and cancel out the "-2". No three points are collinear now, so we need to move one point onto the line formed by the other two: "-1", being negative, can be moved towards the origin, so we move it until it crosses the line formed by the two remaining numbers. This moves it $\frac{1}{3}$ of the way to the origin, thus increasing it from "-1" to "$-\frac{3}{2}$"; amazingly, this number, at position $\frac{2}{3}(2, -\frac{1}{3}) = (\frac{4}{3}, -\frac{2}{9})$ is now $\frac{2}{3}$ of the way from the remaining "$\frac{1}{2}$" at $(0, \frac{2}{3})$ to the number "1" at $(2, -\frac{2}{3})$, meaning that we can remove the final three numbers from the board in a single move, winning the game. We thus

made three moves total, two of the Hölder type, one of the L_p monotonicity type. Reexpressing these moves as inequalities yields the desired derivation of our inequality as a product of powers of Hölder and L_p monotonicity inequalities.

The above example demonstrates how transformative it is to know that the only possible ways of making progress proving a given inequality are by two simple possibilities, thus transforming inequality proving into winning a 2d game with two types of moves. As we show in Theorem 2, this process can be completed automatically in polynomial time via linear programming; but in practice looking at the "2d game board" is often all that is necessary, even for intricate counterintuitive inequalities like the one above.

III. An instance-optimal testing algorithm

In this section we describe our instance-by-instance optimal algorithm for verifying the identity of a distribution, based on independent draws from a distribution. We begin by providing the definitions and terminology that will be used throughout the remainder of the paper. In Section III-B we describe our very simple tester, and give some intuitions and motivations behind its form.

A. Definitions

We use $[n]$ to denote the set $\{1, \ldots, n\}$, and denote a distribution of support size n by $p = p_1, \ldots, p_n$, where p_i is the probability of the ith domain element.

We denote the Poisson distribution with expectation λ by $Poi(\lambda)$, which has probability density function $poi(\lambda, i) = \frac{e^{-\lambda}\lambda^i}{i!}$. We make heavy use of the standard "Poissonization" trick. That is, rather than drawing k samples from a fixed distribution p, we first select $k' \leftarrow Poi(k)$, and then draw k' samples from p. Given such a process, the number of times each domain element occurs is independent, with the distribution of the number of occurrences of the ith domain element distributed as $Poi(k \cdot p_i)$. The independence yielded from Poissonization significantly simplifies many kinds of analysis. Additionally, since $Poi(k)$ is closely concentrated around k: from both the perspective of upper bounds as well as lower bounds, at the cost of only a subconstant factor, one may assume without loss of generality that one is given $Poi(k)$ samples rather than exactly k.

Much of the analysis in this paper centers on L_p norms, where for a vector q, we use the standard notation $||q||_c$ to denote $\left(\sum_i q_i^c\right)^{1/c}$. The notation $||q||_c^b$ is just the bth power of $||q||_c$. For example, $||q||_{2/3}^{2/3} = \sum_i q_i^{2/3}$.

As in Definition 1, we use $p_{-\epsilon}$ to denote the vector of probabilities $p_{\geq s} = p_s, p_{s+1}, \ldots$ defined by sorting the probabilities $p_1 \leq p_2 \leq \ldots$ and letting s be the maximum integer such that $\sum_{i<s} p_i \leq \epsilon$. Additionally, we use $p^{-\max}$ to denote the vector of probabilities with the maximum probability omitted. Hence the frequently used notation $p_{-\epsilon}^{-\max}$ is the vector of probabilities obtained from p by both removing the largest entry, and removing the smallest entries until the weight of the small entries removed is at most ϵ.

B. An optimal tester

Our testing algorithm is extremely simple, and takes the form of a simple statistic that is similar to Pearson's chi-squared statistic, though differs in two crucial ways. Given a set of k samples, with X_i denoting the number of occurrences of the ith domain element, and p_i denoting the probability of drawing the ith domain element from distribution p, the Pearson chi-squared statistic is given as $\sum_i \frac{1}{p_i}(X_i - kp_i)^2$. Adding a constant does not change the behavior of the statistic, and it will prove easier to compare with our statistic if we subtract k from each term, yielding the following:

$$\sum_i \frac{(X_i - kp_i)^2 - kp_i}{p_i}. \tag{3}$$

In the Poissonized setting (where the number of samples is drawn from a Poisson distribution of expectation k), if the samples are drawn from distribution p, then the expectation of this chi-squared statistic is 0 because in that case X_i is distributed according to a Poisson distribution of expectation kp_i, and hence has variance kp_i. Our testing algorithm is, essentially, obtained by modifying this statistic in two ways: replacing the second occurrence of kp_i with X_i, and changing the scaling factor from $1/p_i$ to $1/p_i^{2/3}$:

$$\sum_i \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}}. \tag{4}$$

Note that this statistic still has the property that its expectation is 0 if the samples are drawn from distribution p. The following examples motivate these two modifications.

Example 5. *Let p be the distribution with $p_1 = p_2 = 1/4$, and the remaining half of its probability mass composed of $n/2$ domain elements, each occurring with probability $1/n$. If we draw $k = n^{2/3}$ samples from p, the contribution of the $n/2$ small elements to the variance of Pearson's statistic (Equation 3) is $\approx \frac{n}{2}(n^{-1/3}n^2) = \Omega(n^{8/3})$, and the standard deviation would be $\Omega(n^{4/3})$. If the k samples were not drawn from p, and instead were drawn from distribution q that is identical to p, except with $p_1 = 1/8$ and $p_2 = 3/8$, then the expectation of Pearson's statistic would be $O(n^{4/3})$, though this signal might be buried by the $\Omega(n^{4/3})$ standard deviation due to the small domain elements.*

The above example illustrates that the scaling factor $1/p_i$ in Pearson's chi-squared statistic places too much weight on the small elements, and motivates a smoother scaling factor. There does not seem to be any intuition for the $2/3$ exponent in our statistic—it comes out of optimizing the interplay between various inequalities in the analysis, and is cleanly revealed by our inequality prover of Section II. Intuitive reasoning from the perspective of the tester seems to lead to

a scaling factor of $p_i^{1/2}$, whereas intuitive reasoning from the perspective of the lower bounds seems to lead to a scaling factor of $p_i^{3/4}$. Both intuitions turn out to be misleading, and the correct scaling of $p_i^{2/3}$ was unexpected.

The following example illustrates a second benefit of our statistic of Equation 4 over the chi-squared statistic:

Example 6. *Let p be the distribution with $p_1 = 1-1/n$, and the remaining $1/n$ probability mass is evenly split among n domain elements each with probability $1/n^2$. If we draw $100 \cdot n$ samples, we are likely to see roughly 100 ± 10 of the "rare" domain elements, each exactly once. Such domain elements will have a huge contribution to the variance of Pearson's chi-squared statistic—a contribution of $\Omega(n^2)$. On the other hand, these domain elements contribute almost nothing to the variance of our statistic, because the contribution of such domain elements is essentially $(X_i^2 - X_i)p_i^{-2/3}$, which is 0 if X_i is 0 or 1 and with overwhelming probability, none of these "rare" domain elements will occur more than once. Hence our statistic is extremely robust to seeing rare things either 0 or 1 times, and this significantly reduces the variance of our statistic.*

We now formally define our tester. The tester essentially just computes the statistic of Equation 4, though one also needs to shave off a small $O(\epsilon)$ portion of the distribution p before computing it, and also verify that not too much probability mass lies on this supposedly small portion that was removed.

Throughout the remainder of the paper, we will assume, without loss of generality, that the domain elements of p are sorted in increasing order of probability. Let s be the largest integer such that $\sum_{i<s} p_i \leq \epsilon/8$, and for each domain element i let X_i be the number of times element i occurs in the sample. Note that $p_{\geq s}$ is by definition the same as $p_{-\epsilon/8}$ as defined in Definition 1, though it will be easier to work explicitly with s in the proofs.

AN INSTANCE-OPTIMAL TESTER

Given a parameter $\epsilon > 0$ and a set of k samples drawn from q, let X_i represent the number of times the ith domain element occurs in the samples. Assume wlog that the domain elements of p are sorted in increasing order of probability, and let s be the largest integer such that $\sum_{i<s} p_i \leq \epsilon/8$:

1) If $\sum_{i \geq s, i \neq \arg\max p_i} \left[(X_i - kp_i)^2 - X_i\right] p_i^{-2/3} > 4k||p_{\geq s}^{-\max}||_{2/3}^{1/3}$, or
2) If $\sum_{i<s} X_i > \frac{3}{16}\epsilon k$, then output "DIFFERENT", else output "SAME"

For convenience, we restate Theorem 1, characterizing the performance of the above tester.

Theorem 1. *There exist constants c_1, c_2 such that for any $\epsilon > 0$ and any known distribution p, for any unknown distribution q, our tester will distinguish $q = p$ from $||p-q||_1 \geq \epsilon$ with probability $2/3$ when run on a set of at least $c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon/16}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ samples drawn from q, and no tester can do this task with probability at least $2/3$ with a set of fewer than $c_2 \cdot \max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ samples.*

In general, the "yes" case of the theorem, where $q = p$, can be relaxed to a "tolerant testing" condition $||p-q||_1 \leq O(\frac{1}{k})$ where $k = c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon/16}^{-\max}||_{2/3}}{\epsilon^2}\right\}$ is the number of samples used. This kind of tolerant testing result is true for *any* tester, because statistical distance is subadditive on product distributions, so a change of $\frac{c}{k}$ in the distribution p can induce a change of at most c on the distribution of the output of any testing algorithm that uses k samples. A more refined analysis of our tester (or a tester tailored to the tolerant regime) yields better bounds in some cases. However, the problem of distinguishing $||p-q|||_1 \leq \epsilon_1$ from $||p-q|||_1 \geq \epsilon_2$ enters a very different regime when ϵ_1 is not much smaller than ϵ_2, and many more samples are required. (These problems are very related to the task of *estimating* the distance from q to the known distribution p.) For any constants $\epsilon_1 < \epsilon_2$, it requires $\Theta(\frac{n}{\log n})$ samples to distinguish $||p-q|||_1 \leq \epsilon_1$ from $||p-q|||_1 \geq \epsilon_2$ when p is the uniform distribution on n elements, many more than the $\sqrt{n}$ needed here [16], [17].

Before discussing the proof approach, we provide some intuition behind the form of the sample complexity, $\max\left\{\frac{1}{\epsilon}, \frac{||p_{-\epsilon}^{-\max}||_{2/3}}{\epsilon^2}\right\}$. The maximum with $\frac{1}{\epsilon}$ only very rarely comes into play: the $\frac{2}{3}$ norm of a vector is always at least its 1 norm, so the max with $\frac{1}{\epsilon}$ only takes over from $||p_{-\epsilon}^{-\max}||_{2/3}/\epsilon^2$ if p is of the very special form where removing its max element and its smallest ϵ mass leaves less than ϵ probability mass remaining; the max expression thus prevents the sample size in the theorem from going to 0 in extreme versions of this case.

The subscript and superscript in $||p_{-\epsilon}^{-\max}||_{2/3}$ each *reduce* the final value, and mark two ways in which the problem might be "unexpectedly easy". To see the intuition behind these two modifications in the vector of probabilities, note that if the distribution p contains a single domain element p_m that comprises the majority of the probability mass, then in some sense it is hard to hide changes in p: at least half of the discrepancy between p and q must lie in other domain elements, and if these other domain elements comprise just a tiny fraction of the total probability mass, then the fact that half the discrepancy is concentrated on a tiny fraction of the distribution makes recognizing such discrepancy easier.

On the other hand, having many small domain elements makes the identity testing problem harder, as indicated by the $L_{2/3}$ norm, however only "harder up to a point". If most of the $L_{2/3}$ norm of p comes from a portion of

the distribution with tiny L_1 norm, then it is also hard to "hide" much discrepancy in this region: if a portion of the domain consisting of $\epsilon/3$ total mass in p has discrepancy ϵ between p and q, then the probability mass of these elements in q must total at least $\frac{2}{3}\epsilon$ by the triangle inequality, namely at least twice what we would expect if $q = p$; this discrepancy is thus easy to detect in $O(\frac{1}{\epsilon})$ samples. Thus discrepancy cannot hide in the very small portion of the distribution, and we may effectively ignore the small portion of the distribution when figuring out how hard it is to test discrepancy.

In these two ways—represented by the subscript and superscript of $p_{-\epsilon}^{-\max}$ in our results—the identity testing problem may be "easier" than the simplified $O(\frac{||p||_{2/3}}{\epsilon^2})$ bound. But our corresponding lower bound shows that these are the only ways.

C. Analysis of the tester

The core of the proof of the algorithmic direction of Theorem 1 is an application of Chebyshev's inequality: first arguing that if the samples were drawn from a distribution q with $||p-q||_1 \geq \epsilon$, then the expectation of the statistic in question is large in comparison to the variance, and if the samples were drawn from p, then the variance is sufficiently small so as to not overlap significantly with the likely range of the statistic in the case that $||p-q||_1 \geq \epsilon$. In order to prove the desired inequalities relating the expectation and the variance, we reexpress these inequalities in terms of the two sequences of positive numbers $p = p_1, p_2, \ldots$, and $\Delta = \Delta_1, \Delta_2, \ldots$, with $\Delta_i := |p_i - q_i|$, leading to an expression that is the sum of five inequalities essentially of the canonical form $\prod_i \left(\sum_j p_j^{a_i} \Delta_j^{b_i}\right)^{c_i} \geq 1$. The machinery of Section II thus yields an easily verifiable derivation of the desired inequalities as a product of positive powers of Hölder type inequalities, and L_p monotonicity inequalities. See the full version for proofs.

IV. Lower bounds

In this section we show how to construct distributions that are very hard to distinguish from a given distribution p despite being far from p, establishing the lower bound portion of Theorem 1. Explicitly, we will construct a distribution over distributions, that we will call Q_ϵ, such that most distributions in Q_ϵ are far from p, yet k samples from a randomly chosen member of Q_ϵ will be distributed very close to the distribution of k samples from p. Analyzing the statistics of such sampling processes can be enormously involved (see for example the lower bounds of [16], which involve deriving new and general central limit theorems in high dimensions).

In this paper, however, we show that the statistics of k samples from a randomly chosen distribution from Q_ϵ can be captured much more directly, by a product distribution over the i domain elements of a "coin flip between Poisson distributions." Thus we can analyze this process dimension-by-dimension and sum the distances. That is, if d_i is the distance between what happens for the ith domain element given k samples from p versus k samples from the product distribution "capturing" Q_ϵ, we can sum these up to bound the probability of distinguishing p from Q_ϵ by $\sum_i d_i$. However, this is not good enough for us; since the actual probability of distinguishing these two cases for an ideal tester is more like the L_2 norm of these d_i distances instead of the L_1 norm, to achieve a tight result we need something like $\sqrt{\sum_i d_i^2}$.

To accomplish this, we analyze all distances below via the *Hellinger distance*,

$$H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}.$$

Hellinger distance has two properties perfectly suited for our task: its *square* is subadditive on product distributions (meaning it combines via the L_2 norm instead of the L_1 norm), and the Hellinger distance (times $\sqrt{2}$) bounds the statistical distance.

We first give a lemma characterizing the Hellinger distance between the "coin flip between Poisson distributions" mentioned above and a regular Poisson distribution. We then show how a product distribution of these coin flip distributions forms a powerful class of testing lowerbounds, Theorem 4, which has already found use in [7].

Let $Poi(\lambda \pm \epsilon)$ denote the probability distribution with pdf over nonnegative integers i: $\frac{1}{2}poi(\lambda+\epsilon) + \frac{1}{2}poi(\lambda-\epsilon)$, which is only defined for $\epsilon \leq \lambda$.

Lemma 1. *$H(Poi(\lambda), Poi(\lambda \pm \epsilon)) \leq c \cdot \frac{\epsilon^2}{\lambda}$ for constant c.*

This lemma, proved in the full version of the paper, is a crucial ingredient in the proof of the following general lower bound.

Theorem 4. *Given a distribution p, and associated values ϵ_i such that $\epsilon_i \in [0, p_i]$ for each domain element i, define the distribution over distributions Q_ϵ by the process: for each domain element i, randomly choose $q_i = p_i \pm \epsilon_i$, and then normalize q to be a distribution. Then there exists a constant c such that it takes at least $c\left(\sum_i \frac{\epsilon_i^4}{p_i^2}\right)^{-1/2}$ samples to distinguish p from Q_ϵ with success probability $2/3$. Further, with probability at least $1/2$, the L_1 distance between a random distribution from Q_ϵ and p is at least $\min\{(\sum_{i \neq \arg\max \epsilon_i} \epsilon_i), \frac{1}{2}\sum_i \epsilon_i\}$.*

The lower bound portion of Theorem 1 follows from the above theorem by appropriately choosing the sequence ϵ_i. The details are contained in the full version of the paper.

Proof of Theorem 4: Consider the following related distributions, which emulate the number of times each domain element is seen if we take $Poi(k)$ samples: first randomly generate $\bar{q}_i = p_i \pm \epsilon_i$ without normalizing, and

then for each i draw a sample from $Poi(\bar{q}_i \cdot k)$; compare this to, for each i, drawing a sample from $Poi(p_i \cdot k)$. We note that with probability at least $\frac{1}{2}$, we have $\sum_i \bar{q}_i \geq 1$; further, with probability at least $\frac{1}{2}$ a Poisson distribution with parameter at least k will yield a sample at least k. Thus with probability at least $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, the number of samples from the first Poisson process emulating Q_ϵ will be at least k; with probability $\frac{1}{2}$ the number of samples from the simpler second Poisson process emulating p will be at least k. Thus with probability at least $\frac{1}{8}$ we have "a set of at least k samples" from both distributions.

If it were possible to distinguish p from Q_ϵ in k samples with probability $2/3$, then we could distinguish these two Poisson processes with probability $\frac{1}{2} + \frac{1}{6 \cdot 8}$. However, these two Poisson processes are both product distributions, and we can thus compare them from the fact that the squared Hellinger distance is subadditive on product distributions. For each component i, the squared Hellinger distance is $H(Poi(kp_i), Poi(k[p_i \pm \epsilon_i]))^2$ which by Lemma 1 is at most $c_1 k^2 \frac{\epsilon^4}{p_i^2}$. Summing over i and taking the square root yields a bound on the Hellinger distance of $k \left(c_1 \sum_i \frac{\epsilon^4}{p_i^2} \right)^{1/2}$, which thus bounds the L_1 distance. Thus for small enough c, when k satisfies the bound of the theorem, the statistical distance between a set of k samples drawn from p versus drawn from a random distribution of Q_e must be arbitrarily small, and the two cannot be distinguished.

We now analyze the second part of the theorem, bounding the distance between a distribution $q \leftarrow Q_\epsilon$ and p. We note that the total excess probability mass in the process of generating q that must subsequently be removed (or added, if it is negative) by the normalization step is distributed as $\sum_i \pm\epsilon_i$, and thus by the triangle inequality, the L_1 distance between q and p is at least as large as a sample from $\sum_i \epsilon_i - |\sum_i \pm\epsilon_i|$. We thus show that with probability at least $1/2$, a random value from $|\sum_i \pm\epsilon_i|$ is at most either $\max_i \epsilon_i$ or $\frac{1}{2}\sum_i \epsilon_i$.

Consider the sequence ϵ_i as sorted in descending order. We have two cases. Suppose $\epsilon_1 \geq \frac{1}{2}\sum_i \epsilon_i$. Consider the random number $|\sum_i \pm\epsilon_i|$, where without loss of generality the plus sign is chosen for ϵ_1. With probability at least $1/2$, the sum of the remaining elements will be ≤ 0; further, by the assumption of this case, this sum cannot be smaller than $-2\epsilon_1$. Thus the sum of all the elements has magnitude at most ϵ_1 with probability at least $1/2$.

In the other case, $\epsilon_1 < \frac{1}{2}\sum_i \epsilon_i$. Consider randomly choosing signs $s_i \in \{-1, +1\}$ for the elements iteratively, stopping *before* choosing the sign for the first element j for which it would be possible for $\left| (\sum_{i<j} s_i \epsilon_i) \pm \epsilon_j \right|$ to exceed $\frac{1}{2}\sum_i \epsilon_i$. Since by assumption $\epsilon_1 < \frac{1}{2}\sum_i \epsilon_i$, we have $j \geq 2$. Without loss of generality, assume $\sum_{i<j} s_i \epsilon_i \geq 0$. We have $\sum_{i<j} s_i \epsilon_i < \frac{1}{2}\sum_i \epsilon_i$, and (by symmetry) with probability at most $1/2$ the sum of the remaining elements with randomly chosen signs will be positive. Further, since $s_1\epsilon_1 + s_2\epsilon_2 + \ldots + s_{j-1}\epsilon_{j-1} + \epsilon_j \geq \frac{1}{2}\sum_i \epsilon_i$, we have $s_1\epsilon_1 + s_2\epsilon_2 + \ldots + s_{j-1}\epsilon_{j-1} - \sum_{i \geq j} \epsilon_i \geq -\frac{1}{2}\sum_i \epsilon_i$, for otherwise if this last inequality was "$<$" we could subtract these last two equations to conclude $\epsilon_j + \sum_{i \geq j} \epsilon_i > \sum_i \epsilon_i$, which contradicts the facts that $s_1 \geq s_j$ and $j \geq 2$. Thus a random choice of the remaining signs starting with s_j will yield a total sum at most $\frac{1}{2}\sum_i \epsilon_i$, with probability at least $1/2$, as desired. ■

REFERENCES

[1] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. In *Conference on Learning Theory (COLT)*, 2011.

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

[3] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: lower bounds and applications. In *Symposium on Theory of Computing (STOC)*, 2001.

[4] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 2005.

[5] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS*, 2001.

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

[7] S.O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, 2014.

[8] M. Charikar, S. Chaudhuri, R. Motwani, and V.R. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, 2000.

[9] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, ECCC*, 2000.

[10] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, 2006.

[11] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[12] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. on Inf. Theory*, 50(9):2200–2203, 2004.

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

[14] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

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

[16] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$–sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, 2011.

[17] G. Valiant and P. Valiant. The power of linear estimators. In *FOCS*, 2011.

[18] P. Valiant. Testing symmetric properties of distributions. In *STOC*, 2008.