AN AUTOMATIC INEQUALITY PROVER AND INSTANCE OPTIMAL IDENTITY TESTING*

GREGORY VALIANT† AND PAUL VALIANT‡

Abstract. We consider the problem of verifying the identity of a distribution: Given the description of a distribution over a discrete finite or countably infinite support, $p = (p_1, p_2, \ldots)$, 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 the known distribution p and error parameter ϵ, 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. This result 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 $O(\frac{\sqrt{n}\,\mathrm{polylog}\, n}{\epsilon^4})$ and matching the (tight) results for the case that p is the uniform distribution of 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 of triples $(a, b, c)_i = (a_1, b_1, c_1), \ldots, (a_r, b_r, c_r)$ 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 triples $(a, b, c)_i = (1, 0, \frac{1}{2})$, $(0, 1, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, -1)$. Our characterization is constructive and algorithmic, leveraging linear programming to prove or refute an inequality, which would otherwise have to be investigated, through trial and error, by hand. We hope the computational nature of our characterization will be useful to others, and facilitate analyses like the one here.

Key words. Hypothesis testing, identity testing, instance optimal, Holder's inequality

AMS subject classifications. 68Q32, 26D15, 62G10

1. 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

*A preliminary version of this work appeared at the IEEE Symposium on Foundations of Computer Science, 2014.

Funding: Gregory's work was supported by NSF CAREER Award CCF-1351108, and Paul's work was supported by a Sloan fellowship. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing, at UC Berkeley.

†Stanford University, Stanford, CA (gvaliant@cs.stanford.edu, http://theory.stanford.edu/~valiant/).

‡Brown University, Providence, RI (pvaliant@gmail.com, http://cs.brown.edu/~pvaliant/).

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 $p = \mathrm{Unif}[n]$), 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 theoretical computer science, 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" are 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 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. We refer the reader to [22] for an example of subsequent work in this direction.

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

Definition 1. *For a probability distribution p over a discrete support, let $p^{-\max}$ denote the vector of probabilities obtained from p by removing the entry corresponding to the element of largest probability (with ties broken arbitrarily if there are multiple such elements). For $\epsilon > 0$, define $p_{-\epsilon}$ to be the vector obtained from p by removing the domain elements of smallest probability mass under p, and stopping just before more than ϵ probability mass is removed.*

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

Throughout, we use the standard notation for the L_p norm of a vector: given a vector x, and a real number α we define the α norm of x as

$$\|x\|_\alpha = \left(\sum_i x_i^\alpha\right)^{1/\alpha}$$

Our main result is the following:

Theorem 2. *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* $c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\right\}$ *samples drawn from* q*, and no tester can do this task with probability at least* $2/3$ *with a set of fewer than* $c_2 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-2\epsilon}\|_{2/3}}{\epsilon^2}\right\}$ *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 $\|p_{-2\epsilon}\|_{2/3}$ 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 $p_{-2\epsilon}$, though not from $p_{-\epsilon/16}$).[1]

Because our tester is constant-factor tight, the subscript and superscript on p and the max with $\frac{1}{\epsilon}$ in the sample complexity $\max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-O(\epsilon)}\|_{2/3}}{\epsilon^2}\right\}$ 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 superscript only reduce the value of the norm: $\|p^{-\max}_{-2\epsilon}\|_{2/3} < \|p_{-2\epsilon}\|_{2/3} \leq \|p_{-\epsilon/16}\|_{2/3} \leq \|p\|_{2/3}$, 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 [6], and matching the tight bound on the number of samples required for testing the uniform distribution given in [17].

The core of our testing algorithm is an extremely simple statistic that is similar to Pearson's chi-squared statistic. 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 $\sum_i \frac{(X_i-kp_i)^2-kp_i}{p_i}$. Our testing algorithm is, essentially, obtained by modifying this statistic in two crucial ways: replacing the second occurrence of kp_i with X_i (which has expectation kp_i when drawing samples from p), and changing the scaling factor from $1/p_i$ to $1/p_i^{2/3}$:

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

Our simple testing algorithm is stated below:

[1]In the language of the abstract, Theorem 2 defines a function $f(p,\epsilon)$ characterizing the sample complexity of testing the identity of p, tight up to a factor of 32 in the error ϵ and some constant c_1/c_2 in the number of samples. Interestingly, since the function $f(p,\epsilon)$ grows at least inversely in ϵ as ϵ goes to 0, we can merge the two constants into a single multiplicative constant in the error ϵ and say that the right number of samples for testing the identity of p to within ϵ must lie between $f(p, 32\frac{c_1}{c_2}\cdot\epsilon)$ and $f(p,\epsilon)$. This is a cleaner result, in some sense; however, of the two parameters—the accuracy ϵ and the sample size k—it is often perhaps more important to have precise control of the accuracy, so we wanted to emphasize that while our results are constant-factor-tight, the constant, 32, in front of ϵ is explicit, and can be made small.

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 non-increasing order of probability. Define $s = \min\{i : \sum_{j>i} p_j \leq \epsilon/8\}$, and let $M = \{2, \ldots, s\}$, and $S = \{s+1, s+2, \ldots\}$. (Note that $p_M = p^{-\max}_{-\epsilon/8}$.)

1. If $\sum_{i \in M} \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}} > 4k\|p_M\|_{2/3}^{1/3}$, or
2. If $\sum_{i \in S} X_i > \frac{3}{16}\epsilon k$, then output "$\|p-q\|_1 \geq \epsilon$", else output "$p = q$".

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 of triples $(a,b,c)_i = (a_1,b_1,c_1), \ldots, (a_r,b_r,c_r)$ 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, the class of valid inequalities 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 triples* $(a,b,c)_i = (1,0,\frac{1}{2}),\ (0,1,\frac{1}{2}),\ (\frac{1}{2},\frac{1}{2},-1)$. *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* $(a,b,c)_i = (1,0,\lambda),\ (0,1,1-\lambda),\ (\lambda,1-\lambda,-1)$.

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 two triples* $(a,b,c)_i = (p,0,1),(1,0,-p)$. *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 3. *For a fixed sequence of triples* $(a,b,c)_i = (a_1,b_1,c_1), \ldots (a_r,b_r,c_r)$, *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*

Hölder inequalities of the form

$$\left(\sum_j x_j^{a'} y_j^{b'}\right)^{\lambda} \cdot \left(\sum_j x_j^{a''} y_j^{b''}\right)^{1-\lambda} \geq \sum_j x_j^{\lambda a' + (1-\lambda)a''} y_j^{\lambda b' + (1-\lambda) b''},$$

and L_p monotonicity inequalities of the form $\left(\sum_j x_j^a y_j^b\right)^{\lambda} \leq \sum_j x_j^{\lambda a} y_j^{\lambda b}$*, where* $\lambda \in [0,1]$.

We state this theorem for pairs of sequences $(x)_j, (y)_j$, of positive numbers, although an analogous statement (Theorem 4 stated in Section 2) holds for any number of positive 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 3 is algorithmic in nature; in fact, we describe an algorithm which, when given the sequence of triples $(a, b, c)_i$ 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 sequence of triples $(a, b, c)_i$—see Example 3).

The characterization of Theorem 3 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.html.

1.1. 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 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., [7, 5, 12, 15, 16, 4, 9]), and on establishing lower bounds [18, 23]. 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 [20, 21] resolved the worst-case sample complexities of estimating many of these symmetric properties. Also see [19] 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 [11] 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 $\Theta(\frac{\sqrt{n}}{\epsilon^2})$ on the number of samples subsequently given by Paninski [17]. For the more general problem of verifying the identity of an arbitrary distribution, Batu et al. [6], showed that for worst-case distributions of support size n, $O(\frac{\sqrt{n}\,\mathrm{polylog}\,n}{\epsilon^4})$ samples are sufficient. Since the publication of this current paper, Diakonikolis et al. [10], considered the problem of identity testing under various assumptions about the *shape* of the distribution, including, for example, assuming the distribution is monotone, unimodal, multimodal, or piecewise constant, etc., relative to an ordering of the domain elements; for distributions assumed to be piecewise constant with t pieces, they show a tester with $O(\frac{\sqrt{t}}{\epsilon^2})$ samples, which, letting $t = n$ yields a $O(\frac{\sqrt{n}}{\epsilon^2})$-sample tester in our setting, which has worst-case optimal dependence on n and ϵ (but is not instance-optimal).

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 can be used to test identity.

The form of our tester may be seen as rather similar to those in [1, 2, 8], which considered testing whether two distributions were close or not when *both* distributions are unknown. The testers in those papers 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. These testers are considerable simpler than many of the proposed testers in other works (including [10] and the initial pioneering work [6]), which proceed by subdividing the domain into a super-constant number of partitions, and applying tests to each partition separately. From a technical perspective, our lower bounds leverage Hellinger distance to introduce a flexible class of lower bound instances, which yield the tight results of this work, and were also employed to give the lower bounds in [8].

1.2. Organization. We begin with our characterization of the class of inequalities, as we feel that this tool may be useful to the broader community; this first section is entirely self-contained. Section 3.1 contains the definitions and terminology relevant to the distribution testing portion of the paper, and Section 3.2 describes our very simple instance-optimal distribution identity testing algorithm, and provides some context and motivation for the algorithm. Section 4 discusses the lower bounds, establishing the optimality of our tester.

2. A class of inequalities generalizing Hölder's inequality 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 3.

The basic question we resolve is: for what sequences of triples $(a,b,c)_i$ 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 $(a,b,c)_i$, 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 (positive powers of) 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 3 in this more general form. The proof of this more general theorem is similar to that of its two-sequence analog, Theorem 3.

THEOREM 4. *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 Hölder inequalities of the form $\left(\sum_j \left(\prod_{k=1}^d x_{k,j}^{a'_k}\right)\right)^{\lambda} \left(\sum_j \left(\prod_{k=1}^d x_{k,j}^{a''_k}\right)\right)^{1-\lambda} \geq \sum_j \left(\prod_{k=1}^d x_{k,j}^{\lambda a'_k + (1-\lambda)a''_k}\right)$, and L_p monotonicity inequalities of the form $\left(\sum_j \left(\prod_{k=1}^d x_{k,j}^{a'_k}\right)\right)^{\lambda} \leq \sum_j \left(\prod_{k=1}^d x_{k,j}^{\lambda a'_k}\right)$, where $\lambda \in [0,1]$, and where a'_k, a''_k can be any real numbers.*

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 sequence of triples $(a,b,c)_i$ (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 triples $(a,b,c)_i = (-2,0,-1)$, $(-1,0,3)$, $(0,0,-2-\epsilon)$, $(1,0,3)$, $(2,0,-1)$. 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.[2]

In the following section we give an overview of the linear programming based proof of Theorem 3, and then give the formal proof in Section 2.2. In Section 2.3 we

[2]Showing that counterexample sequences must be essentially this long requires technical machinery from the proof of Theorem 3, however one can glean intuition by evaluating the inequality on the given sequence—n, $\frac{1}{n}$, followed by n ones.

provide an intuitive interpretation of the computation being performed by the linear program.

2.1. Proof overview of Theorem 3. Our proof is based on constructing and analyzing a certain linear program, whose variables, which we denote by ℓ_i, represent $\log \sum_j x_j^{a_i} y_j^{b_i}$ for each i in the index set of triples $(a, b, c)_i$. 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 2-dimensional function $\ell(a, b)$ the *norm graph* of the sequences $(x)_j, (y)_j$, 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 $\sum_i c_i \cdot \ell(a_i, b_i) \geq 0$ for every norm graph ℓ that can be realized via sequences $(x)_j, (y)_j$.

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 the graph 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, which we 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 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, further, the function ℓ is the maximum of a finite number of affine functions (functions 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 minimum objective function of the linear program is nonnegative, 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.

The following section provides the proof details for the above overview.

2.2. Proof of Theorem 3. Given r triples, $(a_1, b_1, c_1), \ldots, (a_r, b_r, c_r)$, consider the linear program with r variables denoted by $\ell_1, \ldots, \ell_r$ with objective function $\min \sum_i c_i \cdot \ell_i$. For each index $k \in [r]$ we add linear constraints to enforce that the point (a_k, b_k, ℓ_k) in $\mathbb{R}^3$ lies on the lower convex hull of the points (a_i, b_i, ℓ_i) *and* the extra point $(2a_k, 2b_k, 2\ell_k)$. Recall that the parameters (a_i, b_i) are constants, so we may use them arbitrarily to set up the linear program. Explicitly, for each triple, pair, or singleton from the set $\{(a_i, b_i) : i \neq k\} \cup \{(2a_k, 2b_k)\}$ that have a unique convex combination that equals (a_k, b_k), we add a constraint that the corresponding combination of their associated z-values (i.e. the corresponding ℓ_i or $2\ell_k$) must be greater than or equal to ℓ_k. The total number of constraints is thus $O(r^4)$. We note that these are homogeneous constraints—there are no additive constants. Intuitively, we are expressing all our constraints on the linearized norm graph in this convex hull form: the Hölder inequalities are naturally convexity constraints, and by adding these "fictitious" points $(2a_k, 2b_k, 2\ell_k)$, the L_p monotonicity inequalities can now also be treated as convexity constraints.

We now begin our proof of one direction of Theorem 3—that if the above linear program has objective function value 0, then the desired inequality can be expressed as the product of a finite number of Hölder and L_p monotonicity inequalities. As a first step, we establish that each of the above constraints can be expressed as a positive linear combination of these two types of inequalities:

LEMMA 5. *Each of the above-described constraints can be expressed as a positive linear combination of the logarithms of Hölder and L_p monotonicity inequalities.*

Proof. Consider, first, the case when the convex combination does not involve the special point $(2a_k, 2b_k)$. Thus there are indices $i1, i2, i3$ and nonnegative constants $\lambda_1, \lambda_2, \lambda_3$ with $\lambda_1 + \lambda_2 + \lambda_3 = 1$ for which $\lambda_1(a_{i1}, b_{i1}) + \lambda_2(a_{i2}, b_{i2}) + \lambda_3(a_{i3}, b_{i3}) = (a_k, b_k)$ and we want to conclude a kind of "three-way Hölder inequality", that $\lambda_1\ell(a_{i1}, b_{i1}) + \lambda_2\ell(a_{i2}, b_{i2}) + \lambda_3\ell(a_{i3}, b_{i3}) \geq \ell(a_k, b_k)$, for any norm graph ℓ. If two of the three λ's are 0 (without loss of generality $\lambda_2 = \lambda_3 = 0$) then $\lambda_1 = 1$ and $(a_{i1}, b_{i1}) = (a_k, b_k)$ making the inequality trivially $\ell(a_k, b_k) \geq \ell(a_k, b_k)$. If only one of the λ's is 0, without loss of generality $\lambda_3 = 0$ and $\lambda_1 + \lambda_2 = 1$, making the desired inequality a standard Hölder inequality,

$$\lambda_1\ell(a_{i1}, b_{i1}) + (1-\lambda_1)\ell(a_{i2}, b_{i2}) \geq \ell\Big(\lambda_1 a_{i1} + (1-\lambda_1)a_{i2}, \lambda_1 b_{i1} + (1-\lambda_1)b_{i2}\Big). \tag{3}$$

In the case that all three λ's are nonzero, we derive the result by replacing λ_1 with $\bar{\lambda}_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}$ in Equation 3 and multiplying both sides of the inequality by $\lambda_1 + \lambda_2$, and then adding the following Hölder inequality:

$$(\lambda_1+\lambda_2)\ell\Big(\bar{\lambda}_1 a_{i1} + (1-\bar{\lambda}_1)a_{i2}, \bar{\lambda}_1 b_{i1} + (1-\bar{\lambda}_1)b_{i2}\Big) + \lambda_3\ell(a_{i3}, b_{i3}) \geq \ell(a_k, b_k). \tag{4}$$

Finally, we consider the case where $(2a_k, 2b_k, 2\ell(a_k, b_k))$ is used; we only consider the triple case as the other cases are easily dealt with. Thus we have that a convex combination with coefficients $\lambda_1 + \lambda_2 + \lambda_3 = 1$ of the points (a_{i1}, b_{i1}), (a_{i2}, b_{i2}), $(2a_k, 2b_k)$ equals (a_k, b_k). We thus must derive the somewhat odd inequality $\lambda_1\ell(a_{i1}, b_{i1}) + \lambda_2\ell(a_{i2}, b_{i2}) + 2\lambda_3\ell(a_k, b_k) \geq \lambda(a_k, b_k)$. As above, substitute $\bar{\lambda}_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}$ for λ_1 in Equation 3 and multiply by $\lambda_1 + \lambda_2$; this time, add to it $\lambda_1 + \lambda_2$ times the L_p monotonicity inequality

$$\frac{1-2\lambda_3}{\lambda_1+\lambda_2}\ell(a_k, b_k) \leq \ell\Big(\frac{1-2\lambda_3}{\lambda_1+\lambda_2}a_k, \frac{1-2\lambda_3}{\lambda_1+\lambda_2}b_k\Big). \tag{5}$$

Everything is seen to match up since the points at which the ℓ functions on the right hand sides of Equations 3 and 5 are evaluated are equal (since $(1-2\lambda_3)a_k = \lambda_1 a_{1i} + \lambda_2 a_{2i}$ from the original interpolation). ∎

Given the above lemma, the proof of one direction of Theorem 3 now follows easily—essentially following from step 4 of the proof overview given in the previous section.

LEMMA 6. *If the objective value of the linear program is non-negative, then it must be zero, and the inequality $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i}$ can be expressed as a product of at most $O(r^4)$ Hölder and L_p monotonicity inequalities.*

Proof. Recall that since the linear program is homogeneous (each constraint has a right hand side of 0), the optimal objective value cannot be larger than 0, and hence if the objective value is not negative, it 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 nonnegative coefficients for each of the primal inequality constraints, such that when we add up these constraints scaled by these coefficients, we find $\sum_i c_i \cdot \ell_i \geq 0$—a lower bound on our primal objective function. Recall that, by construction, if each ℓ_i is interpreted as the value of a norm graph at point (a_i, b_i), then Lemma 5 shows that 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 Hölder and L_p monotonicity inequalities, as claimed. ◻

We now flesh out steps 2 and 3 of the proof overview of the previous section to establish the second direction of the theorem—namely that if the solution to the linear program is negative, we can construct a pair of sequences $(x)_j, (y)_j$ that violates the inequality. We accomplish this in two steps. The first step is to show that for any feasible point, $(\ell)_i$, of the linear program, one can construct a function $\ell(a, b) : \mathbb{R}^2 \to \mathbb{R}$ defined on the entire plane with the property that the function is convex and has the secants through-or-above the origin property, and satisfies $\ell(a_i, b_i) = \ell_i$, where ℓ_i is the assignment of the linear program variable corresponding to a_i, b_i.

Lemma 7. *For any feasible point $(\ell)_i$ of the linear program, we can construct a linearized norm graph $\ell(a, b) : \mathbb{R}^2 \to \mathbb{R}$, which will be the maximum of r affine functions $z_i(a, b) = \alpha_i a + \beta_i b + \gamma_i$ with $\gamma_i \geq 0$, such that the function is convex, and for any $i \in [r]$, $\ell(a_i, b_i) = \ell_i$.*

Proof. We explicitly construct ℓ as the maximum of r linear functions. Recall that for each index k we constrained (a_k, b_k, ℓ_k) to lie on the lower convex hull of all the points (a_i, b_i, ℓ_i) and the special point $(2a_k, 2b_k, 2\ell_k)$. Thus through each point (a_k, b_k, ℓ_k) construct a plane that passes through or below all these other points; define $\ell(a, b)$ to be the maximum of these r functions. For each $k \in [r]$ we have $\ell(a_k, b_k) = \ell_k$ since the kth plane passes through this value, and every other plane passes through or below this value. The maximum of these planes is clearly a convex function. Finally, we note that each plane passes through-or-above the origin since a plane that passes through (a_k, b_k, ℓ_k) and through-or-below $(2a_k, 2b_k, 2\ell_k)$ must pass through or above the origin; hence for all $i \in [r]$, $\gamma_i \geq 0$. ◻

The second step of the proof consists of showing that we can use the function $\ell(a, b)$ of the above lemma to construct sequences $(x)_j, (y)_j$ that instantiate solutions of the linear program arbitrarily well, up to a scaling factor:

Lemma 8. *For a feasible point of the linear program, expressed as an r-tuple of values $(\ell)_i$, and any $\epsilon > 0$, for sufficiently small $\delta > 0$ there exist finite sequences $(x)_j, (y)_j$ such that for all $i \in [r]$,*

$$|\ell_i - \delta \log \sum_j x_j^{a_i} y_j^{b_i}| < \epsilon.$$

Proof. Consider the linearized norm graph $\ell(a, b)$ of Lemma 7 that extends $\ell(a_i, b_i)$ to the whole plane, constructed as the maximum of r planes $z_i(a, b) = \alpha_i a + \beta_i b + \gamma_i$, with $\gamma_i \geq 0$.

Consider, for parameter t_i to be defined shortly, the sequences $(x)_j, (y)_j$ consisting of t_i copies respectively of $e^{\alpha_i/\delta}$ and $e^{\beta_i/\delta}$. Hence, for all a, b we have that

$$\delta \log \sum_j x_j^a y_j^b = \alpha_i a + \beta_i b + \delta \log t_i.$$

Since $\gamma_i \geq 0$, if we let $t_i = \text{round}(e^{\gamma_i/\delta})$ we can approximate γ_i arbitrarily well for small enough δ. Finally, we concatenate this construction for all i. Namely, let $(x)_j, (y)_j$ consist of the concatenation, for all i, of $t_i = \text{round}(e^{\gamma_i/\delta})$ copies respectively of $e^{\alpha_i/\delta}$ and $e^{\beta_i/\delta}$. The values of $\sum_j x_j^a y_j^b$ will be the sum of the values of these r components, thus at least the maximum of these r components, and at most r times the maximum. Thus the values of $\delta \log \sum_j x_j^a y_j^b$ will be within $\delta \log r$ of δ times the logarithm of the max of these components. Since each of the r components approximates the corresponding affine function z_i arbitrarily well, for small enough δ, the function $\delta \log \sum_j x_j^a y_j^b$ is thus an ϵ-good approximation to the function ℓ, and in particular is an ϵ–good approximation to $\ell(a_i, b_i)$ when evaluated at (a_i, b_i), for each i. ∎

The following lemma completes the proof of Theorem 3:

LEMMA 9. *Given a feasible point of the linear program that has a negative objective function value, there exist finite sequences $(x)_j, (y)_j$ which falsify the inequality $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i} \geq 1$.*

Proof. Letting $v > 0$ denote the negative of the objective function value corresponding to feasible point $(\ell)_i$ of the linear program, define $\epsilon = \frac{v}{\sum_i |c_i|}$, and let δ_ϵ and sequences $(x)_j, (y)_j$ be those guaranteed by Lemma 8 to satisfy $|\ell_i - \delta_\epsilon \log \sum_j x_j^{a_i} y_j^{b_i}| < \epsilon$, for all $i \in r$. Multiplying this expression by c_i for each i, summing, and using the triangle inequality yields

$$\left| \sum_i c_i \ell_i - \delta_\epsilon \left(\sum_i c_i \log \sum_j x_j^{a_i} y_j^{b_i} \right) \right| < v,$$

and hence $\sum_i c_i \log \sum_j x_j^{a_i} y_j^{b_i} < 0$, and the lemma is obtained by exponentiating both sides. ∎

2.3. A geometric interpretation of inequality derivations. We provide a pleasing and intuitive interpretation of the problem being solved by the linear program in the proof of Theorem 3. This interpretation is most easily illustrated via an example, and we use one of the inequalities that we encounter in Section 3 in the the analysis of our instance-optimal tester.

EXAMPLE 4. *The 4th component of Lemma 10 (in Section 3.3) consists of showing the inequality*

$$(6) \qquad \left(\sum_j x_j^2 y_j^{-2/3}\right)^2 \left(\sum_j x_j^2 y_j^{-1/3}\right)^{-1} \left(\sum_j x_j\right)^{-2} \left(\sum_j y_j^{2/3}\right)^{3/2} \geq 1,$$

where in the notation of the lemma, the sequence x corresponds to Δ and the sequence y corresponds to p. In the notation of Theorem 3, 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 3 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 halfway 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!)

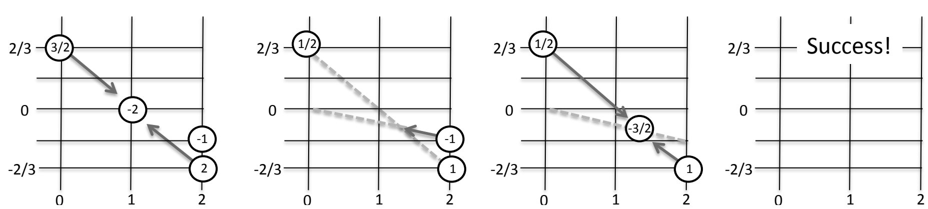

Fig. 1. *Depiction of a successful sequence of "moves" in the game corresponding to the inequality $\left(\sum_j x_j^2 y_j^{-2/3}\right)^2 \left(\sum_j x_j^2 y_j^{-1/3}\right)^{-1} \left(\sum_j x_j\right)^{-2} \left(\sum_j y_j^{2/3}\right)^{3/2} \geq 1$, showing that the inequality is true. The first diagram illustrates the initial configuration of positive and negative weights, together with the "Hölder-type move" that takes one unit of weight from each of the points at $(0, 2/3)$ and $(2, -2/3)$ and moves it to the point $(1, 0)$, canceling out the weight of -2 that was initially at $(1, 0)$. The second diagram illustrates the resulting configuration, together with the "L_p monotonicity move" that moves the -1 weight at location $(2, -1/3)$ towards the origin by a factor of $2/3$ while scaling it by a factor of $3/2$, resulting in a point at $(4/3, -2/9)$ with weight $-3/2$, which is now collinear with the remaining two points. The third diagram illustrates the final "Hölder-type move" that moves the two points with positive weight to their weighted average, zeroing out all weights.*

The answer, as illustrated in Figure 1 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 (Equation 6) as a product of powers of Hölder and L_p monotonicity inequalities, explicitly, as the product of the following three inequalities, *which are respectively 1) the square of a Cauchy-Schwarz inequality, 2) the* $3/2$ *power of an* L_p *monotonicity inequality for* $\lambda = 2/3$*, and 3) the* $3/2$ *power of a Hölder inequalty for* $\lambda = 2/3$*:*

$$\begin{aligned}
\left(\sum_j x_j^2 y_j^{-2/3}\right)\left(\sum_j x_j^0 y_j^{2/3}\right)\left(\sum_j x_j^1 y_j^0\right)^{-2} &\geq 1 \\
\left(\sum_j x_j^{4/3} y_j^{-2/9}\right)^{3/2}\left(\sum_j x_j^2 y_j^{-1/3}\right)^{-1} &\geq 1 \\
\left(\sum_j x_j^2 y_j^{-2/3}\right)\left(\sum_j x_j^0 y_j^{2/3}\right)^{1/2}\left(\sum_j x_j^{4/3} y_j^{-2/9}\right)^{-3/2} &\geq 1
\end{aligned}$$

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 have shown in Theorem 3, 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.

3. 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 the distribution. We begin by providing the definitions and terminology that will be used throughout the remainder of the paper. In Section 3.2 we describe our very simple tester, and give some intuitions and motivations behind its form.

3.1. 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. Throughout, we assume that all samples are drawn independently from the distribution in question.

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 (this goes back to at least Kolmogorov's 1933 paper [13]; see Chapter 5.4 of [14]). 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, $\|q\|_{2/3}^{2/3} = \sum_i q_i^{2/3}$.

3.2. 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:

$$\sum_i \frac{(X_i - kp_i)^2 - kp_i}{p_i}. \tag{7}$$

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 (which has expectation kp_i when drawing samples from p and thus does not change the statistic in expectation), and changing the scaling factor from $1/p_i$ to $1/p_i^{2/3}$:

$$\sum_i \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}}. \tag{8}$$

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 where 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 7) 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, burying a drastic change in the distribution (that could be detected with $O(1)$ samples). Thus we are motivated to consider a smoother scaling factor. There does not seem to be a simple 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 2. 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 $p_i^{2/3}$—resulting from balancing the upper and lower bound desiderata—was unexpected.

The following example illustrates a second benefit of our statistic of Equation 8 over the chi-squared statistic, resulting from changing kp_i to X_i:

Example 6. *Let p be the distribution with $p_1 = 1 - 1/n$, and where 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 $((X_i - kp_i)^2 - X_i)p_i^{-2/3} \approx (X_i^2 - X_i)p_i^{-2/3}$, 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 and prove Theorem 2. The tester essentially just computes the statistic of Equation 8, 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.

> 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 non-increasing order of probability. Define $s = \min\{i : \sum_{j>i} p_j \le \epsilon/8\}$, and let $M = \{2, \ldots, s\}$, and $S = \{s+1, s+2, \ldots\}$. (Note that $p_M = p_{-\epsilon/8}^{-\max}$.)
>
> 1. If $\sum_{i \in M} \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}} > 4k \|p_M\|_{2/3}^{1/3}$, or
> 2. If $\sum_{i \in S} X_i > \frac{3}{16}\epsilon k$, then output "$\|p - q\|_1 \ge \epsilon$", else output "$p = q$".

For convenience, we restate Theorem 2, characterizing the performance of the above tester.

Theorem 2. *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 $\|p - q\|_1 \ge \epsilon$ 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 $c_2 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p_{-2\epsilon}^{-\max}\|_{2/3}}{\epsilon^2}\right\}$ samples.*

Before proving the theorem, we provide some intuition behind the form of the sample complexity, $\max\left\{\frac{1}{\epsilon}, \frac{\|p_{-c\epsilon}^{-\max}\|_{2/3}}{\epsilon^2}\right\}$. 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 $\|p_{-c\epsilon}^{-\max}\|_{2/3}/\epsilon^2$ if p is of the very special form where removing its max element and its smallest $c\epsilon$ 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 $\|p_{-c\epsilon}^{-\max}\|_{2/3}$ 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 $p_{-c\epsilon}^{-\max}$ 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.

Remark on "tolerant testing". We note that the "yes" case of the theorem, where $q = p$, can always 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 [20, 21].

3.3. Analysis of the tester. The core of the proof of the algorithmic direction of Theorem 2 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 its standard deviation, whereas if the samples were drawn from $q = p$, then the expectation is 0 and the standard deviation is sufficiently small so that the distribution of the statistic will not overlap significantly with the previous case (where $\|p-q\|_1 \geq \epsilon$). 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 2 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. For the sake of presenting a self-contained complete proof of Theorem 2, we write out these derivations explicitly below.

We now begin the analysis of the performance of the above tester, establishing the upper bounds of Theorem 2. When $\|p-q\|_1 \geq \epsilon$, we note that at most half of the discrepancy is accounted for by the most frequently occurring domain element of p, since the total probability masses of p and q must be equal (to 1), and thus $\geq \epsilon/2$ discrepancy must occur on the remaining elements. We split the analysis into two cases: when a significant portion of the remaining $\epsilon/2$ discrepancy falls above s then we show that case 1 of the algorithm will recognize it; otherwise, if $\|p_{<s}-q_{<s}\|_1 \geq (3/8)\epsilon$, then case 2 of the algorithm will recognize it.

We first analyze the mean and variance of the left hand side of the first condition of the tester, under the assumption (as discussed in Section 3.1) that a Poisson-distributed number of samples, $Poi(k)$ is used. This makes the number of times each domain element is seen, X_i, be distributed as $Poi(kq_i)$, and makes all X_i independent of each other. It is thus easy to calculate the mean and variance of each term. Explicitly, defining $\Delta_i = p_i - q_i$ we have

$$\underset{X_i \leftarrow Poi(kq_i)}{E}\left[[(X_i - kp_i)^2 - X_i]p_i^{-2/3}\right] = k^2\Delta_i^2 p_i^{-2/3}$$

and

$$\underset{X_i \leftarrow Poi(kq_i)}{Var}\left[[(X_i - kp_i)^2 - X_i]p_i^{-2/3}\right] = \left[2k^2(p_i-\Delta_i)^2 + 4k^3(p_i-\Delta_i)\Delta_i^2\right]p_i^{-4/3}$$

Note that when $p = q$, the expectation is 0, since $\Delta_i \equiv 0$. However, in the case that a significant portion of the ϵ deviation between p and q occurs in the region above s, we show that for suitable k, the variance is somewhat less than the square of the expectation, leading to a reliable test for distinguishing this case from the $p = q$ case.

The motivation for the convoluted steps in the derivations in the following lemma comes entirely from the general inequality result of Theorem 3, though as guaranteed by that theorem, the resulting inequalities can all be derived by elementary means without reference to the theorem.

As defined in the tester, considering the elements of p to be sorted in decreasing order by probability, we let s be the smallest integer so that $\sum_{i>s} \leq \epsilon/8$. For notational convenience, we define the set $M = \{2, \ldots, s\}$, so that p_M consists of those elements of p that have "medium" probabilities—not the largest element, and not the smallest elements that comprise $\leq \epsilon/8$ probability. We define M so that we may explicitly analyze the corresponding discrepancies Δ_M. (Note that the probabilities in the distribution q will typically not be sorted, and may not be similar to the corresponding probabilities in p).

The following lemma shows that the variance of case 1 of our estimator can be made arbitrarily smaller than the square of its expectation, which we will use for a Chebyshev bound proof in Proposition 11 below.

Lemma 10. *For any $c \geq 1$, if $k = c \cdot \max\{\frac{\|p_M\|_{2/3}^{1/3}}{p_s^{1/3}\cdot(\epsilon/8)}, \frac{\|p_M\|_{2/3}}{(\epsilon/8)^2}\}$ and if at least $\epsilon/8$ of the discrepancy falls in the medium region, namely $\sum_{i\in M}|\Delta_i| \geq \epsilon/8$, then*

$$\sum_{i\in M}\left[2k^2(p_i-\Delta_i)^2 + 4k^3(p_i-\Delta_i)\Delta_i^2\right]p_i^{-4/3} < \frac{16}{c}\left[\sum_{i\in M} k^2\Delta_i^2 p_i^{-2/3}\right]^2$$

Proof. Dividing both sides by k^4, the left hand side has terms proportional to $(p_i - \Delta_i)/k$ and its square. We bound such terms via the triangle inequality and the definition of k as $(p_i-\Delta_i)/k \le \left(p_i\frac{(\epsilon/8)^2}{\|p_M\|_{2/3}}+|\Delta_i|\frac{p_s^{1/3}(\epsilon/8)}{\|p_M\|_{2/3}^{1/3}}\right)/c$. Expanding, yields the left hand side divided by k^4 bounded as the sum of 5 terms:

$$\sum_{i\in M}\frac{2}{c^2}\left(p_i^{2/3}\frac{(\epsilon/8)^4}{\|p_M\|_{2/3}^2}+2|\Delta_i|p_i^{-1/3}\frac{p_s^{1/3}(\epsilon/8)^3}{\|p_M\|_{2/3}^{4/3}}+\Delta_i^2p_i^{-4/3}\frac{p_s^{2/3}(\epsilon/8)^2}{\|p_M\|_{2/3}^{2/3}}\right)$$
$$+\frac{4}{c}\left(\Delta_i^2p_i^{-1/3}\frac{(\epsilon/8)^2}{\|p_M\|_{2/3}}+|\Delta_i^3|p_i^{-4/3}\frac{p_s^{1/3}(\epsilon/8)}{\|p_M\|_{2/3}^{1/3}}\right).$$

We bound each of the five terms separately by $\left[\sum_{i\in M}\Delta_i^2p_i^{-2/3}\right]^2$, using the fact that $\frac{1}{c^2}\le\frac{1}{c}$, and sum the constants $2(1+2+1)+4(1+1)$ to yield 16 on the right hand side.

1. Cauchy-Schwarz yields $\sum_{i\in M}\Delta_i^2p_i^{-2/3} \ge \left(\sum_{i\in M}|\Delta_i|\right)^2\Big/\left(\sum_{i\in M}p_i^{2/3}\right) \ge (\frac{\epsilon}{8})^2/\|p_M\|_{2/3}^{2/3}$. Squaring this inequality and noting that, by definition, $\sum_{i\in M}p_i^{2/3}=\|p_M\|_{2/3}^{2/3}$ bounds the first term as desired.

2. We bound $\frac{\epsilon}{p_s^{1/3}}=\frac{\epsilon}{\|\Delta_M\|_1}\sum_{i\in M}|\Delta_i|p_s^{-1/3}\ge\frac{\epsilon}{\|\Delta_M\|_1}\sum_{i\in M}|\Delta_i|p_i^{-1/3}$ since $p_i\ge p_s$ for $i\in M$. Multiplying this inequality by the square of the Cauchy-Schwarz inequality of the previous case: $\left(\sum_{i\in M}\Delta_i^2p_i^{-2/3}\right)^2\ge\|\Delta_M\|_1^4/\|p_M\|_{2/3}^{4/3}$ and the bound $\|\Delta_M\|_1^3\ge(\frac{\epsilon}{8})^3$ yields the desired bound on the second term.

3. Simplifying the third term via $p_i^{-4/3}p_s^{2/3}\le p_i^{-2/3}$ lets us bound this term as the product of the Cauchy-Schwarz inequality of the first case: $\sum_{i\in M}\Delta_i^2p_i^{-2/3}\ge\|\Delta_M\|_1^2/\|p_M\|_{2/3}^{2/3}$ and the bound $\|\Delta_M\|_1^2\ge(\frac{\epsilon}{8})^2$.

4. Here and in the next case we use the basic fact that for $\beta>\alpha>0$ and a (nonnegative) vector z we have $\|z\|_\beta\le\|z\|_\alpha$ (with equality only when z has at most one nonzero entry). Thus $\sum_{i\in M}\Delta_i^2p_i^{-1/3}\le\left(\sum_{i\in M}\Delta_i^{4/3}p_i^{-2/9}\right)^{3/2}$, and this last expression is bounded via (the 3/2 power of) Hölder's inequality for $\lambda=2/3$ by $\left(\sum_{i\in M}\Delta_i^2p_i^{-2/3}\right)\left(\sum_{i\in M}p_i^{2/3}\right)^{1/2}$. Multiplying this inequality by the Cauchy-Schwarz inequality of the first case: $\|\Delta_M\|_1^2/\|p_M\|_{2/3}^{2/3}\le\sum_{i\in M}\Delta_i^2p_i^{-2/3}$ and the bound $(\frac{\epsilon}{8})^2\le\|\Delta_M\|_1^2$ yields the desired bound on the fourth term.

5. The norm inequality from the previous case also yields

$$\sum_{i\in M}\Delta_i^3p_i^{-4/3}\le\left(\sum_{i\in M}\Delta_i^2p_i^{-8/9}\right)^{3/2}\le p_s^{-1/3}\left(\sum_{i\in M}\Delta_i^2p_i^{-2/3}\right)^{3/2}.$$

Multiplying by the square root of the Cauchy-Schwarz bound of the first case,

$$\|\Delta_M\|_1/\|p_M\|_{2/3}^{1/3}\le\left(\sum_{i\in M}\Delta_i^2p_i^{-2/3}\right)^{1/2}$$

and the bound $\frac{\epsilon}{8}\le\|\Delta_M\|_1$ yields the desired bound on the fifth term.

We now prove the upper bound portion of Theorem 2.

PROPOSITION 11. *There exists a constant c_1 such that for any $\epsilon > 0$ and any known distribution p, for any unknown distribution q on the same domain, our tester will distinguish $q = p$ from $\|p - q\|_1 \geq \epsilon$ with probability $2/3$ using a set of $k = c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\right\}$ samples.*

Proof. We first show that if $p = q$ then the tester will recognize this fact with high probability.

Consider the first test of the algorithm, whether

$$\sum_{i \in M} \left[(X_i - kp_i)^2 - X_i\right] p_i^{-2/3} > 4k\|p_M\|_{2/3}^{1/3}.$$

As calculated above, the expectation of the left hand side is 0 in this case, and the variance is $2k^2\|p_M\|_{2/3}^{2/3}$. Thus Chebyshev's inequality yields that this random variable will be greater than $2\sqrt{2}$ standard deviations from its mean with probability at most $1/8$, and thus the first test will be accurate with probability at least $7/8$ in this case.

For the second test, whether $\sum_{i \in S} X_i > \frac{3}{16}\epsilon k$, recall that S was defined to contain those elements of p with probabilities smaller than the "medium" elements M, and, explicitly, have total probability mass $\|p_S\| \leq \epsilon/8$. Denote this total mass by m. Thus $\sum_{i \in S} X_i$ is distributed as $Poi(mk)$, which has mean and variance both $mk \leq \frac{\epsilon k}{8}$. Thus Chebyshev's inequality yields that the probability that this quantity exceeds $\frac{3}{16}\epsilon k$ is at most $\left(\frac{\sqrt{mk}}{(3/16)\epsilon k - mk}\right)^2 \leq \left(\frac{\sqrt{\epsilon k}}{\sqrt{8}(1/16)\epsilon k}\right)^2 = \frac{2^5}{\epsilon k}$. Hence provided $k \geq \frac{2^8}{\epsilon}$, this probability will be at most $1/8$. For the sake of what follows, we actually make k at least twice as large as this, setting $c_1 \geq 2^9$ so that, from the definition of k, we have $k = c_1 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\right\} \geq \frac{2^9}{\epsilon}$.

We now consider the case when $\|p - q\|_1 \geq \epsilon$, and show that the tester is also correct in this setting. Consider the element with largest probability under distribution p, and note that at most half of the discrepancy $\|p - q\|_1$ can be due to the difference in probabilities assigned to this one element, since the total probability masses of p and q are equal (to 1). Thus at least half the discrepancy between p and q occurs on the remaining elements, which consist of the elements in $S \cup M$. Hence $\|(p - q)_{S \cup M}\|_1 \geq \epsilon/2$. We consider two cases. If $\|(p - q)_S\|_1 \geq \frac{3}{8}\epsilon$, namely if most of the at least $\epsilon/2$ discrepancy occurs on the small elements, then since $\|p_S\|_1 \leq \frac{1}{8}\epsilon$ by assumption, the triangle inequality yields that $\|q_S\|_1 \geq \frac{1}{4}\epsilon$. Consider the second test in this case. Analogously to the argument above, Chebyshev's inequality shows that this test will pass except with probability at most $\frac{64}{\epsilon k}$. Hence since $k \geq \frac{2^9}{\epsilon}$ from the previous paragraph, we have that the algorithm will be successful in this case with probability at least $7/8$.

In the remaining case, $\|(p - q)_M\|_1 \geq \frac{1}{8}\epsilon$, we apply Lemma 10. We first show that the number of samples $k = c_1 \frac{\|p^{-\max}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}$ is at least as many as needed for the lemma, $c \cdot \max\left\{\frac{\|p_M\|_{2/3}^{1/3}}{p_s^{1/3}(\epsilon/8)}, \frac{\|p_M\|_{2/3}}{(\epsilon/8)^2}\right\}$, provided $c_1 \geq 128c$. The second component of this maximum is trivially less than or equal to k, since by definition $\|p_M\|_{2/3} = \|p^{-\max}_{-\epsilon/8}\|_{2/3} \leq \|p^{-\max}_{-\epsilon/16}\|_{2/3}$. To bound the first component, we let r (analogously to s) be defined as the smallest integer such that $\sum_{i > r} p_i \leq \epsilon/16$, recalling that the probabilities p_i are sorted in decreasing order. Since $\sum_{i \geq s} p_i = \sum_{i \in S \cup \{s\}} p_i \geq \epsilon/8$, the difference of these expressions yields $\sum_{i=s}^{r} p_i \geq \epsilon/16$. Since each p_i in this last sum is at most p_s, we have that $p_i^{-1/3} \geq p_s^{-1/3}$ for such i, which yields $\sum_{i=s}^r p_i^{2/3} \geq \frac{\epsilon}{16 p_s^{1/3}}$. Thus $\|p_{-\epsilon/16}^{-\max}\|_{2/3}^{2/3} = \sum_{i=2}^r p_i^{2/3} \geq \sum_{i=s}^r p_i^{2/3} \geq \frac{\epsilon}{16 p_s^{1/3}}$, where the second-to-last inequality assumes $s \neq 1$. Multiplying by the inequality $\|p_{-\epsilon/16}^{-\max}\|_{2/3}^{1/3} \geq \|p_{-\epsilon/8}^{-\max}\|_{2/3}^{1/3}$ yields the bound. (In the unusual case that $s = 1$, the set $M = \{2, \ldots, s\}$ is empty, and thus Lemma 10 is trivially true, requiring 0 samples, which we trivially have.)

We thus invoke Lemma 10, which shows that, for any $c \geq 1$, the expectation of the left hand side of the first test, $\sum_{i \in M} \left[(X_i - kp_i)^2 - X_i\right] p_i^{-2/3}$, is at least $\sqrt{c/16}$ times its standard deviation; further, we note that the triangle-inequality expression by which we bounded the standard deviation is minimized when $p = q$, in which case, as noted above, the standard deviation is $\sqrt{2}k\|p_M\|_{2/3}^{1/3}$. Thus the expression on the right hand side of the first test, $4k\|p_M\|_{2/3}^{1/3}$, is always at least $\sqrt{c/16} - 2\sqrt{2}$ standard deviations away from the mean of the left hand side. Thus for $c \geq 512$, Chebyshev's inequality yields that the first test will correctly report that p and q are different with probability at least 7/8.

Thus by the union bound, in either case $p = q$ or $\|p - q\|_1 \geq \epsilon$, the tester will correctly report it with probability at least $\frac{3}{4}$. ∎

4. 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 2. 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 [20], 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 univariate distributions that are 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 $\sqrt{\sum_i d_i^2}$ instead of $\sum_i d_i$.

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. See [3] for a more in-depth discussion of Hellinger distance and its applications to hypothesis testing lower bounds.

We first prove a technical but ultimately straightforward 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 13, which has already found use in [8]. We then assemble the pieces using some inequalities, to show the lowerbound portion of Theorem 2.

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 $\epsilon \le \lambda$.

LEMMA 12. *$H(Poi(\lambda), Poi(\lambda \pm \epsilon)) \le c \cdot \frac{\epsilon^2}{\lambda}$ for constant c.*

Proof. Assume throughout this proof that $\epsilon \le \frac{1}{2}\sqrt{\lambda}$, for otherwise the lemma is trivially true.

We bound

$$H(Poi(\lambda), Poi(\lambda \pm \epsilon))^2 = \frac{1}{2}\sum_{i \ge 0}\left(\sqrt{\frac{e^{-\lambda}\lambda^i}{i!}} - \sqrt{\frac{1}{2}\left[\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} + \frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right]}\right)^2$$

term-by-term via the inequality $|\sqrt{a}-\sqrt{b}| \le \frac{|a-b|}{\sqrt{b}}$. We let $a = \frac{e^{-\lambda}\lambda^i}{i!}$ and $b = \frac{1}{2}\left[\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} + \frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right]$ for some specific i, and sum over i later. We bound the numerator of $\frac{|a-b|}{\sqrt{b}}$ by noting that

$$|a-b| = \left|\frac{e^{-\lambda}\lambda^i}{i!} - \frac{1}{2}\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} - \frac{1}{2}\frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right|$$

is bounded by $\frac{1}{2}\epsilon^2$ times the maximum magnitude of the second derivative with respect to x of $poi(x,i)$ for $x \in [\lambda-\epsilon, \lambda+\epsilon]$. Explicitly, $\frac{d^2}{dx^2}\frac{e^{-x}x^i}{i!} = poi(x,i)\frac{(i-x)^2-i}{x^2}$.

For the denominator of $\frac{|a-b|}{\sqrt{b}}$ we will first bound it in the case when $\lambda \ge 1$, in which case since $\epsilon \le \frac{1}{2}\sqrt{\lambda}$, there is an absolute constant c such that for any $x \in [\lambda-\epsilon, \lambda+\epsilon]$ we have $poi(x,i) \le c \cdot b = \frac{1}{2}c\big[Poi(\lambda-\epsilon)+Poi(\lambda+\epsilon)\big]$. Let x^* be the value of x in the interval $[\lambda-\epsilon, \lambda+\epsilon]$ where $poi(x,i)$ is maximized.Thus the denominator $\sqrt{b}$ is at least $\sqrt{\frac{1}{c}poi(x^*,i)}$.

We combine the bounds of the previous two paragraphs to conclude the case $\lambda \ge 1$. Thus we have $\frac{|a-b|}{\sqrt{b}} \le \frac{\sqrt{c}}{2}\epsilon^2\sqrt{poi(x^*,i)}\max_{x\in[\lambda-\epsilon,\lambda+\epsilon]}\left|\frac{(i-x)^2-i}{x^2}\right|$. Since $\lambda - \epsilon \ge \frac{1}{2}$ in our case, this last expression is thus bounded as $c_2\epsilon^2\sqrt{poi(x^*,i)}\frac{(i-\lambda)^2+i}{\lambda^2}$ for some constant c_2. We thus sum the square of this expression, over all $i \ge 0$, to obtain our bound on the (square of the) Hellinger distance. Since $poi(x^*,i)$ dies off exponentially outside an interval of width $O(\sqrt{\lambda})$, we may bound the sum over all i as just a constant times the sum over an interval of width $\sqrt{\lambda}$ centered at x^*. We note that $poi(x^*,i)$ is bounded by a constant multiple of $\frac{1}{\sqrt{\lambda}}$; since we are considering i within $\frac{1}{2}\sqrt{\lambda}$ of x^*, which is within $\frac{1}{2}\sqrt{\lambda}$ of λ by definition, we have that i is bounded by a constant times λ, as is $(i-\lambda)^2$. Thus, in total for the square of the Hellinger distance, we have $\sqrt{\lambda}$ terms that are each bounded as $\left(c_2\epsilon^2\sqrt{poi(x^*,i)}\frac{(i-\lambda)^2+i}{\lambda^2}\right)^2 \le c_3\epsilon^4\frac{1}{\sqrt{\lambda}}\frac{\lambda^2}{\lambda^4} = c_3\frac{\epsilon^4}{\lambda^2\sqrt{\lambda}}$ for some constant c_3. Multiplying by the number of terms, $\sqrt{\lambda}$, yields the desired bound.

For the case $\lambda < 1$, we note that the second derivative of $poi(x,i)$ is globally bounded by a constant, bounding the numerator of $\frac{|a-b|}{\sqrt{b}}$ by $O(\epsilon^2)$. To bound the denominator, we note that, for $\lambda < 1$, the value $b = \frac{1}{2}\left[\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} + \frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right]$ is

$\Omega(1)$ for $i=0$, it is $\Omega(\lambda)$ for $i=1$, and it is $\Omega(\lambda^2)$ for $i=2$, thus yielding a bound of $O(\frac{\epsilon^4}{\lambda^2})$ on each of the first three terms in the expression for H^2. For $i \geq 3$ we have, for $x \in (0, 2\lambda]$ that $\frac{d^2}{dx^2}poi(x,i) = poi(x,i)\frac{(i-x)^2-i}{x^2} = O(\frac{\lambda^{i-2}i^2}{i!})$. Thus the numerator of $\frac{|a-b|}{\sqrt{b}}$ is bounded by ϵ^2 times this. To bound the denominator, we have that $b \geq \frac{1}{2}poi(\lambda+\epsilon, i) = \Omega(\frac{\lambda^i}{i!})$, leading to a combined bound of $\frac{|a-b|}{\sqrt{b}} = O(\epsilon^2\lambda^{i/2-2}\frac{i^2}{\sqrt{i!}})$, which is bounded as $O(\frac{\epsilon^2}{\lambda}\frac{i^2}{\sqrt{i!}})$ since $i \geq 3$ and $\lambda < 1$. Summing up the square of this over all $i \geq 3$ clearly yields $O(\frac{\epsilon^4}{\lambda^2})$, the desired bound.

Thus in all cases the square of the Hellinger distance is $O(\frac{\epsilon^4}{\lambda^2})$, yielding the lemma.

This lemma is a crucial ingredient in the proof of the following general lower bound.

THEOREM 13. *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 2 follows from the above theorem by appropriately choosing the sequence ϵ_i.

Proof of Theorem 13. For the first part of the theorem, we first analyze the trivial case where $\sum_i \epsilon_i^2 \geq \frac{1}{64}$. The inequality $\sum_i p_i^2 \leq 1$ (L_p monotonicity) and Cauchy-Schwarz yield that $\sum_i \frac{\epsilon_i^4}{p_i^2} \geq \sum_i p_i^2 \sum_i \frac{\epsilon_i^4}{p_i^2} \geq \left(\sum_i \epsilon_i^2\right)^2 \geq \frac{1}{64^2}$, which means the number of samples requested by the theorem can be made 1 by setting $c \leq \frac{1}{64}$; and clearly at least 1 sample is needed to distinguish different distributions, yielding the theorem in this case.

Otherwise, we assume $\sum_i \epsilon_i^2 < \frac{1}{64}$. Consider the following distributions, which emulate the number of times each domain element is seen in Q_ϵ and p if we take $Poi(2k)$ samples: first randomly generate $\bar{q}_i = p_i \pm \epsilon_i$ without normalizing, and then for each i draw a sample from $Poi(\bar{q}_i \cdot 2k)$; compare this to, for each i, drawing a sample from $Poi(p_i \cdot 2k)$. Since $\sum_i \bar{q}_i$ has mean 1 and variance $\sum_i \epsilon_i^2 < \frac{1}{64}$, by Chebyshev's inequality, we have $\sum_i \bar{q}_i \geq \frac{1}{2}$ with probability at least $\frac{15}{16}$. Provided $\sum_i \bar{q}_i \geq \frac{1}{2}$, then the expected number of samples drawn (when, as described above, for each i we draw a sample from from $Poi(\bar{q}_i \cdot 2k)$) is at least k, and thus with probability at least $\frac{1}{2}$, at least k samples will be drawn. Thus via this Poisson process, with probability $\frac{1}{2}$, we have emulated drawing a sample of size k from a distribution that corresponds to Q_ϵ at least $\frac{15}{16}$ of the time.

Correspondingly, we emulate p by the simple Poisson process of drawing $Poi(2k)$ samples from p, and throwing out all but k samples; there will be at least k samples with probability greater than $\frac{1}{2}$.

Assume for the sake of contradiction that there is a hypothetical tester that could distinguish p from Q_ϵ in k samples with probability $2/3$, then this tester could be used to distinguish the following two processes with probability $\frac{1/2+2/3}{2} = \frac{7}{12}$:

1. Draw $\bar{q}_i = p_i \pm \epsilon_i$
   (a) If $\sum_i \bar{q}_i < \frac{1}{2}$ then with probability $\frac{1}{2}$ output "FAIL" and with probability $\frac{1}{2}$ output "Q"

(b) Otherwise, for each i generate a sample from $Poi(\bar{q}_i \cdot 2k)$; if fewer than k total samples are generated, output "FAIL", otherwise flip a biased coin and either output a randomly chosen k of the generated samples, or "FAIL" so that the total probability of outputting "FAIL" in this case equals $\frac{1}{2}$.

2. Or, draw a sample of size $Poi(2k)$ from p, and if fewer than k total samples are generated, output "FAIL", otherwise flip a biased coin and either output a randomly chosen k of the generated samples, or "FAIL" so that the total probability of outputting "FAIL" in this case equals $\frac{1}{2}$.

The tester is simulated on the samples if the chosen process above outputs samples, yielding an opinion "P" or "Q"; if the chosen process above outputs "FAIL", then a random one of "P" or "Q" is chosen; and if the (first) process outputs "Q", then this is output overall. This tester succeeds with probability at least the average of $\frac{1}{2}$ and $\frac{2}{3}$, since the above processes outputs "FAIL" with probability $\frac{1}{2}$ yielding a random guess about "P" or "Q", and otherwise either generate a faithful sample from the corresponding distribution, or in Case 1a outputs the answer directly, and is thus at least as accurate as the $\frac{2}{3}$-accurate tester.

The same tester will perform within $\frac{1}{32}$ of the success rate above if we remove Case 1a and replace it with Case 1b, since this change affects the outcome only if $\sum_i \bar{q}_i < \frac{1}{2}$ and simultaneously "FAIL" is not chosen, which happens with probability $\frac{1}{16} \cdot \frac{1}{2} = \frac{1}{32}$, yielding an accuracy at least $\frac{7}{12} - \frac{1}{32} > \frac{1}{2}$.

We thus derive a contradiction by showing that we cannot distinguish the following two processes with constant probability bounded above 1/2: 1) for each i, draw a sample from $Poi((p_i \pm \epsilon_i) \cdot 2k)$; versus 2) for each i, draw a sample from $Poi(p_i \cdot 2k)$. 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 12 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 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_ϵ is bounded as $O(c)$, and thus for small enough constant c 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. ∎

We apply this result as follows.

COROLLARY 14. *There is a constant c' such that for all probability distributions p and each $\alpha > 0$, there is no tester that, via a set of $c' \cdot \left(\sum_{i\neq m} \frac{\min\{p_i, \alpha p_i^{2/3}\}^4}{p_i^2}\right)^{-1/2}$ samples can distinguish p from distributions with L_1 distance $\frac{1}{2}\sum_{i\neq m}\min\{p_i, \alpha p_i^{2/3}\}$ from p with probability 0.6, where m is the index of the element of p with maximum probability.*

Note that for sufficiently small α, the min is superfluous and the bound on the number of samples becomes $\frac{c'}{\alpha^2 \|p^{-\max}\|_{2/3}^{1/3}}$ and the L_1 distance bound becomes $\frac{1}{2}\alpha\|p^{-\max}\|_{2/3}^{2/3}$, which more intuitively rephrases the result in terms of basic norms, for this range of parameters.

Proof. Consider defining the vector of ϵ_i's by letting $\epsilon_i = \min\{p_i, \alpha p_i^{2/3}\}$ for $i \neq m$, and $\epsilon_m = \max_{i\neq m}\epsilon_i$; hence if the domain is sorted with $p_1 \geq p_2 \geq \ldots$, then for $i \geq 2$ we set $\epsilon_i = \min\{p_i, \alpha p_i^{2/3}\}$, and then set $\epsilon_1\epsilon_2$. Theorem 13 yields that p and Q_ϵ cannot be distinguished given a set of $\sqrt{2}c \cdot \left(\sum_{i\neq m} \frac{\min\{p_i, \alpha p_i^{2/3}\}^4}{p_i^2}\right)^{-1/2}$ samples where c is the constant from Theorem 13. Also from Theorem 13, with probability at least 1/2, the distance between p and an element of Q_ϵ is at least the min of $\sum_{i\neq m}\min\{p_i, \alpha p_i^{2/3}\}$ and $\frac{1}{2}\sum_i \min\{p_i, \alpha p_i^{2/3}\}$, which we trivially bound by $\frac{1}{2}\sum_{i\neq m}\min\{p_i, \alpha p_i^{2/3}\}$. We derive a contradiction as follows. If a tester with the parameters of this corollary existed, then repeating it a constant number of times and taking the majority output would amplify its success probability to at least 0.9; such a tester could be used to violate Theorem 13 via the procedure: given a set of samples drawn from either p or Q_ϵ, run the tester, and if it outputs "Q_ϵ" then output "Q_ϵ", and if it outputs "p" then flip a coin and with probability 0.7 output "p" and otherwise output "Q_ϵ". If the distribution is p then our tester will correctly output this with $0.9 \cdot 0.7 > 0.6$ probability. If the distribution was drawn from Q_ϵ then with probability at least 1/2 the distribution will be far enough from p for the tester to apply (as noted above, by Theorem 13) and report this with probability 0.9; otherwise the tester will report "Q_ϵ" with probability at least $1 - 0.7 = 0.3$. Thus the tester will correctly report "Q_ϵ" with probability at least $\frac{0.9+0.3}{2} = 0.6$ in all cases, the desired contradiction. ∎

We now prove the lower bound portion of Theorem 2.

PROPOSITION 15. *There exists a constant c_2 such that for any $\epsilon \in (0,1)$ and any known distribution p, no tester can distinguish for an unknown distribution q whether*

$q = p$ or $\|p-q\|_1 \geq \epsilon$ with probability $\geq 2/3$ when given a set of samples of size $c_2 \cdot \max\left\{\frac{1}{\epsilon}, \frac{\|p^{-\max}_{-2\epsilon}\|_{2/3}}{\epsilon^2}\right\}$.

Proof. We note, trivially, that the distributions of the vectors of k samples from two distributions that are ϵ far apart are themselves at most $k\epsilon$ far apart; thus for an appropriate constant c_2, at least $c_2 \cdot \frac{1}{\epsilon}$ samples are needed to distinguish such distributions, showing the first part of our max bound.

To show that the second term in the maximum is also a lower bound on the necessary sample size, we apply Corollary 14. Consider the probabilities p_i to be sorted in decreasing order, so that p_1 is the maximum probability element. Define α to be the value which satisfies $\frac{1}{2}\sum_{i\geq 2}\min\{p_i, \alpha p_i^{2/3}\} = \epsilon$, and let s be the smallest integer such that $\sum_{i>s} p_i \leq 2\epsilon$. We note that for $i \in \{2,\ldots,s\}$ the min is never p_i, or else (since p_i are sorted in descending order and the inequality $p_i \leq \alpha p_i^{2/3}$ gets stronger for smaller p_i), the sum would be at least $\sum_{i\geq s} p_i$ which is greater than 2ϵ by definition of s. Thus $\alpha\sum_{i=2}^{s} p_i^{2/3} = \sum_{i=2}^{s}\min\{p_i, \alpha p_i^{2/3}\} \leq \sum_{i\geq 2}\min\{p_i,\alpha p_i^{2/3}\} = 2\epsilon$, which yields $\alpha \leq 2\|p_{\{2,\ldots,s\}}\|_{2/3}^{-2/3}\epsilon$. The lower bound on k from Corollary 14 is thus bounded (since the min of two quantities can only increase if we replace one by a weighted geometric mean of both of them) as $c'\cdot\left(\sum_{i\geq 2}\frac{\min\{p_i,\alpha p_i^{2/3}\}^4}{p_i^2}\right)^{-1/2} = c'\cdot\left(\sum_{i\geq 2}\min\{p_i^2, \alpha^4 p_i^{2/3}\}\right)^{-1/2} \geq c'\cdot\left(\alpha^3\sum_{i\geq 2}\min\{p_i,\alpha p_i^{2/3}\}\right)^{-1/2}$. We bound this last expression by bounding α^3 by the cube of our bound $\alpha \leq 2\|p_{\{2,\ldots,s\}}\|_{2/3}^{-2/3}\epsilon$ and then plugging in the definition $\frac{1}{2}\sum_{i\geq 2}\min\{p_i,\alpha p_i^{2/3}\} = \epsilon$ to yield a lower bound on k of $c'\cdot\left(16\|p_{\{2,\ldots,s\}}\|_{2/3}^{-2}\epsilon^4\right)^{-1/2} = \frac{c'}{4}\cdot\frac{\|p_{\{2,\ldots,s\}}\|_{2/3}}{\epsilon^2}$. A constant number of repetitions lets us amplify the accuracy of the tester from the 0.6 of Corollary 14 to the 2/3 of this theorem. ∎

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*, Proc. 25th Conference on Learning Theory (COLT), 23 (2012), pp. 22.1–22.18.

[3] Z. Bar-Yossef, *The Complexity of Massive Data Set Computations*, PhD thesis, Berkeley, CA, USA, 2002. AAI3183783.

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

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

[6] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, *Testing random variables for independence and identity*, in IEEE Symposium on Foundations of Computer Science (FOCS), 2001.

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

[8] S. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, *Optimal algorithms for testing closeness of discrete distributions*, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2014, pp. 1193–1203.

[9] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya, *Towards estimation error guarantees for distinct values*, in Symposium on Principles of Database Systems (PODS), 2000.

[10] I. Diakonikolas, D. M. Kane, and V. Nikishkin, *Testing identity of structured distributions*, in Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '15, Philadelphia, PA, USA, 2015, Society for Industrial and Applied Mathematics, pp. 1841–1854, http://dl.acm.org/citation.cfm?id=2722129.2722252.

[11] O. Goldreich and D. Ron, *On testing expansion in bounded-degree graphs*, in Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

[12] S. Guha, A. McGregor, and S. Venkatasubramanian, *Streaming and sublinear approximation of entropy and information distances*, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2006.

[13] A. N. Kolmogorov, *On the empirical determination of a distribution law*, Giornale dell'Istituto Italiano degli Attuari, 4 (1933), pp. 83–91.

[14] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*, Cambridge University Press, 2005.

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

[16] L. Paninski, *Estimating entropy on m bins given fewer than m samples*, IEEE Trans. on Information Theory, 50 (2004), pp. 2200–2203.

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

[18] 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 (2009), pp. 813–842.

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

[20] 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 Proceedings of the ACM Symposium on Theory of Computing (STOC), 2011.

[21] G. Valiant and P. Valiant, *The power of linear estimators*, in IEEE Symposium on Foundations of Computer Science (FOCS), 2011.

[22] G. Valiant and P. Valiant, *Instance optimal learning of discrete distributions*, in Proceedings of the ACM Symposium on Theory of Computing (STOC), 2016.

[23] P. Valiant, *Testing symmetric properties of distributions*, in Symposium on Theory of Computing (STOC), 2008.