Electronic Colloquium on Computational Complexity, Report No. 111 (2013)

Instance-by-instance optimal identity testing

Gregory Valiant
Stanford University
gregory.valiant@gmail.com

Paul Valiant
Brown University
pvaliant@gmail.com

August 16, 2013

Abstract

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

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

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

The characterization is of a perhaps non-traditional nature in that it uses linear programming to compute a derivation that may otherwise have to be sought through trial and error, by hand. We do not believe such a characterization has appeared in the literature, and hope its computational nature will facilitate analyses like the one here.

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

ISSN 1433-8092

website more often?). How many visitors must you observe to decide this, and, algorithmically, how do you decide this? We consider this basic question of verifying the identity of a distribution, also known as the problem of "identity testing against a known distribution". This problem has been well studied, and yielded the punchline that it is 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, 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 number of samples required to verify the identity of a distribution, as a function of the distribution in question. To more cleanly present our results, we introduce the following notation.

Definition 1. *For a probability distribution p, let $p^{-\max}$ denote the vector of probabilities obtained by removing the entry corresponding to the element of largest probability.*

Definition 2. *For a vector p and $\epsilon > 0$, define $p_{-\epsilon}$ to be the vector obtained from p by iteratively removing the smallest domain elements and stopping before more than ϵ probability mass is removed.*

Alternatively, consider the vector of probabilities, $(p)_i$, assumed to be sorted in increasing order, and let s be the largest integer for which $\sum_{i<s} p_i \leq \epsilon$. Set $p_{<s}$ to be 0.

Our main result is the following:

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

In short, over the entire range of potential distributions p, our tester is optimal, up to constant factors in ϵ and the number of samples. The distinction of "constant factors in ϵ" is needed, as $||p_{-\epsilon/16}||_{2/3}$ might *not* be within a constant factor of $||p_{-\epsilon}||_{2/3}$ if, for example, the majority of the $2/3$-norm of p comes from tiny domain elements that only comprise an ϵ fraction of the 1-norm. Note that $||p^{-\max}_{-\epsilon}||_{2/3} < ||p_{-\epsilon}||_{2/3} \leq ||p_{-\epsilon/16}||_{2/3} \leq ||p||_{2/3}$, and hence $||p||_{2/3}$ is always an upper bound on the number of samples required; for less pathological p, all these quantities are within a small constant factor of each other, in which case the theorem says that $\Theta(\frac{||p||_{2/3}}{\epsilon^2})$ is the optimal number of samples.

Because our tester is constant-factor tight, however, the subscript and superscript in the sample complexity $||p^{-\max}_{-\epsilon}||_{2/3}/\epsilon^2$ both mark real phenomena, and are not just artifacts of the analysis. Explicitly, the subscript and superscript each *reduce* the final value, and mark two ways in which the problem might be "unexpectedly easy". 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, because high discrepancy on a region of tiny total probability mass must necessarily greatly increase the probability mass on this region. We can thus hope to estimate the probability mass of this region in q and thus detect such an occurrence.

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

We note that, 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 bound of $O(\sqrt{n}/\epsilon^2)$ on the number of samples required to test such distributions, tightening the previous bound of $O(\frac{\sqrt{n}\text{polylog } n}{\epsilon^4})$ from [3], and matching the tight bound on the number of samples required for testing the uniform distribution [10]. Of course there are distributions p supported on $[n]$ for which identity testing can be done with far fewer samples, for example a uniform distribution on a tiny fraction of the elements (recall that in this model, the distribution p is known by the tester), though traditional testers will fail to take advantage of this.

By contrast, our results are of a new style that we call "instance-by-instance optimal": if, for a specific distribution p, it is possible to conduct identity testing efficiently, then our tester will do so, in essentially the best possible way. Having p explicitly provided to the tester enables our approach, but it is tantalizing to ask whether this style of "instance-by-instance" optimal testers may be extended beyond this setting.

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

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

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

Such inequalities are typically proven by hand, via trial and error. One basic tools 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 1, and substituting arbitrary combinations of x and y for X and Y yields families of inequalities of the form: $\left(\sum_j x_j^{a_1} y_j^{b_1}\right)^{\lambda} \left(\sum_j x_j^{a_2} y_j^{b_2}\right)^{1-\lambda} \left(\sum_j x_j^{\lambda a_1 + (1-\lambda) a_2} y_j^{\lambda b_1 + (1-\lambda) b_2}\right)^{-1} \geq 1$, and we can multiply inequalities of this form together to get further cases of the inequality in Equation 1. 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 general class of inequalities. 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.

We show that the cases where Equation 1 holds are exactly those cases expressible as a product of inequalities of the above two forms.:

Theorem 2. *The inequality* $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i} \geq 1$ *holds for all finite sequences of positive numbers* $(x)_j, (y)_j$ *if and only if it can be expressed as a finite product of positive powers of the Hölder inequalities* $\left(\sum_j x_j^{a'} y_j^{b'}\right)^{\lambda} \left(\sum_j x_j^{a''} y_j^{b''}\right)^{1-\lambda} \geq \sum_j x_j^{\lambda a' + (1-\lambda)a''} y_j^{\lambda b' + (1-\lambda)b''}$, *and the* L_p *monotonicity inequalities* $\left(\sum_j x_j^a y_j^b\right)^{\lambda} \leq \sum_j x_j^{\lambda a} y_j^{\lambda b}$, *for* $\lambda \in [0, 1]$.

This characterization seems to be a useful and general tool, and seems absent from the literature.

1.1 Related Work

Over the past fifteen years, there has been a body of work exploring the general question of how to estimate or test properties of distributions using fewer samples than would be necessary to actually learn the distribution in question. Such properties include "symmetric" properties (properties whose value is invariant to relabeling domain elements) such as entropy, support size, and distance metrics between distributions (such as L_1 distance), with work on both the algorithmic side (e.g. [4, 2, 7, 8, 9, 1, 5]), and on establishing lower bounds [11, 15]. 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 [13, 14] resolved the worst-case sample complexities of estimating many of these symmetric properties.. Also see [12] for a recent survey.

The specific question of verifying 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 [6] first considered the problem of distinguishing whether a set of samples was drawn from the uniform distribution of support n versus from a distribution that is least ϵ far from the uniform distribution, with the tight bound of $\Theta(\frac{\sqrt{n}}{\epsilon^2})$ subsequently given by Paninski [10]. For the more general problem of verifying an arbitrary distribution, Batu et al. [3], showed that for worst-case distributions of support size n, $O(\frac{\sqrt{n}\text{polylog } n}{\epsilon^4})$ samples are sufficient.

1.2 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 distributions 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 made heavy use of the standard "Poissonization" trick. That is, rather than drawing k samples from a fixed distribution p, we first select $k' \leftarrow Poi(k)$, and then draw k' samples from p. Given such a process, the number of times each domain element occurs is independent, with the distribution of the number of occurrences of the ith domain element

distributed as $Poi(k \cdot p_i)$. This independence yielded from this Poissonization significantly simplifies 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 $(\sum_i q_i^c)^{1/c}$. The notation $||q||_c^b$ is just the bth power of $||q||_c$. For example, $||q||_{2/3}^{2/3} = \sum_i q_i^{2/3}$.

As mentioned in Definitions 1 and 2, we use $p_{-\epsilon}$ to denote the vector of probabilities $p_{\geq s} = p_s, p_{s+1}, \ldots$ defined by sorting the probabilities $p_1 \leq p_2 \leq ...$ and letting s be the maximum integer such that $\sum_{i<s} p_i \leq \epsilon$. Additionally, we use $p^{-\max}$ to denote the vector of probabilities with the maximum probability omitted. Hence the frequently used notation $p_{-\epsilon}^{-\max}$ is the vector of probabilities obtained from p by both removing the largest entry, and removing the smallest entries until the weight of the small entries removed is at most ϵ.

2 An optimal tester

Assume the domain elements of p are sorted in increasing order of probability. Let s be the largest integer such that $\sum_{i<s} p_i \leq \epsilon/8$, and for each domain element i let X_i be the number of times element i occurs in the sample. Note that $p_{\geq s}$ is by definition the same as $p_{-\epsilon/8}$ as defined above, though we prefer to explicitly work with s in what follows, and thus will not use the $p_{-\epsilon}$ notation.

Given a set of k samples drawn from q, with X_i representing the number of times the ith domain element occurs, and a parameter $\epsilon > 0$:

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

2.1 Analysis of the tester

We now analyze the performance of the above tester, establishing the upper bounds of Theorem 1. When $||p-q||_1 \geq \epsilon$, note that at most $\epsilon/2$ 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). 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 $||p_{\geq s} - q_{\geq s}|| \geq 3/8$, 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 1.2) 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

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

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 greater than the square of the expectation. Note that when $p = q$, the expectation is 0, since $\Delta_i \equiv 0$.

The motivation for the convoluted steps in the derivations in the following lemma comes entirely from the general inequality result of Theorem 2, 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, let s be the largest integer such that $\sum_{i<s} p_i \leq \epsilon/8$, where we take the elements of p to be sorted by probability.

Lemma 1. *For any $c \geq 1$, if $k = c \cdot \max\{\frac{||p_{\geq s}^{-\max}||_{2/3}^{1/3}}{p_s^{1/3} \cdot (\epsilon/8)}, \frac{||p_{\geq s}^{-\max}||_{2/3}}{(\epsilon/8)^2}\}$ and if at least $\epsilon/8$ of the discrepancy falls above s, namely $\sum_{i \geq s, i \neq \arg\max p_i} |\Delta_i| \geq \epsilon/8$, then*

$$\sum_{i \geq s, i \neq \arg\max p_i} \left[2k^2(p_i - \Delta_i)^2 + 4k^3(p_i - \Delta_i)\Delta_i^2\right] p_i^{-4/3} < \frac{16}{c}\left[\sum_{i \geq s, i \neq \arg\max p_i} k^2 \Delta_i^2 p_i^{-2/3}\right]^2$$

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 from the triangle inequality and the definition of k as $(p_i - \Delta_i)/k \leq \left(p_i \frac{(\epsilon/8)^2}{||p_{\geq s}^{-\max}||_{2/3}} + |\Delta_i| \frac{p_s^{1/3}(\epsilon/8)}{||p_{\geq s}^{-\max}||_{2/3}^{1/3}}\right)/c$. Expanding, yields the left hand side divided by k^4 bounded as the sum of 5 terms:

$$\sum_{i \geq s, i \neq \arg\max p_i} \frac{2}{c^2}\left(p_i^{2/3} \frac{(\epsilon/8)^4}{||p_{\geq s}^{-\max}||_{2/3}^2} + 2|\Delta_i| p_i^{-1/3} \frac{p_s^{1/3}(\epsilon/8)^3}{||p_{\geq s}^{-\max}||_{2/3}^{4/3}} + \Delta_i^2 p_i^{-4/3} \frac{p_s^{2/3}(\epsilon/8)^2}{||p_{\geq s}^{-\max}||_{2/3}^{2/3}}\right)$$
$$+ \frac{4}{c}\left(\Delta_i^2 p_i^{-1/3} \frac{(\epsilon/8)^2}{||p_{\geq s}^{-\max}||_{2/3}} + |\Delta_i^3| p_i^{-4/3} \frac{p_s^{1/3}(\epsilon/8)}{||p_{\geq s}^{-\max}||_{2/3}^{1/3}}\right).$$

We bound each of the five terms separately, using the fact that $\frac{1}{c^2} \leq \frac{1}{c}$, and sum the constants $2(1 + 2 + 1) + 4(1 + 1)$ to yield 16 on the right hand side.

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

2. We bound $\frac{\epsilon}{p_s^{1/3}} = \frac{\epsilon}{||\Delta_{\geq s}^{-\max}||_1} \sum_{i \geq s, i \neq \arg\max p_i} |\Delta_i| p_s^{-1/3} \geq \frac{\epsilon}{||\Delta_{\geq s}^{-\max}||_1} \sum_{i \geq s, i \neq \arg\max p_i} |\Delta_i| p_i^{-1/3}$. Multiplying this inequality by the square of the Cauchy-Schwarz inequality of the previous case: $\left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3}\right)^2 \geq ||\Delta_{\geq s}^{-\max}||_1^4 / ||p_{\geq s}^{-\max}||_{2/3}^{4/3}$ and the bound $||\Delta_{\geq s}^{-\max}||_1^3 \geq (\frac{\epsilon}{8})^3$ yields the desired bound on the second term.

3. Simplifying the third term via $p_i^{-4/3} p_s^{2/3} \leq p_i^{-2/3}$ lets us bound this term as the product of the Cauchy-Schwarz inequality of the first case: $\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3} \geq ||\Delta_{\geq s}^{-\max}||_1^2 / ||p_{\geq s}^{-\max}||_{2/3}^{2/3}$ and the bound $||\Delta_{\geq s}^{-\max}||_1^2 \geq (\frac{\epsilon}{8})^2$.

4. Here and in the next case we use the basic fact that for $\beta > \alpha > 0$ and a (nonnegative) vector z we have $||z||_\beta \leq ||z||_\alpha$ (with equality only when z has at most one nonzero entry). Thus $\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-1/3} \leq \left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^{4/3} p_i^{-2/9}\right)^{3/2}$, which Hölder's inequality bounds by $\left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3}\right) \left(\sum_{i \geq s, i \neq \arg\max p_i} p_i^{2/3}\right)^{1/2}$. Multiplying this inequality by the Cauchy-Schwarz inequality of the first case: $||\Delta_{\geq s}^{-\max}||_1^2 / ||p_{\geq s}^{-\max}||_{2/3}^{2/3} \leq \sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3}$ and the bound $(\frac{\epsilon}{8})^2 \leq ||\Delta_{\geq s}^{-\max}||_1^2$ yields the desired bound on the fourth term.

5. The norm inequality from the previous case also yields

$$\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^3 p_i^{-4/3} \leq \left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-8/9}\right)^{3/2} \leq p_s^{-1/3} \left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3}\right)^{3/2}.$$

Multiplying by the square root of the Cauchy-Schwarz bound of the first case, $||\Delta_{\geq s}^{-\max}||_1 / ||p_{\geq s}^{-\max}||_{2/3}^{1/3} \leq \left(\sum_{i \geq s, i \neq \arg\max p_i} \Delta_i^2 p_i^{-2/3}\right)^{1/2}$ and the bound $\frac{\epsilon}{8} \leq ||\Delta_{\geq s}^{-\max}||_1$ yields the desired bound on the fifth term.

□

We now prove the upper bound portion of Theorem 1.

Proposition 1. *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 $k = c_1 \frac{||p_{-\epsilon/16}^{-\max}||_{2/3}}{\epsilon^2}$ 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 $\sum_{i \geq s, i \neq \arg\max p_i} \left[(X_i - kp_i)^2 - X_i\right] p_i^{-2/3} > 4k||p_{\geq s}^{-\max}||_{2/3}^{1/3}$. As calculated above, the expectation of the left hand side is 0 in this case, and the variance is $2k^2||p_{\geq s}^{-\max}||_{2/3}^{2/3}$. 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 $\sum_{i<s} X_i > \frac{3}{16}\epsilon k$, recall that s was defined so that the total probability mass among elements $< s$ is at most $\epsilon/8$. Denote this total mass by m. Thus $\sum_{i<s} X_i$ 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 at most 1/8. Note that for a suitable c_1, since $\epsilon \leq 1$ (otherwise the testing problem is trivial), we trivially have that $k = c_1 \frac{||p_{-\epsilon/8}^{-\max}||_{2/3}}{\epsilon^2} \geq \frac{2^8}{\epsilon}$.

We now show that when $||p - q||_1 \geq \epsilon$ the tester will correctly recognize this too. Note that at most $\epsilon/2$ of this discrepancy can be explained by the discrepancy in the probability of the most probable element of p since the total probability masses of p and q are equal (to 1). There are two cases. If $||(p-q)_{<s}^{-\max}||_1 \geq \frac{3}{8}\epsilon$, namely if most of the remaining at least $\epsilon/2$ discrepancy occurs for elements $< s$, 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 for an appropriate constant c_1 the algorithm will be successful in this case with probability at least 7/8.

In the remaining case, $||(p-q)^{-\max}_{\geq s}||_1 \geq \frac{1}{8}\epsilon$, and we apply Lemma 1. 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, $c \cdot \max\{\frac{||p^{-\max}_{\geq s}||^{1/3}_{2/3}}{p_s^{1/3}(\epsilon/8)}, \frac{||p^{-\max}_{\geq s}||_{2/3}}{(\epsilon/8)^2}\}$, provided $c_1 \geq 128c$. The second component of the maximum is trivially bounded since by definition $||p^{-\max}_{\geq s}||_{2/3} = ||p^{-\max}_{-\epsilon/8}||_{2/3} \leq ||p^{-\max}_{-\epsilon/16}||_{2/3}$. To bound the first component, we let r (analogously to s) be defined as the largest integer such that $\sum_{i<r} p_i \leq \epsilon/16$. Since $\sum_{i\leq s} p_i \geq \epsilon/8$, the difference of these expressions yields $\sum_{i=r}^{s} p_i \geq \epsilon/16$. 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 $\sum_{i=r}^{s} p_i^{2/3} \geq \frac{\epsilon}{16 p_s^{1/3}}$. Thus $||p^{-\max}_{-\epsilon/16}||^{2/3}_{2/3} = \sum_{i\geq r, i\neq \arg\max p_i} p_i^{2/3} \geq \frac{\epsilon}{16 p_s^{1/3}}$. Multiplying by the inequality $||p^{-\max}_{-\epsilon/16}||^{1/3}_{2/3} \geq ||p^{-\max}_{-\epsilon/8}||^{1/3}_{2/3}$ yields the bound.

We thus invoke Lemma 1, which shows that, for any $c \geq 1$, the expectation of the left hand side of the first test, $\sum_{i\geq s, i\neq \arg\max p_i} \left[(X_i - kp_i)^2 - X_i\right] p_i^{-2/3}$, 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 $\sqrt{2}k||p^{-\max}_{\geq s}||^{1/3}_{2/3}$. Thus the expression on the right hand side of the first test, $4k||p^{-\max}_{\geq s}||^{1/3}_{2/3}$, 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 $\frac{3}{4}$. □

3 Lower bounds

Let $Poi(\lambda \pm \epsilon)$ denote the probability distribution with pdf over nonnegative integers i: $\frac{1}{2}poi(\lambda + \epsilon) + \frac{1}{2}poi(\lambda - \epsilon)$, which is only defined for $\epsilon \leq \lambda$. Recall the Hellinger distance $H(p, q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$.

Lemma 2. $H(Poi(\lambda), Poi(\lambda \pm \epsilon)) \leq O(\frac{\epsilon^2}{\lambda})$

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

We bound $H(Poi(\lambda), Poi(\lambda \pm \epsilon))^2 = \frac{1}{2}\sum_{i\geq 0}\left(\sqrt{\frac{e^{-\lambda}\lambda^i}{i!}} - \sqrt{\frac{1}{2}\left[\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} + \frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right]}\right)^2$ term-by-term via the inequality $|\sqrt{a} - \sqrt{b}| \leq \frac{|a-b|}{\sqrt{b}}$. We let $a = \frac{e^{-\lambda}\lambda^i}{i!}$ and $b = \frac{1}{2}\left[\frac{e^{-\lambda-\epsilon}(\lambda+\epsilon)^i}{i!} + \frac{e^{-\lambda+\epsilon}(\lambda-\epsilon)^i}{i!}\right]$. We will make use of the bound that there is an absolute constant c such that for any $x \in [\lambda - \epsilon, \lambda + \epsilon]$ we have $poi(x, i) \leq c \cdot b$.

We note 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, $\frac{d^2}{dx^2}\frac{e^{-x}x^i}{i!} = poi(x, i)\frac{(i-x)^2 - i}{x^2}$. Let x^* be the value of x in the interval $[\lambda - \epsilon, \lambda + \epsilon]$ where $poi(x, i)$ is maximized. Note that the denominator $\sqrt{b}$ is at least $\sqrt{\frac{1}{c}poi(x^*, i)}$. For $\lambda \geq 1$ we

thus have $\lambda - \epsilon \geq \frac{1}{2}$, and thus we may bound $\frac{|a-b|}{\sqrt{b}} \leq \frac{\sqrt{c}}{2}\epsilon^2\sqrt{poi(x^*,i)}\max_{x\in[\lambda-\epsilon,\lambda+\epsilon]}\left|\frac{(i-x)^2-i}{x^2}\right| = O(\epsilon^2\sqrt{poi(x^*,i)}\frac{(i-\lambda)^2+i}{\lambda^2})$. Summing the square of this, over all $i \geq 0$, where as defined above, x^* is the value of x in the interval $[\lambda-\epsilon,\lambda+\epsilon]$ where $poi(x,i)$ is maximized, and $\epsilon \leq \frac{1}{2}\sqrt{\lambda}$ yields $O(\frac{\epsilon^4}{\lambda^2})$ because $poi(\lambda,i)$ dies off rapidly outside an interval of width $O(\sqrt{\lambda})$, as attested by the moments of the Poisson distribution.

For the case $\lambda < 1$, note that the second derivative of $poi(x,i)$ is globally bounded by a constant, and thus for $i \in \{0,1,2\}$ we this bound and the bound $b = \Omega(\lambda^2)$ to conclude that the first 3 terms in the expression for H^2 are bounded as $O(\frac{\epsilon^4}{\lambda^2})$. For $i \geq 3$ we have, for $x \in (0, \lambda + \frac{1}{2}\sqrt{\lambda})$ that $\frac{d^2}{dx^2}poi(x,i) = poi(x,i)\frac{(i-x)^2-i}{x^2} = O(\frac{e^{-x}x^{i-2}i^2}{i!}) = O((\lambda+\epsilon)^{i-2}\frac{i^2}{i!})$. Since $b \geq \frac{1}{2}poi(\lambda+\epsilon,i)$, we have that the bound on the square root of the ith term is $O(\epsilon^2(\lambda+\epsilon)^{i/2-2}\frac{i^2}{\sqrt{i!}})$. The sum of the squares of these terms clearly is $o(\frac{\epsilon^4}{\lambda^2})$, since $\lambda < 1$ and $\epsilon \leq \frac{1}{2}\sqrt{\lambda}$.

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

□

This lemma yields the following general lower bound.

Theorem 3. *Given a distribution p, and associated values ϵ_i such that $\epsilon_i \in [0, p_i]$, 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. 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 $\min\{(\sum_i \epsilon_i) - \max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$.*

Proof. Consider the following related distributions, which emulate the number of times each domain element is seen if we take $Poi(k)$ samples: first randomly generate $\bar{q}_i = p_i \pm \epsilon_i$ without normalizing, and then for each i draw a sample from $Poi(\bar{q}_i \cdot k)$; compare this to, for each i, drawing a sample from $Poi(p_i \cdot k)$. We note that with probability at least $\frac{1}{2}$, we have $\sum_i \bar{q}_i \geq 1$; further, with probability at least $\frac{1}{2}$ a Poisson distribution with parameter at least k will yield a sample at least k. Thus with probability at least $\frac{1}{8}$ we have "a set of at least k samples" from both distributions. Thus if it were possible to distinguish p from Q_ϵ in k samples with probability $2/3$, then we could distinguish these two Poisson processes with probability $\frac{1}{2} + \frac{1}{6\cdot 8}$. However, note that 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 i, the squared Hellinger distance is $H(Poi(kp_i), Poi(k[p_i \pm \epsilon_i]))^2$ which by Lemma 2 is at most $c_1 k^2 \frac{\epsilon^4}{p_i^2}$. Summing over i and taking the square root yields a bound on the Hellinger distance of $k\left(c_1\sum_i \frac{\epsilon^4}{p_i^2}\right)^{1/2}$, which thus bounds the L_1 distance. Thus for small enough c, when k satisfies the bound of the theorem, the statistical distance between a set of k samples drawn from p versus drawn from a random distribution of Q_e is arbitrarily small, and the two cannot be distinguished.

To analyze 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. By symmetry, with probability at most $1/2$, the sum of the remaining elements will be positive; otherwise, the sum of the remaining elements cannot be smaller than $-\epsilon_1$. Thus with probability at least $1/2$, we have $|\sum_i \pm\epsilon_i| \leq \epsilon_1$, as desired.

Otherwise $\epsilon_1 < \frac{1}{2}\sum_i \epsilon_i$. 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 $s_1\epsilon_1 + s_2\epsilon_2 + \ldots + s_{j-1}\epsilon_{j-1} + \epsilon_j \geq \frac{1}{2}\sum_i \epsilon_i$, and by assumption $s_1 \geq s_j$ and $j \geq 2$, 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 we would have $\sum_i \epsilon_i \geq \epsilon_1 + \sum_{i\geq j} \epsilon_i \geq \epsilon_j + \sum_{i\geq j} \epsilon i > \sum_i \epsilon_i$, a contradiction. Thus a random choice of the remaining signs 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 1. *There is a constant c' such that for 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 rephrases the result in terms of basic norms, for this range of parameters.

Proof. We apply Theorem 3, letting $\epsilon_i = \min\{p_i, \alpha p_i^{2/3}\}$ for $i \neq m$, and $\epsilon_m = \max_{i\neq m} \epsilon_i$ to show 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. Also from Theorem 3, 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 3 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 do the same, 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 $0.7 > 0.6$ 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 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 1.

Proposition 2. *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 $c_2 \frac{||[p^{-\max} - \epsilon]||_{2/3}}{\epsilon^2}$.*

Proof. We apply Corollary 1. Letting m be the index at which p_i is maximized, consider the value of α for which $\frac{1}{2}\sum_{i \neq m} \min\{p_i, \alpha p_i^{2/3}\} = \epsilon$, and let s be the largest integer such that $\sum_{i<s} p_i \leq \epsilon$, where we assume p_i is sorted in ascending order. We note that for $i \geq s$ the min is never p_i, or else (since p_i are sorted in ascending order and the inequality $p_i \leq \alpha p_i^{2/3}$ gets stronger for smaller p_i), the sum would be at least $\sum_{i \leq s} p_i$ which is greater than ϵ by definition of s. Thus $\alpha \sum_{i=s}^{m-1} p_i^{2/3} = \sum_{i=s}^{m-1} \min\{p_i, \alpha p_i^{2/3}\} \leq \sum_{i \neq m} \min\{p_i, \alpha p_i^{2/3}\} = 2\epsilon$, which yields $\alpha \leq 2||p_{\geq s}^{-\max}||_{2/3}^{-2/3}\epsilon$. The lower bound on k from Corollary 1 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 $c' \cdot \left(\sum_{i \neq m} \frac{\min\{p_i, \alpha p_i^{2/3}\}^4}{p_i^2}\right)^{-1/2} \geq c' \cdot \left(\alpha^3 \sum_{i \neq m} \min\{p_i, \alpha p_i^{2/3}\}\right)^{-1/2} \geq c' \cdot \left(16||p_{\geq s}^{-\max}||_{2/3}^{-2}\epsilon^4\right)^{-1/2} = \frac{c'}{4} \cdot \frac{||p_{\geq s}^{-\max}||_{2/3}}{\epsilon^2}$. A constant number of repetitions lets us amplify the accuracy of the tester from the 0.6 of Corollary 1 to the 2/3 of this theorem. □

4 A class of inequalities generalizing Cauchy-Schwarz

In this section we consider a general class of inequality which we have used repeatedly in Section 2, and which we have not been able to find in the literature in suitable generality.

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

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

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

Such inequalities are typically proven by hand, via trial and error. One basic tools for this is the Cauchy-Schwarz inequality, $\left(\sum_j X_j\right)^{1/2}\left(\sum_j Y_j\right)^{1/2} \geq \sum_j \sqrt{X_j Y_j}$, or the slightly more general Hölder inequality, a weighted version of Cauchy-Schwarz, where for $\lambda \in (0,1)$ we have $\left(\sum_j X_j\right)^{\lambda}\left(\sum_j Y_j\right)^{1-\lambda} \geq \sum_j X_j^{\lambda} Y_j^{1-\lambda}$. Writing this in the form of Equation 2, and substituting arbitrary combinations of x and y for X and Y yields families of inequalities of the form: $\left(\sum_j x_j^{a_1} y_j^{b_1}\right)^{\lambda}\left(\sum_j x_j^{a_2} y_j^{b_2}\right)^{1-\lambda}\left(\sum_j x_j^{\lambda a_1 + (1-\lambda)a_2} y_j^{\lambda b_1 + (1-\lambda)b_2}\right)^{-1} \geq 1$, and we can multiply inequalities of this form together to get further cases of the inequality in Equation 2. This inequality is tight when the two sequences X and Y are proportional to each other.

A second and different basic inequality of our general form, for $\lambda \in [0,1)$, is: $\left(\sum_j X_j\right)^{\lambda} \leq \sum_j X_j^{\lambda}$, which is the fact that the L_p norm is a decreasing function of p. (Intuitively, this is a

slight generalization of the trivial fact that $x^2 + y^2 \leq (x+y)^2$, and follows from the fact that the derivative of x^λ is a decreasing function of x, for positive x). As above, products of powers of x and y may be substituted for X to yield a general class of inequalities. 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, however, is that the cases where Equation 2 holds are exactly those cases expressible as a product of inequalities of the above two forms.

Theorem 2 *The inequality* $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i}\right)^{c_i} \geq 1$ *holds for all finite sequences of positive numbers* $(x)_j, (y)_j$ *if and only if it can be expressed as a finite product of positive powers of the Hölder inequalities* $\left(\sum_j x_j^{a'} y_j^{b'}\right)^{\lambda} \left(\sum_j x_j^{a''} y_j^{b''}\right)^{1-\lambda} \geq \sum_j x_j^{\lambda a' + (1-\lambda)a''} y_j^{\lambda b' + (1-\lambda)b''}$, *and the* L_p *monotonicity inequalities* $\left(\sum_j x_j^a y_j^b\right)^{\lambda} \leq \sum_j x_j^{\lambda a} y_j^{\lambda b}$, *for* $\lambda \in [0,1]$.

Proof. One direction of the implication is trivial. We prove the other direction via two steps: letting I be the size of the index set of i, we construct a linear program on the I-tuple of values ℓ_i representing $\log \sum_j x_j^{a_i} y_j^{b_i}$, and show that if the desired inequality is true then this linear program has objective value 0; we then note that the solution to the *dual* of this linear program is an explicit (finite) combination of the Hölder and L_p monotonicity inequalities that yields a derivation of the inequality as desired.

Given sequences $(x)_j, (y)_j$, consider the function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(a,b) = \log \sum_j x_j^a y_j^b$. The Hölder inequalities explicitly represent the fact that 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 ℓ, intuitively as "any secant of ℓ that passes through the origin must past through or above the origin," explicitly, for all (a',b') and all $\lambda \in (0,1)$ we have $\lambda\ell(a',b') \leq \ell(\lambda a', \lambda b')$. Let $\mathfrak{F}$ represent this family of functions from $\mathbb{R}^2$ to $\mathbb{R}$, namely, those functions that are convex and whose secants pass through-or-above the origin.

Consider, for sequences $(a)_i, (b)_i$ the I-tuples of values $\ell_i = \ell(a_i, b_i)$ that can be extended to a member of $\mathfrak{F}$. We express this set of I-tuples via the following constraints. For each 4-tuple $i1, i2, i3, i4$, if (a_{i1}, b_{i1}) is in the (closed) triangle formed by the other three points, then add the constraint that the point $(a_{i1}, b_{i1}, \ell_{i1})$ lies on-or-below the plane formed by the other three values; also, for each 3-tuple $(a_{i1}, b_{i1}), (a_{i2}, b_{i2}), (a_{i3}, b_{i3})$ where there is a ray from (a_{i1}, b_{i1}) through the edge $(a_{i2}, b_{i2}), (a_{i3}, b_{i3})$ that passes through the origin, the corresponding ray in 3-dimensions passes through-or-above the origin (equivalently, the plane through these 3 points passes through-or-above the origin). For each sequence $(a)_i, (b)_i$ this defines a linear program of at most $\binom{I}{4} + \binom{I}{3}$ constraints.

For a feasible point of the linear program, expressed as an I-tuple of values ℓ_i, and any $\delta > 0$ we show that for sufficiently small $\epsilon > 0$ there exist finite sequences $(x)_j, (y)_j$ such that for all i we have $\epsilon \log \sum_j x_j^{a_i} y_j^{b_i}$ is within δ of ℓ_i, namely that, up to the ϵ scaling, we can instantiate solutions of the linear program arbitrarily well. Consider the lower convex hull of the points (a_i, b_i, ℓ_i); because the constraints impose convexity, each of the i points is on the lower convex hull. For each i, choose a triangle on the lower convex hull passing through (a_i, b_i, ℓ_i) and two other such points such that there is a ray from (a_i, b_i) through the other edge passing through the origin. Let the equation of the plane defined by this triangle be $z_i(a,b) = \alpha_i a + \beta_i b + \gamma_i$. By the second set of constraints, each such plane passes through-or-above the origin, and thus $\gamma_i \geq 0$. Let $C(a,b)$ be the function defined

as the max of these i planes. By the first set of constraints, C passes through each point (a_i, b_i, ℓ_i).

Consider $(x)_j, (y)_j$ consisting of t_i copies respectively of $e^{\alpha_i/\epsilon}$ and $e^{\beta_i/\epsilon}$. In this case, for all a, b we have that $\epsilon \log \sum_j x_j^a y_j^b$ equals $\alpha_i a + \beta_i b + \epsilon \log t_i$. Since $\gamma \geq 0$, we let $t_i = \text{round}(e^{\gamma_i/\epsilon})$ and can approximate γ_i arbitrarily well for small enough ϵ. Finally, we concatenate this construction for all i. Namely, let $(x)_j, (y)_j$ consist of the concatenation, for all i, of $t_i = \text{round}(e^{\gamma_i/\epsilon})$ copies respectively of $e^{\alpha_i/\epsilon}$ and $e^{\beta_i/\epsilon}$. The values of $\sum_j x_j^a y_j^b$ will be the sum of the values of these I components, thus at least the maximum of these I components, and at most I times the maximum. Thus the values of $\epsilon \log \sum_j x_j^a y_j^b$ will be within $\epsilon \log I$ of ϵ times the logarithm of the max of these components. Since each of the I components approximates z_i arbitrarily well, for small enough ϵ, the function $\epsilon \log \sum_j x_j^a y_j^b$ is thus a δ–good approximation to our target C (defined as the maximum over i of z_i) and in particular is a δ–good approximation to $\ell(a_i, b_i)$ when evaluated at (a_i, b_i), for each i.

Recall our goal for the first step of the proof of the theorem, constructing a linear program on $(\ell)_i$ that has objective value 0 if our desired inequality is true.

We have already defined the constraints to the linear program; we now define the objective: minimize $\sum_i \ell_i \cdot c_i$, where recall that c_i are the exponents around each term of the desired inequality. Note that all the constraints of our linear program are homogenous, and thus satisfied for $(\ell)_i$ uniformly 0, so 0 is certainly a feasible objective value. We show the contrapositive of our claim: if the linear program can have negative objective value, then the desired inequality is false.

Consider a solution $(\ell)_i$ to the linear program with negative objective value $-v$, and let $\delta > 0$ be such that $\delta \sum_i |c_i| < v$. Let $\epsilon > 0$ and the sequences $(x)_j, (y)_j$ be as constructed above so that $\epsilon \log \sum_j x_j^{a_i} y_j^{b_i}$ is a δ–good approximation to ℓ_i, for all i. Summing over i and weighting by the coefficients c_i yields that $\sum_i c_i \epsilon \log \sum_j x_j^{a_i} y_j^{b_i}$ is within v of $-v$, and hence is strictly less than 0. Dividing by ϵ and exponentiating yields via the triangle inequality that $\prod_i \left(\sum_j x_j^{a_i} y_j^{b_i} \right)^{c_i} < 1$, namely that our desired inequality is violated, concluding this part of the proof.

We have thus shown that if the desired inequality is true then the linear program over the I-tuple $(\ell)_i$ minimizing objective function $\sum_i \ell_i \cdot c_i$ has optimal value 0 subject to the constraints that 1) For each 4-tuple $i1, i2, i3, i4$, if (a_{i1}, b_{i1}) is in the (closed) triangle formed by the other three points, then add the constraint that its value $\ell(a_{i1}, b_{i1})$ lies on-or-below the plane formed by the other three values; and 2) For each 3-tuple $(a_{i1}, b_{i1}), (a_{i2}, b_{i2}), (a_{i3}, b_{i3})$ where there is a ray from (a_{i1}, b_{i1}) through the edge $(a_{i2}, b_{i2}), (a_{i3}, b_{i3})$ that passes through the origin, the corresponding ray in 3-dimensions passes through-or-above the origin.

By linear programming duality, there is thus a nonnegative linear combination of the constraints that sums up to the inequality $\sum_i \ell_i \cdot c_i \geq 0$. Interpreting, as above, $\ell_i = \ell(a_i, b_i) = \log \sum_j x_j^{a_i} y_j^{b_i}$, and exponentiating, we thus have the desired inequality expressed as a product of powers of the exponentials of the linear program constraints. We conclude by noting that the exponential of each linear program constraint is the product of positive powers of at most 3 of the basic Hölder and L_p inequalities. Explicitly, the first set of inequalities, that a point inside a triangle has value on-or-below the plane defined by the triangle is a consequence of the convexity of the function ℓ, and can be expressed as the sum of two instances of the basic 3-point definition of convexity, which is the Hölder inequality in our context. For the second set of inequalities, if a point (a, b, z) lies on the edge $(a_{i2}, b_{i2}, \ell_{i2}), (a_{i3}, b_{i3}, \ell_{i3})$, then by the convexity of ℓ, $z \geq \ell(a, b)$; thus if the ray from (a_{i1}, b_{i1}) through (a, b) passes through the origin, the ray in 3-dimensions from $(a_{i1}, b_{i1}, \ell_{i1})$ through $(a, b, \ell(a, b))$ passes through-or-above the origin by the L_p monotonicity inequality—assuming the

triangle does not contain the origin—and thus so does the ray through (a, b, z). If the triangle contains the origin, then two applications of convexity imply that $\ell(0, 0)$ must lie on-or-below the triangle, and one application of the L_p monotonicity inequality with $\lambda = 0$ and arbitrary starting point implies that $\ell(0, 0) \geq 0$.

Thus since there are at most $\binom{I}{4} + \binom{I}{3}$ constraints, our desired inequality can be expressed as the product of positive powers of at most three times this many Hölder and L_p inequalities, as claimed.

□

References

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

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

[3] 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.

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

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

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

[7] 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.

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

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

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

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

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

[13] 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.

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

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

ECCC ISSN 1433-8092

http://eccc.hpi-web.de