Testing Symmetric Properties of Distributions

Paul Valiant
MIT
Cambridge, MA, USA
pvaliant@mit.edu

ABSTRACT

We introduce the notion of a *Canonical Tester* for a class of properties on distributions, that is, a tester strong and general enough that "a distribution property in the class is testable if and only if the Canonical Tester tests it". We construct a Canonical Tester for the class of symmetric properties of one or two distributions, satisfying a certain weak continuity condition. Analyzing the performance of the Canonical Tester on specific properties resolves several open problems, establishing lower bounds that match known upper bounds: we show that distinguishing between entropy $< \alpha$ or $> \beta$ on distributions over $[n]$ requires $n^{\alpha/\beta - o(1)}$ samples, and distinguishing whether a pair of distributions has statistical distance $< \alpha$ or $> \beta$ requires $n^{1-o(1)}$ samples. Our techniques also resolve a conjecture about a property that our Canonical Tester does not apply to: distinguishing identical distributions from those with statistical distance $> \beta$ requires $\Omega(n^{2/3})$ samples.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Algorithms, Theory

Keywords

Distribution Testing, Property Testing, Multivariate Statistics, Continuity, Vandermonde Matrices

1. INTRODUCTION

Property testing has been extensively investigated in a variety of settings, in particular, program checking (e.g. [8, 9]), testing of algebraic properties (e.g. [9, 20]), and graph testing (e.g. [12]). This advanced state of knowledge is evidenced by the emergence of general structural theorems, most notably the characterization by Alon et al. of those graph properties testable in constant time [2], making use of the *canonical tester* of [13].

By contrast, the emerging and significant subfield of *distribution testing* is currently a collection of beautiful but specific results, without a common framework.

Distribution Testing and Symmetric Properties.

The quintessential question in distribution testing can be so expressed:

> *Given black-box access to samples from one or more distributions and a property of interest for such distributions, how many samples must one draw to become confident whether the property holds?*

Such questions have been posed for a wide variety of distribution properties, including monotonicity, independence, identity, and uniformity [1, 7, 5], as well as "decision versions" of support size, entropy, and statistical and L_2 distance[4, 6, 11, 14, 10, 16, 18, 17].

The properties of the latter group, and the uniformity property of the former one, are *symmetric*. Symmetric properties are those preserved under renaming the elements of the distribution domain, and in a sense capture the "intrinsic" aspects of a distribution. For example, entropy testing asks one to distinguish whether a distribution has entropy less than α or greater than β, and is thus independent of the names of the elements. As a second example, statistical distance testing asks whether a pair of distributions are close or far apart in the L_1 sense (half the sum of the absolute values of the differences between the probabilities of each element under the two distributions). Again, it is clear that this property does not depend on the specific naming scheme for the domain elements.

Prior Work.

Answering a distribution testing question requires two components, an upper-bound (typically in the form of an algorithm) and a lower-bound, each a functions of n, the number of elements in the distribution domain. Ideally, such upper- and lower-bounds would differ by a factor of $n^{o(1)}$, so as to yield *tight* answers. This is rarely the case in the current literature, however. We highlight three such gaps that we resolve in this paper —see Theorems 1, 2, and 3 respectively, and Section 2 for definitions. The prior state of the art is:

Closeness Testing Distinguishing two identical distribu-

STOC'08, May 17–20, 2008, Victoria, British Columbia, Canada.

tions from two distributions with statistical distance $> \frac{1}{2}$ can be done in $\widetilde{O}(n^{2/3})$ by [6] and cannot be done in $o(\sqrt{n})$ samples [6].

Distance Approximation For constants $0 < \alpha < \beta < 1$, distinguishing distribution pairs with statistical distance less than α from those with distance greater than β can be done in $\widetilde{O}(n)$ samples by [3], and cannot be done in $o(\sqrt{n})$ samples (as above).

Entropy Testing For (large enough) constants $\alpha < \beta$, distinguishing distributions with entropy less than α from those with entropy greater than β can be done in $n^{\alpha/\beta}n^{o(1)}$ samples by [4], and cannot be done in (roughly) $n^{\frac{2}{3}\alpha/\beta}$ samples [18].

1.1 Our Results

We develop a unified framework for optimally answering distribution testing questions for a large class of properties:

The Canonical Tester.

We focus our attention on the class of symmetric properties satisfying the following *continuity condition*: informally, there exists (ϵ, δ) such that changing the distribution by δ induces a change of at most ϵ in the property.[1] For such symmetric properties, we essentially prove that *there is no difference between proving an upper bound and proving a lower bound*. To formalize this notion we make use of a *Canonical Tester*.

The Canonical Tester is a specific algorithm that, on input (the description of) of a property π and $f(n)$ samples from the to-be-tested distribution, answers YES or NO —possibly incorrectly. If $f(n)$ is large enough so that the Canonical Tester accurately tests the property, then clearly the property is testable with $f(n)$ samples; if the Canonical Tester does not test the property, then the property *is not* testable with $f(n)/n^{o(1)}$ samples. Thus to determine the number of samples needed to test π, one need only "use the Canonical Tester to search for the value f".[2]

Applications.

We prove the following three informally stated results, the first and third resolving open problems from [6, 4, 18]. Our techniques can also be easily adapted to reproduce (and slightly extend) the main results of [18]; we sketch this construction at the end of Section 3.1.[3]

THEOREM 1. *Distinguishing two identical distributions from two distributions with statistical distance at least $\frac{1}{2}$ requires $\Omega(n^{2/3})$ samples.*

THEOREM 2. *For any constants $0 < \alpha < \beta < 1$, distinguishing between distribution pairs with statistical distance less than α from those with distance greater than β requires $n^{1-o(1)}$ samples.*

THEOREM 3. *For real numbers $\alpha < \beta$, distinguishing between distributions with entropy less than α from those with entropy greater than β requires $n^{\alpha/\beta - o(1)}$ samples.*

Theorems 2 and 3 result directly from the Canonical Tester; Theorem 1 is proven from one of the structural theorems we develop along the way.

1.2 Our Techniques

To prove our contributions, we rely on results from a variety of fields, including multivariate analysis and linear algebra. However, rather than directly applying these techniques, we are forced to forge two specific tools, described below, that may be of independent interest.

Wishful Thinking.

Prior lower-bounds for testing symmetric properties of distributions have relied on the following crucial observation: since the property is invariant under permutation of the sample frequencies, the tester may as well be invariant under permutation of the *observed* sample frequencies. In other words, the identities of the samples received do not matter, only how many elements appear once, twice, etc. We summarize this as "collisions describe all".

However, analyzing the distribution of different types of collisions has proven to be very difficult. One of our main technical contributions is what we call the *Wishful Thinking Theorem* (Theorem 6). Analyzing the statistics of collisions would be easy if the distributions involved were coordinate-wise independent with simple marginals. The Wishful Thinking Theorem guarantees that treating the collision statistics as such does not introduce any meaningful error, thus making collision analysis "as easy as we might wish".

Importantly, the Wishful Thinking Theorem does not require any continuity condition, and thus can be applied to analyze testing general symmetric properties. Indeed, we apply this result directly to show the bound of Theorem 1.

Low-Frequency Blindness.

Prior work on testing properties of distributions noted that the frequencies of the high-frequency elements of a distribution (typically those expected to appear at least $\log n$ times among the samples) will be well-approximated by the *observed* frequencies of these items in the drawn samples. Thus if we are interested in a continuous property of the distribution, then these approximate frequencies give meaningful information. The question, however, is what to do with the low-frequency elements, which may not even appear in the given sample, despite being in the support of the distribution. Clearly the approximation of the elements not appearing in the sample cannot be taken to be 0 — approximating a distribution with support size n based on k samples would yield a distribution with support at most k, potentially distorting the distribution beyond recognition.

[1] Technically this is *uniform continuity* and not *continuity*; however, since the space of probability distributions over $[n]$ is compact, by the Heine-Cantor theorem every continuous function here is thus also uniformly continuous.

[2] The notion of "Canonical Tester" here is very much related to that used in [13], but ours is in a sense stronger because we have exactly *one* —explicitly given— canonical tester for each property, while [13] defines a class of canonical testers and shows that at least one of them must work for each property.

[3] As a side note, it would have been nice if there were an illustrative example where we could invoke the Canonical Testing theorem to derive a better algorithm for a well-studied problem; however, previous algorithmic work has been so successful that all that remains is for us to provide matching lower bounds.

Our second technique leverages continuity to show that, no matter how we analyze them, there is no way to meaningfully extract information from low-frequency elements: we call this the *Low-Frequency Blindness Theorem* (Theorem 5). This result considerably simplifies the design of a Canonical Tester: the high-frequency elements can be easily well-approximated; the low-frequency ones can be ignored. (See Section 5 for a more thorough discussion of how we use continuity and how our techniques relate to previous work, specifically [18].)

2. DEFINITIONS

For the sake of simplicity of presentation we work primarily with properties of single distributions in this extended abstract (*e.g.* entropy). For the full presentation see [21].

For positive integers n we let $[n]$ denote the integers $\{1, \ldots, n\}$. All logarithms are base 2. We denote elements of vectors with functional notation —as $v(i)$ for the ith element of v.

DEFINITION 1. *A* distribution *on* $[n]$ *is a function* $p : [n] \to [0, 1]$ *such that* $\sum_i p(i) = 1$. *We use* $\mathcal{D}_n$ *to denote the set of all distributions on* $[n]$.

Throughout this work we use n to denote the size of the domain of a distribution.

DEFINITION 2. *A* property *of a distribution is a function* $\pi : \mathcal{D}_n \to \mathbb{R}$. *A* binary property *of a distribution is a function* $\beta : \mathcal{D}_n \to \{\text{"yes"}, \text{"no"}, \emptyset\}$.

Any property π and pair of real numbers $a < b$ induces a binary property π_a^b defined as: if $\pi(p) > b$ then $\pi_a^b(p) =$"yes"; if $\pi(p) < a$ then $\pi_a^b(p) =$"no"; otherwise $\pi_a^b(p) = \emptyset$.

DEFINITION 3. *Given a binary property* π_a^b *on distributions and a function* $k : \mathbb{Z}^+ \to \mathbb{Z}^+$, *an algorithm* T *is a "*π_a^b*-tester with sample complexity* $k(\cdot)$*" if, for any distribution* p, *algorithm* T *on input* $k(n)$ *random samples from* p *will accept with probability greater than* $\frac{2}{3}$ *if* $\pi_a^b(p) =$*"yes", and accept with probability less than* $\frac{1}{3}$ *if* $\pi_a^b(p) =$*"no". The behavior is unspecified when* $\pi_a^b(p) = \emptyset$.

The metric we use to compare vectors is the L_1 norm, $|v| \triangleq \sum_i |v(i)|$. For the special case of probability distributions we define the *statistical distance* between p^+, p^- as $\frac{1}{2}|p^+ - p^-|$. (In some references the normalization constant $\frac{1}{2}$ is omitted.) We may now define our notion of continuity:

DEFINITION 4. *A property* π *is* (ϵ, δ)*-weakly-continuous if for all distributions* p^+, p^- *satisfying* $|p^+ - p^-| \le \delta$ *we have* $|\pi(p^+) - \pi(p^-)| \le \epsilon$.

Finally, we define symmetric properties:

DEFINITION 5. *A property* π *symmetric if for all distributions* p *and all permutations* σ *we have* $\pi(p) = \pi(p \circ \sigma)$.

3. THE CANONICAL TESTER AND APPLICATIONS

To motivate the rest of the paper we introduce the Canonical Tester here. Given a binary property $\pi_a^b : \mathcal{D}_n \to \{\text{"yes"}, \text{"no"}, \emptyset\}$, k samples from $[n]$ represented as the histogram $s : [n] \to \mathbb{Z}^+$ counting the number of times each element has been sampled, and a threshold $\theta \in \mathbb{Z}^+$, then the *k-sample* $\mathcal{T}^\theta$ *tester for* π_a^b returns an answer "yes" or "no" according to the following steps.

DEFINITION 6 (CANONICAL TESTER T^θ FOR π_a^b).

(1) For each i *such that* $s(i) > \theta$ *insert the constraint* $p(i) = \frac{s(i)}{k}$, *otherwise insert the constraint* $p(i) \in [0, \frac{\theta}{k}]$.

(2) Insert the constraint $\sum_i p(i) = 1$.

(3) Let P *be the set of solutions to these constraints.*

(4) If the set $\pi_a^b(P)$ *(the image of elements of* P *under* π_a^b*) contains "yes" but not "no" then return "yes"; if* $\pi_a^b(P)$ *contains "no" but not "yes" then return "no"; otherwise answer arbitrarily.*

We note that the Canonical Tester is defined as a function not an algorithm, bypassing issues of computational complexity. The tradeoffs between computational and sample complexity are a potential locus for much fruitful work, but are beyond the scope of this paper.

As a brief illustration of the procedure of the Canonical Tester, consider the operation of the Canonical Tester with threshold $\theta = 2$ on input 10 samples drawn from the set [5]: $(1, 2, 2, 1, 1, 1, 4, 5, 5, 5)$. The histogram of these samples is the function s mapping $1 \to 4$ (since "1" occurs four times), $2 \to 2$, $3 \to 0$, $4 \to 1$, and $5 \to 3$. Since both "1" and "5" occur more than $\theta = 2$ times, Step 1 adds the equality constraints $p(1) = \frac{4}{10}$ and $p(5) = \frac{3}{10}$, and inequality constraints for the remaining elements $p(2), p(3), p(4) \in [0, \frac{2}{10}]$. The Canonical Tester then finds all probability distributions p that satisfy these constraints, and in Step 4 determines whether these constraints induce a unique value for the property π_a^b.

Our main result is that (for appropriately chosen θ) the Canonical Tester is optimal: "if the Canonical Tester cannot test it, nothing can." The specifics of this claim depend on the continuity property of π. Explicitly:

THEOREM 4 (CANONICAL TESTING THEOREM). *Given a symmetric* (ϵ, δ)*-weakly-continuous property* $\pi : \mathcal{D}_n \to \mathbb{R}$ *and two thresholds* $a < b$, *such that the Canonical Tester* T^θ *for* $\theta = \frac{600 \log n}{\delta^2}$ *on* π_a^b *fails to distinguish between* $\pi > b + \epsilon$ *and* $\pi < a - \epsilon$ *in* k *samples, then no tester can distinguish between* $\pi > b - \epsilon$ *and* $\pi < a + \epsilon$ *in* $k \cdot \frac{\delta}{1000 \cdot 2^{4\sqrt{\log n}}}$ *samples.*

Essentially, the Canonical Tester is optimal up to small additive constants in a and b, and a small ($n^{o(1)}$) factor in the number of samples k.

Discussion.

While it will take us the rest of the paper to prove the Canonical Testing theorem, we note one case where it is reasonably clear that the Canonical Tester does the "right thing". Given a distribution on $[n]$, consider an element whose expected number of occurrences in k samples is somewhat greater than θ. For large enough θ we can appeal to the Law of Large Numbers to see that the *observed* frequency of this element will be (greater than $\frac{\theta}{k}$ so that the Canonical Tester will invoke an equality constraint, and) a very good estimate of its *actual* frequency. Since π is a (weakly) continuous function, evaluating π on a good estimate of the input distribution will yield a good estimate of the property, which is exactly what the Canonical Tester does. Thus the Canonical Tester does the "right thing" with high-frequency elements, and if all the elements are high-frequency will return the correct answer with high probability.

The low-frequency case, however, does not have such a simple intuition. Suppose all the frequencies of the distribution to be tested are at most $\frac{1}{k}$. Then with high probability none of the elements will be observed with high frequency. In this case the Canonical Tester constructs the set $\hat{P}$ defined by the constraints $\forall i, p(i) \in [0, \frac{\theta}{k}]$, $\sum_{i=1}^{n} p(i) = 1$ effectively *discarding all its input data*! Thus for every "low-frequency distribution" the Canonical Tester induces the same set $\hat{P}$, from which Step 4 will generate the *same* output. How can such a tester possibly be optimal?

By necessity, it must be the case that "no tester can extract useful information from low-frequency elements". We call this result the Low-Frequency Blindness theorem, which constitutes our main lower bound. The Canonical Testing theorem shows that these lower bounds are tight, and in fact match the upper bounds induced by the operation of the Canonical Tester.

THEOREM 5 (LOW FREQUENCY BLINDNESS). *Given a symmetric property π on distributions on $[n]$ that is (ϵ, δ)-weakly-continuous and two distributions, p^+, p^- that are identical for any index occurring with probability at least $\frac{1}{k}$ in either distribution but where $\pi(p^+) > b$ and $\pi(p^-) < a$, then no tester can distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$ in $k \cdot \frac{\delta}{1000 \cdot 2^{4\sqrt{\log n}}}$ samples.*

To prove this theorem we (1) derive a general criterion for when two distributions are indistinguishable from k samples, and (2) exhibit a procedure for generating a pair of distributions $\hat{p}^+, \hat{p}^-$ that satisfy this indistinguishability condition and where $\pi(\hat{p}^+)$ is large yet $\pi(\hat{p}^-)$ is small (greater than $b-\epsilon$ and less than $a+\epsilon$ respectively). We call the indistinguishability criterion the Wishful Thinking theorem (Theorem 6), in part because the criterion involves a particularly intuitive comparison of the moments of the two distributions; the second component is the matching moments theorem (Theorem 7), which shows how we may slightly modify p^+, p^- into a pair $\hat{p}^+, \hat{p}^-$ whose moments match each other so that we may apply the Wishful Thinking theorem.

3.1 Applications

We prove Theorems 2 and 3 here, and further, outline how to reproduce the results of [18] on estimating the distribution support size. (Theorem 1 is shown at the end of Section 4.) As noted above, these results yield lower-bounds matching previously known upper bounds; thus we do not need the full power of the Canonical Testing theorem to generate optimal algorithms, but may simply apply our lower bound, the Low-Frequency Blindness theorem.

We note one thing that the reader may find very strange about the following proofs: to apply the Low Frequency Blindness theorem we construct distributions p^+, p^- that have very different values of the property π and then invoke the theorem to conclude that the property cannot be approximated; however, this does *not* mean that p^+ and p^- are themselves hard to distinguish between —in the examples below they are often in fact quite easy to distinguish.

In practice, it may be hard to come up with such indistinguishable distributions, and for this reason we set up the machinery of this paper to save the property testing community from this step: as noted above, internal to the proof of the Low Frequency Blindness theorem (specifically the Matching Moments theorem) is a procedure that constructs a pair of distributions $\hat{p}^+, \hat{p}^-$ with property values almost exactly those of p^+, p^- respectively, but which *are* indistinguishable. In this manner we can now prove property testing lower-bounds without having to worry about indistinguishability.

The Entropy Approximation Bound.

We prove a more formal statement of Theorem 3, making use of the fact (proven in the full version of this paper [21]) that entropy is $(1, \frac{1}{2\log n})$-weakly-continuous.

LEMMA 1. *For any real number $\gamma > 1$, the entropy of a distribution on $[n]$ cannot be approximated within γ factor using $O(n^\theta)$ samples for any $\theta < \frac{1}{\gamma^2}$, even restricting ourselves to distributions with entropy at least $\frac{\log n}{\gamma^2} - 2$.*

PROOF. Given a real number $\gamma > 1$, let p^- be the uniform distribution on $\frac{1}{4} n^{1/\gamma^2}$ elements, and let p^+ be the uniform distribution on all n elements. We note that p^- has entropy $\frac{\log n}{\gamma^2} - 2$ and p^+ has entropy $\log n$. Further, all of the frequencies in p^+ and p^- are less than $\frac{1}{k}$ where $k = \frac{1}{4} n^{1/\gamma^2}$. We apply the Low Frequency Blindness Theorem with $\epsilon = 1$ to conclude that, since entropy is $(1, \frac{1}{2\log n})$-weakly-continuous, distinguishing distributions with entropy at least $(\log n) - 1$ from those with entropy at most $\frac{\log n}{\gamma^2} - 1$ requires $n^{1/\gamma^2 - o(1)}$ queries, which implies the desired result. □

The Statistical Distance Bound.

PROOF OF THEOREM 2. We note that statistical distance is a symmetric property, and by the triangle inequality is (ϵ, ϵ)-weakly-continuous for any $\epsilon > 0$. We invoke the Low Frequency Blindness Theorem as follows: Let $p_1^- = p_2^-$ be the uniform distribution on $[n]$, let p_1^+ be uniform on $[\frac{n}{2}]$, and let p_2^+ be uniform on $\{\frac{n}{2}+1, \ldots, n\}$. We note that the statistical distance of p_1^- from p_2^- is 0, since they are identical, while p_1^+ and p_2^+ have distance 1. Further, each of the frequencies in these distributions is at most $\frac{2}{n}$. We apply the Low Frequency Blindness Theorem with $\epsilon = \delta = \min\{\alpha, 1-\beta\}$ and $k = \frac{n}{2}$ to yield the desired result. □

The Distribution Support Size Bound.

Distribution Support Size, as defined in [18] is the problem of estimating the support size of a distribution on $[n]$ given that no element occurs with probability in $(0, \frac{1}{n})$ —that is, if it has nonzero probability then it has probability at least $\frac{1}{n}$. We note that for any $\delta > 0$ the support size function is $(n\delta, \delta)$-weakly-continuous, and further, for any constants $a < b < 1$, uniform distributions with support size na or nb are "low frequency" for any number of samples $k = o(n)$. Thus, letting $\delta < \frac{b-a}{2}$ the Low Frequency Blindness theorem implies that that distinguishing support size $> nb$ from $< na$ requires $n^{1-o(1)}$ samples... modulo one small detail: as noted above, distribution support size is only defined on certain distributions, and one must check that our proof techniques maintain this constraint. We defer the details to the full version of this paper [21].

3.2 Further Directions

It is not immediately clear why *symmetric* and *weakly-continuous* are related to the Canonical Tester, since syntactically the tester could conceivably be applied to a much wider class of properties.[4] Indeed we suspect that this tester —or something very similar— may be shown optimal for

[4]We note that if a property is drastically discontinuous then

more general properties. However, neither the symmetry nor the continuity condition can be relaxed entirely:

- Consider the problem of determining whether a distribution has more than $\frac{2}{3}$ of its weight on its first half or its second half. Specifically, on distributions of support $[n]$ let $\pi(p) = |p(\{1, \ldots, \lfloor \frac{n}{2} \rfloor\})|$, where we want to distinguish $\pi < \frac{1}{3}$ from $\pi > \frac{2}{3}$. We note that π is continuous but not symmetric. The optimal tester for this property draws a *single* sample, answering according to whether this sample falls in the first half or second half of the distribution. Further, this tester will likely return the correct answer even when each frequency in p is in $[0, \frac{2}{n}]$. However, the Canonical Tester will *discard* all such samples unless $\frac{\theta}{k} < \frac{2}{n}$, that is, unless the number of samples is almost n. Thus there is a gap of roughly n between the performance of the Canonical Tester and that of the best tester for this property.
- The problem of Theorem 1, determining whether a pair of distributions is identical or far apart can be transformed into an approximation problem by defining $\pi(p_1, p_2)$ to be -1 if $p_1 = p_2$ and $|p_1 - p_2|$ otherwise, and asking to test $\pi_{-1/2}^{1/2}$. We note that π is clearly symmetric, but *not* continuous. It can be seen that the Canonical Tester requires $\widetilde{\Theta}(n)$ samples (this follows trivially from our Theorem 2), which is $\sim n^{1/3}$ worse than the bound of $\widetilde{O}(n^{2/3})$ provided by [6] (and proven optimal by out Theorem 1).

4. THE WISHFUL THINKING THEOREM

It is intuitively obvious that the order in which samples are drawn from a distribution can be of no use to a property tester, and we have already implicitly used this fact by noting that a property tester may be given, instead of a vector of samples, just the *histogram* of the samples —the number of times each element appears. This is an important simplification because it eliminates extraneous information from the input representation, thus making the behavior of the property tester on such inputs easier to analyze. For the class of symmetric properties, however, a further simplification is possible: instead of representing the input by its histogram, we represent it by the *histogram of its histogram*, an object that appears in the literature under the name "fingerprint" [3].

To give an explicit example, consider the sample sequence $(3, 1, 2, 2, 5, 1, 2)$; the histogram of this is the sequence $(2, 3, 1, 0, 1)$, expressing that 1 occurs two times, 2 occurs three times, 3 occurs once, etc.; the histogram of this histogram is the sequence $(2, 1, 1)$ indicating that two elements occur once (3,5), one element occurs twice (1) and one element occurs three times (2) —the zeroth entry, expressing those elements not occurring, is ignored. This is the fingerprint: a vector whose ith entry denotes the number of elements that experience i-way collisions.

To motivate this, we note that for a symmetric property —that is, a property invariant under relabelings of the elements— a distribution which takes value 1 half of the time, 2 a quarter of the time and 3 a quarter of the time has the same property as a distribution that takes value 1 a quarter of the time, 2 half of the time, and 3 a quarter of the time. It is not relevant to the tester that "1" occurs more times than "2" or vice versa; the only useful information is that (for example) one element appears twice, and two elements appear once; in short, the only useful information is the "collision statistics", which is exactly what the histogram of the histogram captures. (See for example [3, 6].)

Our goal in this section is to establish a general condition for when two low-frequency distributions are indistinguishable by k-sample symmetric property testers, which we do by establishing a general condition for when the distribution of k-sample fingerprints of two distributions are statistically close, a result that we call the Wishful Thinking theorem. To motivate the main result of this section, we present a "wishful thinking" analysis, of the relevant quantity: the statistical distance between the distributions of the k-sample fingerprints induced by two distributions p^+, p^- respectively. None of the following derivation is technically correct except for its conclusion, which we prove via a different (technically correct!) method in the rest of this section.

> Consider the contribution of the ith element of a distribution p to the ath entry of the fingerprint: 1 when i is sampled a times out of k samples, 0 otherwise. Since each sample draws i with probability $p(i)$, the probability of drawing i *at all* in k samples is roughly $k \cdot p(i)$, and we (wishfully) approximate the probability of i being drawn a times as this quantity to the ath power, $k^a \cdot p(i)^a$. Thus the binary random variable representing the contribution of i to the ath fingerprint entry has mean and mean-squared equal to (roughly) $k^a \cdot p(i)^a$, where, since p is low-frequency, this is also essentially the variance. Assuming (wishfully) that the contributions from different i are *independent*, we sum the mean and variance over all i to find that the distribution of the value of the ath fingerprint entry has mean and variance both equal to $k^a \sum_{i=1}^n p(i)^a$, a quantity recognizable as proportional to the ath *moment* of p; denote this by m_a. Thus to compare the ath fingerprint entries induced by p^+ and p^- respectively, we may (wishfully) just compare the mean and variance of the induced distributions. Intuitively, the induced distributions are close if the difference between their means is much less than the square root of the variance of either: we estimate the statistical distance as $\frac{|m_a^+ - m_a^-|}{\sqrt{m_a^+}}$. Thus to estimate the statistical distance between the entire fingerprints, we sum over a: $\sum_a \frac{|m_a^+ - m_a^-|}{\sqrt{m_a^+}}$. If this expression is much less than 1, then p^+ and p^- are not distinguishable by a symmetric tester in k samples.

In this intuitive analysis we made use of "wishful thinking" once trivially to simplify small constants, but more substantially, twice to eliminate high-dimensional dependencies of distributions: we assumed that the contributions of different elements i to the ath fingerprint entry were independent; and we assumed that the distributions of different fingerprint entries were independent. As noted above, despite how convenient these claims are, neither of them is true. (Intuitively one may think of the first independence assumption as being related to the question of whether one application of the histogram function preserves entry-independence —in general it does not— and the second independence assumption

essentially anything is a "Canonical Tester" for it, since such a property is *not testable at all*. So the tester we present is canonical for weakly-continuous and "drastically discontinuous" properties. The situation in between remains open.

as being related to issues arising from the second application of the histogram function.) To address the first kind of dependency, we appeal to the standard technique of *Poissonization* (see [4]). The second dependency issue will be analyzed by appeal to a recent multivariate analysis bound.

Definition 7. *A Poisson process with parameter $\lambda \geq 0$ is a distribution over the nonnegative integers where the probability of choosing c is defined as* $\mathrm{poi}(c; \lambda) \triangleq \frac{e^{-\lambda}\lambda^c}{c!}$. *We denote the corresponding random variable as $Poi(\lambda)$. For a vector $\vec{\lambda} \geq 0$ of length t we let $Poi(\vec{\lambda})$ denote the t-dimensional random variable whose ith component is drawn independently from the univariate $Poi(\vec{\lambda}(i))$ for each i.*

Definition 8. *A k-*Poissonized tester *T is a function that correctly classifies a property on a distribution p with probability $\frac{7}{12}$ on input samples generated in the following way:*

- *Draw $k' \leftarrow Poi(k)$.*
- *Return k' samples from p.*

We have the following standard lemma:

Lemma 2. *If there exists a k-sample tester T for a property π_a^b then there exists a k-Poissonized tester T' for π_a^b.*

The reason for applying this Poissonization transform is the following elementary fact: taking $Poi(k)$ samples from p, the number of times element i is sampled is (1) independent of the number of times any other element is sampled, and (2) distributed according to $Poi(k \cdot p(i))$. In other words, the histogram of these samples may be computed entry-by-entry: for the ith entry return a number drawn from $Poi(k \cdot p(i))$. We have resolved the first interdependence issue of the wishful-thinking argument.

To resolve the second interdependence issue, pushing the element-wise independence through the second application of the histogram function, we show how we may *approximate* the distribution of the fingerprint of $Poi(k)$ samples by an element-wise independent distribution (which will turn out to be a multivariate Poisson distribution itself). To express this formally, we note that the fingerprint of $Poi(k)$ samples from p is an example of what is sometimes called a "generalized multinomial distribution", and then invoke a result that describes when generalized multinomial distributions may be approximated by multivariate Poisson distributions.

Definition 9. *The* generalized multinomial distribution *parameterized by matrix ρ, denoted M^ρ, is defined by the following random process: for each row ρ_i of ρ, draw a column from the distribution ρ_i; return a row vector recording the total number of samples falling into each column (the histogram of the samples).*

Lemma 3. *For any distributions p with support $[n]$ and positive integer k, the distribution of fingerprints of $Poi(k)$ samples from p is the generalized multinomial distribution M^ρ where matrix ρ has n rows, columns indexed by fingerprint index a, and (i, a) entry equal to* $\mathrm{poi}(a; k \cdot p(i))$, *that is, the ath row of ρ expresses the distribution $Poi(k \cdot p(i))$.*

Proof. As noted above, the ith element of the histogram of drawing $Poi(k)$ samples from p is drawn (independently) from the distribution $Poi(k \cdot p(i))$. The generalized multinomial distribution M^ρ simply draws these samples for each i and returns the histogram, which is distributed as the histogram of the histogram of the original $Poi(k)$ samples, as desired. □

We introduce here the main result from Roos[19] which states that generalized multinomial distributions may be well-approximated by multivariate Poisson processes.

Roos's Theorem [19]. *Given a matrix ρ, letting $\vec{\lambda}(a) = \sum_i \rho(i, a)$ be the vector of column sums, we have*

$$|M^\rho - \mathrm{Poi}(\vec{\lambda})| \leq 8.8 \sum_a \frac{\sum_i \rho(i, a)^2}{\sum_i \rho(i, a)}.$$

Thus the multivariate Poisson distribution is a good approximation for the fingerprints, provided ρ satisfies a smallness condition. We provide a straightforward bound of this distance from the fact that the underlying distribution p is "low-frequency" as follows:

Lemma 4. *Given a distribution p, an integer k, and a real number $0 < \epsilon \leq \frac{1}{2}$ such that $\forall i, p(i) \leq \frac{\epsilon}{k}$, if ρ is the matrix with (i, a) entry* $\mathrm{poi}(a; k \cdot p(i))$ *then $\sum_a \frac{\sum_i \rho(i,a)^2}{\sum_i \rho(i,a)} \leq 2\epsilon$.*

Proof. We note that $\rho(i, a) = \mathrm{poi}(a; k \cdot p(i)) = \frac{e^{-k \cdot p(i)}(k \cdot p(i))^a}{a!} \leq (k \cdot p(i))^a \leq \epsilon^a$. Thus

$$\sum_a \frac{\sum_i \rho(i, a)^2}{\sum_i \rho(i, a)} \leq \sum_a \max_i \rho(i, a) \leq \sum_a \epsilon^a \leq 2\epsilon.$$

□

Via the Poissonization technique and Roos's theorem we have thus reduced the problem to that of comparing two multivariate Poisson distributions. To this end we apply the following bound, derived from the fact that the statistical distance between *univariate* Poisson distributions with parameters λ^+, λ^- is at most $2\frac{|\lambda^+ - \lambda^-|}{\sqrt{1+\max\{\lambda^+, \lambda^-\}}}$ and the fact that the statistical distance of two distributions with independent marginals is at most the sum of the respective distances between the marginals:

Lemma 5. *The statistical distance between two multivariate Poisson distributions with parameters $\vec{\lambda}^+, \vec{\lambda}^-$ is bounded as*

$$|\mathrm{Poi}(\vec{\lambda}^+) - \mathrm{Poi}(\vec{\lambda}^-)| \leq 2 \sum_a \frac{|\vec{\lambda}^+(a) - \vec{\lambda}^-(a)|}{\sqrt{1 + \max\{\vec{\lambda}^+(a), \vec{\lambda}^-(a)\}}}.$$

Combining results yields:

Lemma 6. *Given an integer $k > 0$ and two distributions p^+, p^- all of whose frequencies are at most $\frac{1}{500k}$, then, letting $\vec{\lambda}^+(a) = \sum_i$* $\mathrm{poi}(a; k \cdot p(i))$, *if it is the case that*

$$\sum_a \frac{|\vec{\lambda}^+(a) - \vec{\lambda}^-(a)|}{\sqrt{1 + \max\{\vec{\lambda}^+(a), \vec{\lambda}^-(a)\}}} < \frac{1}{25}.$$

then it is impossible to test any symmetric property that is true for p^+ and false for p^- in k samples.

Proof. Combining Lemma 4 with Roos's Theorem we have that for each of p^+ and p^- the distance of the Poisson approximation from the distribution of fingerprints of $Poi(k)$ samples is at most $\frac{2 \cdot 8.8}{500} < \frac{1}{25}$. Thus, by the triangle inequality, the distance between the distribution of fingerprints of $Poi(k)$ samples from p^+ versus p^- is at most $\frac{2}{25}$ plus the bound from Lemma 5, which is also $\frac{2}{25}$, yielding a total distance of at most $\frac{4}{25}$, which is less than $\frac{1}{6}$. Assume for the sake of contradiction that there is a k-sample tester that distinguishes between p^+ and p^-. By Lemma 2 there must thus exist a tester on $Poi(k)$ samples. However, the definition of a Poissonized tester requires that the tester succeed with probability at least $\frac{7}{12}$ on p^+ and succeed with probability at most $\frac{5}{12}$ on p^-, which contradicts the fact that their input distributions have statistical distance strictly less than $\frac{1}{6}$. Thus no such tester can exist. $\square$

As it turns out, we can simplify this bound by replacing $\vec{\lambda}(a)$ here with the ath moments of the distributions, yielding the final form of the Wishful Thinking theorem. The proof involves expressing each $\vec{\lambda}_a$ as a power series in terms of the moments, and is straightforward but technical; we omit it from this abstract (see [21] for details).

Definition 10. *For integer k and distribution p, the k-*based moments *of p are the values $k^a \sum_i p(i)^a$ for $a \in \mathbb{Z}^+$.*

Theorem 6 (Wishful Thinking). *Given a positive integer k and two distributions p^+, p^- all of whose frequencies are at most $\frac{1}{500k}$, then, letting m^+, m^- be the k-based moments of p^+, p^- respectively, if it is the case that*

$$\sum_a \frac{|m^+(a) - m^-(a)|}{\sqrt{1+\max\{m^+(a), m^-(a)\}}} < \frac{1}{50}.$$

then it is impossible to test any symmetric property that is true for p^+ and false for p^- in k samples.

We will find it convenient to work with a finite subset of the moments in Section 5, so we prove as a corollary to the Wishful Thinking Theorem that if we have an even tighter bound on the frequencies of the elements, then we may essentially ignore all moments beyond the first $\sqrt{\log n}$.

Corollary 1. *Given a positive integer k, real number $\epsilon \le \frac{1}{10 \cdot 2^{\sqrt{\log n}}}$ and two distributions p^+, p^- all of whose frequencies are at most $\frac{\epsilon}{k}$, then, letting m^+, m^- be the k-based moments of p^+, p^- respectively, if it is the case that*

$$\sum_{a=2}^{\sqrt{\log n}} \frac{|m^+(a) - m^-(a)|}{\sqrt{1+\max\{m^+(a), m^-(a)\}}} < \frac{1}{120}.$$

then it is impossible to test any symmetric property that is true for p^+ and false for p^- in k samples.

Proof. We derive this from the bound of the Wishful Thinking Theorem. We note that for any distributions p^+, p^-, we have $m^+(0) = m^-(0) = n$, and $m^+(1) = m^-(1) = k$, so thus the terms for $a < 2$ vanish. To bound the terms for $a > \max\{2, \sqrt{\log n}\}$ we note that for such a we have $m^+(a) \le k^a n (\frac{\epsilon}{k})^a = n\epsilon^a \le .1^a$ Thus, since $\frac{|m^+(a)-m^-(a)|}{\sqrt{1+\max\{m^+(a),m^-(a)\}}} \le m^+(a)$, we can bound these terms by $\sum_{a \ge 2} .1^{a+b} < \frac{1}{50} - \frac{1}{120}$, yielding the corollary. $\square$

The Closeness Testing Lower Bound.

We are now in a position to prove Theorem 1, the bound on testing whether two distributions are identical or far apart. The proof is a realization of an outline that appeared in [6], but making essential use of the Wishful Thinking Theorem.

As we have deferred discussion of the two-distribution versions of our results to the full version of this paper (see [21]), we mention briefly that for a pair of distributions (p_1, p_2) we define the k-based moments to be the matrix with entries $m(a,b) = k^{a+b} \sum_i p_1(i)^a \cdot p_2(i)^b$, and note that the Wishful Thinking theorem generalizes naturally to the two-distribution case with the parameters changing from $\frac{1}{500k}$ and $\frac{1}{50}$ to $\frac{1}{1000k}$ and $\frac{1}{250}$ respectively.

Proof of Theorem 1. Let x, y be distributions on $[n]$ defined as follows: for $1 \le i \le n^{2/3}$ let $x(i) = y(i) = \frac{1}{2n^{2/3}}$. For $n/2 < i \le 3/4n$ let $x(i) = \frac{2}{n}$; and for $3n/4 < i \le n$ let $y(i) = \frac{2}{n}$. The remaining elements of x and y are zero.

Let $p_1^+ = p_2^+ = p_1^- = x$, and $p_2^- = y$ and let $k = \frac{n^{2/3}}{800}$. We note that each frequency defined is at most $\frac{1}{1600k}$. Let $m^+_{a,b}$ and $m^-_{a,b}$ be the k-based moments of (p_1^+, p_2^+) and (p_1^-, p_2^-) respectively. We note that since x and y are permutations of each other, whenever one of $a = 0$ or $b = 0$ we have $m^+_{a,b} = m^-_{a,b}$, so the corresponding terms from the Wishful Thinking Theorem vanish. For the remaining terms, $a, b \ge 1$ and we explicitly compute $m^-_{a,b} = \frac{n^{2/3}}{1600^{a+b}}$ and $m^+_{a,b} = \frac{n^{2/3}}{1600^{a+b}} + \frac{n}{4(400n^{1/3})^{a+b}}$, so thus

$$\begin{aligned}
\sum_{a,b} \frac{|m^+_{a,b} - m^-_{a,b}|}{\sqrt{1+\max\{m^+_{a,b}, m^-_{a,b}\}}} &\le \sum_{a,b} \frac{|m^+_{a,b} - m^-_{a,b}|}{\sqrt{m^-_{a,b}}} \\
\le \sum_{a,b\ge 1} \frac{\frac{n}{4(400n^{1/3})^{a+b}}}{\sqrt{\frac{n^{2/3}}{1600^{a+b}}}} &= \sum_{a,b\ge 1} \frac{n^{2/3}}{4(10n^{1/3})^{a+b}} \\
= \frac{1}{400} \sum_{a,b\ge 0} \frac{1}{(10n^{1/3})^{a+b}} &\le \frac{1}{400} \sum_{a,b} \frac{1}{10^{a+b}} < \frac{1}{250}.
\end{aligned}$$

Invoking the Wishful Thinking theorem (two-distribution version) yields the desired result. $\square$

5. THE MATCHING MOMENTS THEOREM

In the previous section we showed essentially that moments are all that matter in the low-frequency setting. In this section we consider the new ingredient of (ϵ, δ)-weak continuity and show that with this ingredient, even moments become useless for distinguishing properties; in short, no useful information can be extracted from the low-frequency portion of a distribution, a claim that will be made explicitly in the final section.

To see how the Wishful Thinking theorem relates to an (ϵ, δ)-weakly-continuous property π, we note that if π_a^b is testable, then for any distribution p^+ with large value of π and distribution p^- with small value of π, we must not only be able to distinguish samples of p^+ from samples of p^-, but further, we must be able to distinguish samples of any distribution in a ball of radius δ about p^+ from samples of any distribution in a ball of radius δ about p^-. By the Wishful Thinking theorem this means that we can test the property only if the images of these balls under the moments function lie far apart. The main result of this section is (essentially) that the images of these balls under the moments function *always overlap*.

We carry out this analysis under the constraint that we desire an intersection point that is itself a low frequency distribution (we relax the constraint to frequency at most $\frac{n^{o(1)}}{k\delta}$), so that we can conclude the argument as follows:

there exists $\hat{p}^+$ near p^+ with moments vector near some fixed $\hat{m}$ and there exists $\hat{p}^-$ near p^- with moments also near $\hat{m}$ such that both $\hat{p}^+$ and $\hat{p}^-$ have frequencies below $\frac{n^{o(1)}}{k\delta}$; thus by the Wishful Thinking theorem, large values of π are indistinguishable from small values of π in $\frac{k\delta}{n^{o(1)}}$ samples. More specifically, there is a *fixed* vector $\hat{m}$ in moments space that each of these spheres lies close to.

In other words, the plan for this section is to show how we can modify low-frequency distributions (1) slightly, (2) into almost-low-frequency distributions so that (3) their moments almost match $\hat{m}$. Recall from Section 4 that the zeroth and first moments already match (being always n and k respectively), so we need only work to match the second and higher moments. Further, the second and higher moments all depend on quadratic or higher powers of the frequencies, so the original moments of the low-frequency distribution will be swamped by the moments of the small "almost-low-frequency" modifications we make.

To give a flavor of how to find these modifications to match the second and higher moments, suppose for the moment that we ignore the constraints that the distribution p has n entries summing to 1, and consider, for arbitrary κ, c, γ, what happens to the κ-based moments if we add c new entries of value $\frac{\gamma}{\kappa}$. By trivial application of the definition, the κ-based moments of the distribution will simply increase by the vector $c\cdot(1, \gamma, \gamma^2, \ldots)$. The crucial fact here is that these moments are a linear function of c. In order to be able to fix the first $\mu = \sqrt{\log n}$ moments we need μ linear equations with μ unknowns: instead of using one value of c and γ we let γ range over $[\mu]$ and let c_γ denote the number of new entries of value $\frac{\gamma}{\kappa}$ we insert. Given the desired value for $\hat{m}$ we solve for the vector c by matrix division: if V is the transform matrix such that the new moments equal $m+V\cdot c$ then, equating this to our moments target $\hat{m}$, we solve for c as $c = inv(V)(\hat{m} - m)$.

There are a few evident concerns with this approach: (1) how do we ensure each c_γ is integral? (2) how do we ensure that each c_γ is positive? (3) how do we ensure each c_γ is small enough that the distribution is not changed much? and (4) how do we reinstate the constraints that the distribution has n entries summing to 1?

The short answers to these questions are: (1) Round to the nearest integer. (2) If we are worried about c being negative, say as low as the negation of $\bar{c} = \max_m inv(V)\cdot m$ we simply set $\hat{m} = V\cdot\bar{c}$ since we are free to choose $\hat{m}$ as we wish. Now $c = inv(V)(\hat{m} - m) = \bar{c} - inv(V)m \geq 0$ by definition of $\bar{c}$, so c is always positive. (3) To bound the size of c we note that the matrix V is in fact an example of a *Vandermonde matrix*, a class which is both well studied and well-behaved; we use standard bounds on the inverse of Vandermonde matrices. And (4) see Definition 12 for the details of the fairly straightforward construction.

(We note that [18] previously used Vandermonde matrices to control moments in a similar context. One principle distinction is that they did not have a "wishful thinking theorem" to motivate the general approach we take here; instead, they essentially seek one special case of the Matching Moments theorem, and apply it to bound the complexity of the particular problem of testing distribution support size.)

We define the particular Vandermonde matrices we use:

Definition 11. *For positive integer μ define the $\mu\times\mu$ matrix V^μ to have entries $V^\mu(i,j) = j^i$.*

As noted above, we need a bound on the size of elements of $inv(V^\mu)$. To compute this we make use of the following standard (if slightly unwieldy) formula:

Lemma 7 (From [15]). *For any vector z of length μ the inverse of the $\mu\times\mu$ Vandermonde matrix with entries $z(j)^i$ has (i,j)th entry*

$$\frac{(-1)^{i+1}\displaystyle\sum_{\substack{1\leq s_1<s_2<\ldots<s_{\mu-i}\leq\mu\\ \forall q, s_q\neq j}}\prod_{q=1}^{\mu-i} z_{s_q}}{\displaystyle\prod_{q\in\{1,\ldots,\mu\}-\{j\}}(z_q - z_j)}.$$

We apply this lemma to bound the inverse of V^μ. We omit the straightforward computation.

Lemma 8. *Each element of $inv(V^\mu)$ has magnitude at most 6^μ.*

We now present the construction for "matching moments".

Definition 12. *Define the function M mapping distribution p on $[n]$, positive integer $k\leq n$, and real number number $0<\delta\leq 1$ to distribution $\hat{p}\leftarrow M^k_\delta(p)$ via the following sequence of modifications to p:*

1. *Let $\delta' = \frac{\delta}{2}$; let I be the largest set of indices i such that $\sum_{i\in I} p(i)\leq\delta'$. Set $\hat{p}$ equal to p on $[n]-I$, and 0 on I.*
2. *Let $\mu = \lfloor\sqrt{\log n}\rfloor$, and let $\kappa = k\cdot\frac{\delta'}{4\mu^3 6^\mu}$; for integers $2\leq a\leq\mu$ let $m(a)$ be the κ-based moments of this modified vector, with $m(1) = 0$ defined separately. Let $\hat{c} = inv(V^\mu)\cdot m$.*
3. *Let $\overline{m}(a)$ be an upper-bound on m which has value 0 for $a = 1$ and value $\frac{\kappa^2}{k}$ otherwise. Let $\overline{V}^{\mu I}$ be a $\mu\times\mu$ matrix with entries 6^μ, and let $\bar{c} = \overline{V}^{\mu I}\cdot\overline{m}$.*
4. *For each $\gamma<\mu$ choose $c(\gamma) = \lfloor\bar{c}(\gamma) - \hat{c}(\gamma)\rfloor$ indices $i\in I$ with $\hat{p}(i) = 0$ and set $\hat{p}(i) = \frac{\gamma}{\kappa}$ for these indices.*
5. *Make $\sum\hat{p}(i) = 1$ by filling in $n\frac{\delta'}{2}$ of the unassigned entries from I uniformly.*

Let $\hat{m}^k_\delta$ be the moments produced by applying this procedure to the uniform distribution.

For these $\hat{m}, M$ we prove:

Theorem 7 (Matching Moments Theorem). *For integers k, n and real number δ, the vector $\hat{m}^k_\delta$ and the function M^k_δ of Definition 12 are such that for distribution p such that $\forall i, p(i)\leq\frac{1}{k}$, letting $\hat{p}\leftarrow M^k_\delta(p)$ and $\hat{k} = \frac{k\delta}{100\cdot 2^{3\sqrt{\log n}}}$ we have*

- *For all $i\in[n]$, $\hat{p}(i)\leq 1/\hat{k}$;*
- *$|p-\hat{p}|\leq\delta$*
- *The $\hat{k}$-based ath moment of $\hat{p}$, for $a\leq\sqrt{\log n}$ equals $\hat{m}$ to within $\frac{1}{10000\log n}$.*

Proof of the Matching Moments Theorem. We first show that the definition of M is valid.

We note that $\hat{m}$ is indeed an upper-bound on m: when $a = 1$ we have $m(1) = \overline{m}(1) = 0$; otherwise, since $p(i)\leq\frac{1}{k}$ for each i, the κ-based moments are bounded as $m(a)\leq\sum_i\hat{p}(i)(\frac{1}{k})^{a-1}\cdot\kappa^a\leq\frac{\kappa^2}{k}\sum_i\hat{p}(i)\leq\frac{\kappa^2}{k}$, as desired. The fact that $\overline{V}^{\mu I}$ bounds the magnitudes of the elements of $inv(V^\mu)$ is Lemma 8. Since $\overline{V}^{\mu I}$ and $\overline{m}$ respectively bound the magnitudes of $inv(V^\mu)$ and m, their product $\bar{c}$ bounds the magnitudes of $\hat{c}$. Thus each of the expressions $\lfloor \bar{c}(\gamma) - \hat{c}(\gamma) \rfloor$ is nonnegative and Step 4 can be carried out.

We now show that Step 5 can be carried out. Note that the total frequency contribution of the elements added in Step 4 is just $\frac{1}{\kappa}$ times the κ-based first moment computed as $V_1^\mu \cdot c$, where V_1^μ denotes the first row of V^μ. We note that V_1^μ has entries 1 through μ, with sum $\frac{\mu(\mu+1)}{2}$. Since $\bar{c}$ bounds the magnitude of $\hat{c}$ and $c = \lfloor \bar{c} - \hat{c} \rfloor$, we have that the entries of c are bounded by corresponding entries of $2\bar{c}$. Further, each of these entries we may compute explicitly from the definition as $2\frac{(\mu-1)\kappa^2 6^\mu}{k}$. Thus the total new weight from Step 4 is at most $\frac{\mu^3 \kappa 6^\mu}{k} = \frac{\delta'}{4}$. By construction, the weight before Step 4 is at least $1 - \delta'$, and cannot exceed this by more than the highest frequency in p, which is at most $\frac{1}{k} \leq \frac{\delta}{100}$. Thus the total weight of $\hat{p}$ is at most $1 - \frac{\delta'}{2}$ by the end of Step 4. Further, because each element we added to the distribution has frequency (much) greater than $\frac{1}{k}$, and each element we removed from p in Step 1 had frequency less than $\frac{1}{k}$, the number of nonzero elements in $\bar{p}$ by Step 4 is no greater than $n(1 - \frac{\delta'}{2})$, so the elements "fit", and we have proven consistency of the construction.

The first property of the theorem follows trivially from the construction.

The second property of the theorem follows from the fact that in Step 1 we removed at most δ' weight from the distribution, and in the remaining steps we only added weight. Thus the distribution has changed by at most $2\delta' = \delta$.

We now examine the moments of the resulting distribution. We note that the first μ moments would be exactly the vector $V^\mu \cdot \bar{c}$ save for two caveats: the rounding in Step 4 and the new elements added in Step 5.

We note that rounding affects the ath κ-based moment by at most (one times) the sum of the absolute values of the entries of the ath row of V^μ, which we represent as $|V_a^\mu|$ and analyze later.

We analyze Step 5 by noting that the total weight added in Step 5, namely the gap between 1 and the weight at the end of Step 4, is controlled by the linear equations, up to rounding errors. Thus the difference between the maximum and minimum weight possibly added is at most the total weight of (one copy each of) the elements $\frac{1}{\kappa}, \frac{2}{\kappa}, \ldots, \frac{\mu}{\kappa}$, which equals $\frac{\mu(\mu+1)}{2\kappa} \leq \frac{\mu^2}{\kappa}$. Since the total weight to be added is at most δ' and the number of entries this weight is divided among is $n\frac{\delta'}{2}$, we bound the gap between the maximum and minimum values of the ath κ-based moment using the inequality $x^a - (x(1-y))^a \leq yax^{a-1}$ by $\kappa^a \frac{\mu^2}{\kappa} a \left(\frac{2}{n}\right)^{a-1} \leq \mu^3 \frac{2\kappa}{n}$. Since $n \geq k$ (otherwise we could not have $\forall i, p(i) \leq \frac{1}{k}$), by definition of κ (Definition 12) this expression is at most 1.

Thus, for any fixed a between 2 and μ the difference between the maximum and minimum κ-based moments reached by M, from any starting distribution p, is at most $1 + |V_a^\mu|$. Since the elements of the ath row of V^μ are the values γ^a for $1 \leq \gamma \leq \mu$, the sum $|V_a^\mu|$ consists of μ integer elements, all at most μ^a and some strictly less, so $1 + |V_a^\mu| \leq \mu^{a+1}$.

To convert this bound on the κ-based moments to a bound on the $\hat{k}$-based moments we multiply by $(\frac{\hat{k}}{\kappa})^a$ where $\frac{\hat{k}}{\kappa} = \frac{8\mu^3 6^\mu}{100 \cdot 2^{3\sqrt{\log n}}} \leq \frac{1}{100\mu^2}$, where the last equality holds for large n asymptotically, and for $n > 3$ by inspection for small integer values of μ. Thus the bound on the variation of the $\hat{k}$-based moments is $\mu^{a+1}(\frac{1}{100\mu^2})^a \leq \frac{1}{10000\mu^2}$ for $a \geq 2$, and 0 for $a < 2$, as desired. $\Box$

6. THE CANONICAL TESTING THEOREM

In this section we prove the main results of this work, the Low Frequency Blindness and Canonical Testing theorems (Theorems 4 and 5 as stated in Section 3). First we show how to combine the results of the previous two sections to show a general class of lower-bounds for testing symmetric weakly-continuous properties. Then we show that these lower-bounds apply in almost exactly those cases where the Canonical Tester fails, providing a tight characterization of the sample complexity for any symmetric weakly-continuous property.

The lower-bound we present completes the argument we have been making in the last few sections that *testers cannot make use of the low-frequency portion of distributions.* Explicitly, if we have two distributions p^+, p^- that are identical on their high-frequency indices then the tester may as well return the same answer for both pairs. Thus if a property takes very different values on p^+ and p^- then it is not testable. We first show this result for the case where neither distribution has high-frequency elements —this lemma is a simple consequence of the combination of the Wishful Thinking and Matching Moments theorems.

LEMMA 9. *Given a symmetric property π on distributions on $[n]$ that is (ϵ, δ)-weakly-continuous and two distributions, p^+, p^- all of whose frequencies are less than $\frac{1}{k}$ but where $\pi(p^+) > b$ and $\pi(p^-) < a$, then no tester can distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$ in $\frac{k\delta}{1000 \cdot 2^{4\sqrt{\log n}}}$ samples.*

PROOF. Consider the distributions obtained by applying the Matching Moments Theorem to p^+, p^-: let $\hat{p}^+ = M_\delta^k(p^+)$ and $\hat{p}^- = M_\delta^k(p^-)$. From the Matching Moments theorem's three conclusions we have that (1) the modified distributions have frequencies at most $1/\hat{k} = \frac{100 \cdot 2^{3\sqrt{\log n}}}{k\delta}$; (2) the statistical distance between each modified distribution and the corresponding original distribution is at most δ, which, since π is (ϵ, δ)-weakly-continuous implies that $\pi(\hat{p}^+) > b - \epsilon$ and $\pi(\hat{p}^-) < a + \epsilon$; and (3) the $\hat{k}$-based moments of $\hat{p}^+$ and $\hat{p}^-$ up to degree $\sqrt{\log n}$ are equal to within $\frac{2}{10000 \log n}$.

We then apply the corollary to the Wishful Thinking Theorem (Corollary 1) for $k = \hat{k}\frac{1}{10 \cdot 2^{\sqrt{\log n}}}$. (The k we use for the Wishful Thinking theorem is different from the k used in the previous paragraph for the Matching Moments theorem; however, we retain $\hat{k}$ from the previous paragraph.) We note that the ath k-based moment is proportional to k^a, so since the $\hat{k}$-based moments of $\hat{p}^+$ and $\hat{p}^-$ match to within $\frac{2}{10000 \log n}$ and since $k < \hat{k}$, the k-based moments also match to within this bound. We may thus evaluate the condition of Corollary 1 as

$$\sum_{a=2}^{\sqrt{\log n}} \frac{|m^+(a) - m^-(a)|}{\sqrt{1 + \max\{m^+(a), m^-(a)\}}} \leq \sum_{a=2}^{\sqrt{\log n}} |m^+(a) - m^-(a)| \leq \frac{2\sqrt{\log n}}{10000 \log n} < \frac{1}{120},$$

and thus Corollary 1 yields the desired conclusion. $\Box$

We now easily derive the full Low Frequency Blindness theorem (Theorem 5).

Proof of the Low Frequency Blindness theorem. The intuition behind the proof is that the high-frequency samples give no useful information to distinguish between p^+, p^-, and the low frequency samples are covered by Lemma 9.

Let H be the set of indices of either distribution occurring with frequency at least $\frac{1}{k}$ and let $p_H = p^-|H(= p^+|H)$, namely the high-frequency portion of p^- and p^+. let $L = [n] - H$, and let $\ell = |p^+(L)|$, namely the probability that p^+ or p^- draws a low-frequency index.

Formally, we construct a property π' that is only a function of distributions on L, but can "simulate" the operation of π on both p^+ and p^-. We show how a tester for π would imply a tester for π', and conclude by invoking Lemma 9 to see that neither tester can exist.

Consider the following property π' on arbitrary distributions p_L with support L: define the function f mapping p_L to the distribution p on $[n]$ such that $p|H = p_H$, $p|L = p_L$, and the probability of being in L, $p(L)$, equals ℓ. Let $\pi'(p_L) = \pi(f(p_L))$.

Assume for the sake of contradiction that there exists a $\bar{k}$-sample tester T for $\pi_{a+\epsilon}^{b-\epsilon}$ (for some $\bar{k}$). We construct a $\bar{k}$-sample tester T' for $\pi'^{b-\epsilon}_{a+\epsilon}$ as follows: let k_L be the result of counting the number of heads in $\bar{k}$ flips of a coin that lands heads with probability ℓ; return the result of running T on input the concatenation of the first k_L samples input to T', and $\bar{k} - k_L$ samples drawn at random from p_H (defined above).

Clearly for any distribution p_L on L, running the above algorithm on $\bar{k}$ samples from p_L will invoke T being run on (a simulation of) $\bar{k}$ samples drawn from $f(p)$; thus since, by assumption, T distinguishes $\pi > b - \epsilon$ from $\pi < a + \epsilon$ we conclude that T' distinguishes $\pi' > b - \epsilon$ from $\pi' < a + \epsilon$.

To finish the argument we show that this cannot be the case. Note that since f is a linear function with coefficients $\ell \leq 1$, the (ϵ, δ)-weak-continuity of π implies the (ϵ, δ)-weak-continuity of π'. Further, we have that $p^+|L$ and $p^-|L$ are both $\ell \cdot k$-low-frequency distributions, where by definition, $\pi'(p^+|L) > b$ and $\pi'(p^-|L) < a$. We thus invoke Lemma 9 on $\pi', p^+|L, p^-|L$, and $\ell \cdot k$ to conclude that no tester can distinguish $\pi' > b-\epsilon$ from $\pi' < a+\epsilon$ in $\frac{\ell k \delta}{1000 \cdot 2^{4\sqrt{\log n}}}$ samples, which implies from the previous paragraph that no tester can distinguish $\pi > b - \epsilon$ from $\pi < a + \epsilon$ in the same number of samples.

To eliminate the ℓ from this bound requires a slightly tighter analysis, for which we refer the reader to [21]. $\square$

We conclude with a proof of the Canonical Testing theorem (Theorem 4), making use of the following lemma:

Lemma 10. *Given a distribution p and parameter θ, if we draw k random samples from p then with probability at least $1 - \frac{4}{n}$ the set P constructed by the Canonical Tester will include a distribution $\hat{p}$ such that $|p - \hat{p}| \leq 24\sqrt{\frac{\log n}{\theta}}$.*

The proof is elementary: use Chernoff bounds on each index i and then apply the union bound to combine the bounds.

Proof of the Canonical Testing theorem. Without loss of generality assume that the Canonical Tester fails by saying "no" at least a third of the time on input samples from some distribution p when in fact $\pi_a^b(p) > b + \epsilon$. From the definition of the Canonical Tester this occurs when, with probability at least $\frac{1}{3}$, the set P constructed contains a distribution p^- such that $\pi(p^-) < a$. From Lemma 10, P contains some p^+ within statistical distance δ from p with probability at least $1 - \frac{4}{n}$. Thus by the union bound there exists a single P with both of these properties, meaning there exist such p^-, p^+ lying in the same P, and thus having the *same* θ-high-frequency components. Since π is (ϵ, δ)-weakly-continuous, $\pi(p^+) > b$. Applying the Low Frequency Blindness Theorem to p^+, p^- yields the desired result. $\square$

7. ACKNOWLEDGEMENTS

I am indebted to Ronitt Rubinfeld for introducing me to these beautiful problems, and for her inexhaustible encouragement and guidance during the long process of preparing these results for publication. I would also like to thank Silvio Micali, Arnab Bhattacharyya, and Ning Xie for the significant time they devoted to helping me edit this document.

8. REFERENCES

[1] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and Almost k-wise Independence. *STOC* 2007.

[2] N. Alon, E. Fischer, I. Newman, and A. Shapira. A combinatorial characterization of the testable graph properties: it's all about regularity. *STOC* 2006.

[3] T. Batu. *Testing Properties of Distributions.* Ph.D. thesis, Cornell University, 2001.

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

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

[6] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. "Testing that distributions are close". *FOCS*, 2000.

[7] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. *STOC*, 2004.

[8] M. Blum and S. Kannan. Designing Programs that Check Their Work. *STOC* 1989.

[9] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences* 47(3):549Ű595, 1993.

[10] A. Chakrabarti, G. Cormode, and A. McGregor. A Near-Optimal Algorithm for Computing the Entropy of a Stream. *SODA* 2007.

[11] M. Charikar, S. Chaudhuri, R. Motwani, and V.R. Narasayya. Towards Estimation Error Guarantees for Distinct Values. *PODS*, 2000.

[12] O. Goldreich, S. Goldwasser, and D. Ron. Property Testing and Its Connection to Learning and Approximation. *FOCS* 1996.

[13] O. Goldreich and L. Trevisan. Three theorems regarding testing graph properties. *Random Struct. Algorithms*, 23(1):23-57, 2003.

[14] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and Sublinear Approximation of Entropy and Information Distances. *SODA* 2006.

[15] A. Klinger. The Vandermonde Matrix. *The American Mathematical Monthly*, 74(5):571-574, 1967.

[16] P. Indyk and A. McGregor. Declaring Independence via the Sketching of Sketches. *SODA* 2008.

[17] P. Indyk and D. Woodruff. Tight Lower Bounds for the Distinct Elements Problem. *FOCS*, 2003.

[18] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem. *FOCS* 2007.

[19] B. Roos. On the Rate of Multivariate Poisson Convergence. *Journal of Multivariate Analysis* 69:120-134, 1999.

[20] R. Rubinfeld and M. Sudan. Robust Characterizations of Polynomials with Applications to Program Testing. *SIAM Journal on Computing* 25(2):252-271, 1996.

[21] P. Valiant. Testing Symmetric Properties of Distributions. *ECCC report TR07-135*, 2007.