Electronic Colloquium on Computational Complexity, Report No. 135 (2007)

Testing Symmetric Properties of Distributions

Paul Valiant

December 26, 2007

Abstract

We introduce the notion of a *Canonical Tester* for a class of properties, that is, a tester strong and general enough that "a property 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.

ISSN 1433-8092

1 Introduction

Property testing has been extensively investigated in a variety of setting, in particular, program checking (starting with [7, 8]), testing of algebraic properties (starting with [20]), and graph testing (starting with [12]). This advanced state of knowledge is evidenced by the emergence of general structural theorems, most the characterization by Alon et. al. of those graph properties testable in constant time [2]

By contrast, the emerging and significant subfield of *distribution testing* is currently a collections 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, 6, 4], as well as "decision versions" of support size, entropy, and statistical and L_2 distance[3, 5, 10, 13, 9, 15, 18, 16].

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 for a second example, uniformity testing asks whether all elements in the distribution have the same probability, or whether this is far from being the case. Again, it is clear that this property does not depend on the specific naming scheme for the domain elements.

Lower- and Upper-Bounds. Answering a distribution testing question requires two components, an upper-bound and a lower-bound, each expressed as 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. For instance, the upper- and lower-bounds of statistical-distance testing differ by a factor of $n^{1/6}$. Similar gaps exist in the published bounds for many other symmetric properties. Perhaps intuitively, the techniques developed for the upper-bounds differ from those developed for the lower-bounds. In the first case, they look for an "algorithm", in the second case, an "impossibility proof".

1.1 Our Results

We prove the following three informally stated results, the first and third resolving open problems from [5, 3, 18]:

Theorem 1. *Distinguishing identical distributions from distributions with statistical distance $\frac{1}{2}$ requires $O(n^{2/3})$ samples.*

Theorem 2. *For any constants $0 < \alpha < \beta < 2$, 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.*

More importantly, perhaps, we prove Theorems 2 and 3 by developing a unified framework for optimally answering distribution testing questions for a large class of properties.[1]

[1] We note that our techniques are general enough that they may be further used to reproduce the main result of [18].

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.[2] 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. If the Canonical Tester successfully 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 binary-search for f".

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 actual frequencies, the tester may as well be invariant under permutation of the *observed* 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 when different types of collisions appear has proven to be very difficult. One of our main technical contributions is what we call the *Wishful Thinking Theorem*. Analyzing the statistics of collisions would be easy if the distributions involved were independent gaussians. The Wishful Thinking Theorem guarantees that treating the collision statistics as independent gaussians 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 used for 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 with frequency at least $\frac{\log n}{n}$) will be well-approximated by the *observed* frequencies of these items in the drawn sample. (If we are interested in a continuous property of the distribution, then an approximation of the distribution is 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, else the distribution may essentially disappear or be distorted beyond recognition.

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 items: we call this the *Low-Frequency Blindness Theorem*. This result considerably simplifies our Canonical Tester: the high-frequency elements it can well-approximate; the low-frequency ones it may ignore.

Continuity and Approximations. We note that, given a function f from distributions to real numbers, there are essentially two ways to formulate a "property" from f: we can ask for testers to distinguish between the cases when $f(p) = a$ (YES) and $f(p) = b$ (NO), or we can ask for testers to distinguish between the cases when $f(p) < a$ (YES) and $f(p) > b$ (NO). In this paper we take the second option —approximation

[2]Technically this is *uniform continuity* and not *continuity*; however, since the space of probability distributions over $[n]$ is compact, every continuous function here is also uniformly continuous.

properties— for the simple reason that this is the domain where continuity can be leveraged. Continuity yields statements of the form "if I nudge p by δ then $f(p)$ will change by at most ϵ." In order to allow for these "nudges" to f we work mainly with approximation properties instead of the strict $= a$ or $= b$ properties.[3] An interesting illustration of the distinction between these types of properties is given by the case of statistical distance: [5] exhibited an algorithm for the strict version taking roughly $n^{2/3}$ samples, a result which we show tight in Theorem 1. however, for the distance *approximation* problem we show that the optimal tester takes roughly n samples.

Roadmap For reasons of space, much of the technical material is moved to the appendix. Each of the main sections in the body of the paper has a section in the appendix that contains the technical details. The sections are arranged as follows: After a quick review of definitions, we introduce and discuss the Canonical Tester as an algorithm. The remaining sections lead up to the proof that is is also a lower bound. We start this process with a review of some elements of the standard toolkit for lower-bounding property testers, which we summarize as the Generalized Positive-Negative Distance Lemma. Following this, we derive the Wishful Thinking Theorem, which has as an immediate application the proof of the $O(n^{2/3})$ lower bound for strict statistical distance testing. The next step is the Matching Moments Theorem. In the final section we prove the Low Frequency Blindness Theorem, which implies the Canonical Testing Theorem; we conclude with applying these results to establish the sample complexity of approximating statistical distance and entropy.

2 Definitions

For positive integers n we let $[n]$ denote the integers $\{1, \ldots, n\}$. For real numbers a, b we let $[a, b]$ denote the interval containing all x between a and b, inclusive. Logarithms are base 2 unless denoted "$\log_e$". We denote elements of vectors with functional notation —as $v(i)$ for the ith element of v— to limit proliferation of subscripts.

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

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

Definition 2. *A property of a single distribution is a function* $\pi : \mathcal{D}_n \to \mathbb{R}$*;*
a property of a distribution pair is a function $\pi : \mathcal{D}_n \times \mathcal{D}_n \to \mathbb{R}$.

As we do not expect a tester to distinguish between a continuum of possible output values, we ask testers to decide *binary properties*:

Definition 3. *A binary property of a single distribution is a function* $\pi : \mathcal{D}_n \to \{$ *"yes", "no",* $\emptyset\}$*;*
a binary property of a distribution pair is a function $\pi : \mathcal{D}_n \times \mathcal{D}_n \to \{$ *"yes", "no",* $\emptyset\}$.

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

Definition 4. *Given a binary property* π *on pairs of distributions, real numbers* $0 < a < b < 1$*, and a function* $k : \mathbb{Z}^+ \to \mathbb{Z}^+$*, an algorithm* T *is a "*(a, b) *U-tester with sample complexity* $k(\cdot)$*" if, for any distribution pair* p_1, p_2*, algorithm* T *on input* $k(n)$ *random samples from* p_1 *and* $k(n)$ *random samples from* p_2 *will accept with probability at least* b *if* $\pi(p_1, p_2) =$ *"yes", and accept with probability at most* a *if* $\pi(p_1, p_2) =$ *"no". The behavior is unspecified when* $\pi(p_1, p_2) = \emptyset$.

[3] Of course, any strict property can be converted into an approximation property by changing f, but this change may not preserve continuity.

When a and b are not specified, we take them to be $\frac{1}{3}$ and $\frac{2}{3}$ respectively. We may refer to $b-a$ as the *soundness* of a tester.

The metric we use on probability distributions p, and vectors more generally is the L_1 norm, $|p| \triangleq \sum_i |p_i|$. In particular, for two probability distributions p^+, p^- we may define the *statistical distance* as $|p^+ - p^-|$. (In some references there is a normalization constant of $\frac{1}{2}$ here.) We may now define our notion of continuity:

Definition 5. *A property π of a single distribution is (ϵ, δ)-weakly-continuous if for all distributions p^+, p^- satisfying $|p^+ - p^-| \leq \delta$ we have $|\pi(p^+) - \pi(p^-)| \leq \epsilon$. A property π of a distribution pair is (ϵ, δ)-weakly-continuous if for all distributions $p_1^+, p_2^+, p_1^-, p_2^-$ satisfying $|p_1^+ - p_1^-| + |p_2^+ - p_2^-| \leq \delta$ we have $|\pi(p_1^+, p_2^+) - \pi(p_1^-, p_2^-)| \leq \epsilon$.*

Finally, we define symmetric properties:

Definition 6. *A property π of a single distribution is symmetric if for all distributions p and all permutations σ we have $\pi(p) = \pi(p \circ \sigma)$. A property of a distribution pair is symmetric if for all distributions p_1, p_2 and all permutations σ we have $\pi(p_1, p_2) = \pi(p_1 \circ \sigma, p_2 \circ \sigma)$.*

3 The Canonical Tester

We introduce Canonical Testing by way of the following observation: given a distribution p and an index i, the number of times the ith element of p will occur in k samples is modeled by the binomial distribution $\mathrm{Bin}(k, p(i))$, which has the property that for $p(i) \gg \frac{1}{k}$, the distribution will return a value close to $\frac{p(i)}{k}$ with high probability; on the other hand if $p(i) \ll \frac{1}{k}$ the distribution will return 0 with high probability. Thus elements of a distribution fall into essentially two regimes: any high-frequency element is well-approximable from random samples, and any low-frequency element is almost invisible to random samples. This naturally motivates a tester that, when an element i is sampled a large number of times, estimates $p(i)$ from this number, and for those elements that are not sampled often, it declares the elements to be low-frequency but inapproximable beyond that; if these estimates for the high-frequency elements and smallness bounds for the low-frequency elements uniquely specify the property to be tested, then it returns this as the answer. We define this tester more formally as follows, where since we primarily work with properties of distribution pairs we define the tester for distribution pairs.[4]

Definition 7 (Canonical Tester)**.** *Given a property π on distribution pairs with support $[n]$, and supposing k samples are drawn from each of a pair of such distributions, with $s_1(i)$ counting the number of times element i is sampled from the first distribution and $s_2(i)$ counting the number of times i is sampled from the second distribution, then the* k-sample $\mathcal{T}_\epsilon^c$ tester *for distinguishing $\pi < a$ from $\pi > b$ returns an answer "$< a$" or "$> b$" from s according to the following steps.*

(1) *For each i such that $s_1(i) > c$ or $s_2(i) > c$ insert the constraints $p_1(i) = \frac{s_1(i)}{k}$ and $p_2(i) = \frac{s_2(i)}{k}$.*

(2) *For the remaining i insert the constraints $p_1(i), p_2(i) \in [0, \frac{c}{k}]$*

(3) *Insert the constraints $\sum_i p_1(i) = 1$ and $\sum_i p_2(i) = 1$.*

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

(5) *If $\forall (p_1, p_2) \in P, \pi(p_1, p_2) \geq a + \epsilon$ then return "$> b$"; if $\forall (p_1, p_2) \in P \pi(p_1, p_2) < b - \epsilon$ then return "$< a$"; if both or neither of these conditions apply, return an arbitrary answer.*

We refer to the portions of p_1, p_2 specified exactly by the constraints in Step 1 as the c-high-frequency approximation*, and to P as the set of* low-frequency extensions *to the c-high-frequency approximation.*

For symmetric, (ϵ, δ)-weakly-continuous properties we will see that the proper parameters of $\mathcal{T}$ are $\epsilon = \epsilon$ and $c = \frac{600 \log n}{\delta^2}$.[5]

To provide some justification for why the system of constraints is reasonable we have the following:

[4] It is syntactically straightforward to extend this tester to apply to cases with greater or fewer distributions by simply changing the set of subscripts on s and p from $\{1, 2\}$ to a different set.

[5] See the theorems of the final section.

Lemma 1. *Given a distribution pair (p_1, p_2) and constant c, drawing k random samples from each distribution, with probability at least $1 - \frac{4}{n}$ the set of low-frequency extensions to the c-high-frequency approximation of (p_1, p_2) will include a pair $(\bar{p_1}, \bar{p_2})$ such that $|p_1 - \bar{p_1}| + |p_2 - \bar{p_2}| \leq 24\sqrt{\frac{\log n}{c}}$.*

The proof is elementary, using Chernoff bounds for each i and then applying the union bound to combine the bounds. We defer it to the appendix.

Discussion. It is not immediately clear why or when this tester will work (i.e. achieve the conditions of Definition 4). Further, this tester is described as a function, not an algorithm, completely bypassing issues of computational complexity. Nevertheless, we show in the course of this work that for any symmetric weakly-continuous property π and suitably chosen c, ϵ, if the k-sample $\mathcal{T}_\epsilon^c$ tester does not correctly test between $\pi < \alpha$ and $\pi > \beta$, essentially nothing will.[6] It is for this reason that we call $\mathcal{T}$ canonical.

It is also not immediately clear why *symmetric* and *weakly-continuous* are related to $\mathcal{T}$, since the tester could conceivably be applied to a much wider class of properties.[7] Indeed we suspect that this tester —or something very similar— may be shown optimal for more general properties. However, neither the symmetry or the continuity condition can be relaxed entirely:

- Consider the problem of determining whether a (single) distribution has more weight on its first half or its second half. Specifically, on distributions of support $[n]$ let $\pi(p) = |p(\{1, \ldots, \lfloor \frac{n}{2} \rfloor\})| - |p(\{\lfloor \frac{n}{2} \rfloor + 1, \ldots, n\})|$. We note that π is continuous but not symmetrical. It is fairly clear that π can be easily approximated from a constant(!) number of samples s by taking the difference between the number of samples in the first half of the distribution and the number in the second half of the distribution, and dividing by the total number of samples. Further, this tester will often return the correct answer even when each frequency in p is in $[0, \frac{2}{n}]$. However, the Canonical Tester will *discard* all such frequencies unless $\frac{c}{k} < \frac{2}{n}$, that is, if the number of samples is essentially 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 determining whether a distribution pair is identical or far apart was analyzed in [5], where they constructed a $\tilde{\theta}(n^{2/3})$-sample tester. (Recall our Theorem 1 for the definition and a matching lower-bound.) This problem 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, where π is seen to be symmetric, but *not* continuous. It can be seen that the Canonical Tester for π requires $\tilde{\theta}(n)$ samples (this follows trivially from our Theorem 2), which is $\sim n^{1/3}$ worse than the optimal tester.

4 Step 0: The Generalized Positive-Negative Distance Lemma

The Positive-Negative Distance Lemma states for general symmetric properties a result that appears in the literature in the context of specific properties such as entropy [3]. This lemma provides a general condition for when properties are not testable; in the rest of the paper build up a sequence of results that lets us apply this condition in those cases where the Canonical Tester fails.

The proof of this lemma synthesizes three different techniques. These techniques address in turn three evident difficulties in deriving lower-bounds: (1) when taking k samples from a distribution, the number of times the first element is sampled is (anti-) correlated with the number of times the second, third, and other elements are sampled; (2) the complete record of each of the $2k$ samples is a lot of data to analyzes; and (3) we have no idea how the tester will make its decision from this data.

This material appears in full in the appendix; we omit all but the crucial definitions here.

Definition 8. *A Poisson process with parameter $\lambda \geq 0$ is a distribution over the nonnegative integers where the probability of choosing c is defined as* $\text{poi}(c; \lambda) \triangleq \frac{e^{-\lambda}\lambda^c}{c!}$. *We denote the random variable as $Poi(\lambda)$. For*

[6] The qualifier "essentially" is made precise in Theorem 7 (we lose a small factor in k and small constants in α and β).

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

vectors $\lambda \in \mathbb{R}^{+k}$ for $k \in \mathbb{Z}^+$ we let $Poi(\lambda)$ denote the k-dimensional random variable whose ith component is drawn from the univariate $Poi(\lambda(i))$ for each i.

Given a tester T with sample complexity k on distributions p_1, p_2, we modify it in an essentially trivial way, but one which changes the analysis drastically. The following process is called *Poissonization*:

1. Draw two numbers k_1, k_2 from the Poisson process $Poi(k)$.
2. Draw k_1 samples from p_1 and k_2 samples from p_2.
3. If either $k_1 < k$ or $k_2 < k$, FAIL.
4. Otherwise, return the result of running T on the first k samples drawn from each distribution.

Definition 9. *Given two multisets of samples S_1, S_2 drawn from distributions with finite support set X, the* fingerprint *of S_1, S_2 is a function $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \to \mathbb{Z}^+$ such that $f(i, j)$ is number of elements of X that appear exactly i times in S_1 and j times in S_2.*

Lemma 2. *For any symmetric property U and random variable κ, if there exists a κ-sample tester T then there exists a κ-sample tester T' which takes as input only the fingerprint of κ samples drawn from each distribution.*

For the rest of this paper when we refer to a tester we will generally consider its input to be in fingerprint form.

Definition 10. *Given distributions p_1, p_2 with support $[n]$ and a positive integer k, define $D^k_{p_1,p_2}$ to be the distribution of fingerprints of the following sampling process:*

1. *Draw two numbers k_1, k_2 from the Poisson process $Poi(k)$.*
2. *Draw k_1 samples from p_1 and k_2 samples from p_2.*

Lemma 3 (Positive-Negative Distance)**.** *If π is a symmetric property testable in k samples then for any positive distribution pair p_1^+, p_2^+ and any negative distribution pair p_1^-, p_2^-, we have $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \geq \frac{1}{12}$.*

5 Step 1: The Wishful Thinking Theorem

In this section we derive a general theorem for upper-bounding expressions of the form $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}|$. Specifically, we focus on the case where each of the probabilities in $p_1^+, p_2^+, p_1^-, p_2^-$ are (sufficiently) less than $\frac{1}{k}$. (We will analyze the case of larger frequencies in the final section. As we will see there, Lemma 12 allows us essentially to analyze these low and high frequency cases separately.)

The Wishful Thinking Theorem yields a bound in terms of the *moments* of each distribution pair, defined as follows: (We use a normalization constant k so as to keep the moments of a reasonable size when distributions' frequencies are all on the order of $\frac{1}{k}$.)

Definition 11. *Given a distribution pair p_1, p_2 and a positive number k, then for each pair of nonnegative integers (a, b) define the k-*based (a, b) moment of (p_1, p_2) *as $\sum_i k^{\alpha+\beta} p_1(i)^\alpha p_2(i)^\beta$.*

We motivate the result of this section and its name with the following "wishful thinking" analysis, of $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}|$. 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.

> Recalling the alternative definition of $D^k_{p_1,p_2}$ provided by Lemma 12, consider the contribution to the (a, b) fingerprint entry provided by the ith elements of p_1 and p_2: by definition this will be 1 with probability $\text{poi}(a, k \cdot p_1(i)) \cdot \text{poi}(b, k \cdot p_2(i))$. Taking small liberties with the definition of the Poisson distribution, we will approximate this as $k^{a+b} p_1(i)^a p_2(i)^b$. Thus the expected value of the (a, b) fingerprint entry is (roughly) $k^{a+b} \sum_i p_1(i)^a p_2(i)^b$, which is the (a, b)th moment of the pair (p_1, p_2). Applying further wishful thinking, we approximate the distribution of (a, b) as

being a gaussian with mean and variance exactly this moment, and take it to be *independent* of all of the other fingerprint entries. Having modeled $D^k_{p_1,p_2}$ as independent gaussians, and using the approximation that a gaussian with mean and variance α, and a gaussian with mean and variance β have statistical distance roughly $\frac{|\alpha-\beta|}{\sqrt{\max\{\alpha,\beta\}}}$, we propose the following theorem:

Theorem 4 (Wishful Thinking)**.** *Given probability distribution pairs p_1^+, p_2^+ and p_1^-, p_2^- and positive number ϵ and integer k such that each probability of any element in each distribution is bounded by $\frac{\epsilon}{k}$, let $m^+_{a,b}$ be the k-based (a,b) moment of (p_1^+, p_2^+) for each $a, b \geq 0$, with $m^-_{a,b}$ defined correspondingly for (p_1^-, p_2^-). Then*

$$|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \leq 40\epsilon + 10\sum_{a,b} \frac{|m^+_{a,b} - m^-_{a,b}|}{\lfloor \frac{a}{2} \rfloor! \lfloor \frac{b}{2} \rfloor! \sqrt{1 + \max\{m^+_{a,b}, m^-_{a,b}\}}}.$$

We note that the expression $\frac{|m^+_{a,b}-m^-_{a,b}|}{\sqrt{1+\max\{m^+_{a,b},m^-_{a,b}\}}}$ is bounded by several simpler expressions, including $|m^+_{a,b} - m^-_{a,b}|$, $\frac{|m^+_{a,b}-m^-_{a,b}|}{\sqrt{\max\{m^+_{a,b},m^-_{a,b}\}}}$, and $|\sqrt{m^+_{a,b}} - \sqrt{m^-_{a,b}}|$, either of which could be used instead in the theorem for the sake of convenience; similarly, the $\lfloor \frac{a}{2} \rfloor! \lfloor \frac{b}{2} \rfloor!$ term could be dropped. In this paper we do not need the full strength of the Wishful Thinking Theorem, but since it may be useful in other contexts, we prove the full theorem.

We start towards a proof of this theorem by noting that the distribution $D^k_{p_1,p_2}$ as constructed via Lemma 12 is an example of what is sometimes known as a *generalized multinomial distribution*, defined in general as the distribution of the histogram of the sum of independent random variables.

Definition 12. *The* generalized multinomial distribution *parameterized by matrix ρ, denoted M^ρ, is defined by the following random process: for each row of ρ, draw a column from the distribution ρ_i; return a row vector recording the total number of samples falling into each column.*

Lemma 4. *For any distributions p_1, p_2 with support $[n]$ and positive integer k, the distribution $D^k_{p_1,p_2}$ is the generalized multinomial distribution M^ρ where matrix ρ has n rows, columns indexed by pairs of nonnegative fingerprint indices (a,b), and $(i,(a,b))$ entry equal to $\mathrm{poi}(a; k \cdot p_1(i))\mathrm{poi}(b; k \cdot p_2(i))$ for each $i \in [n]$ and $a, b \in \mathbb{Z}^+$.*

Proof. Clear from the definition of $D^k_{p_1,p_2}$ in Lemma 12. □

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, with the vector of column sums defined as $\lambda(d) = \sum_i \rho(i,d)$, then*

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

Applying this theorem and the previous lemma we see that $D^k_{p_1,p_2}$ may be well-approximated by a multivariate Poisson process. The bound is shown in the following lemma:

Lemma 5. *Given a probability distribution pair p_1, p_2, a real number $\epsilon > 0$ and a positive integer k such that each probability is at most $\frac{\epsilon}{k}$, let ρ be the matrix such that $D^k_{p_1,p_2} = M^\rho$, and let $\lambda(d) = \sum_i \rho(i,d)$. Then $|D^k_{p_1,p_2} - \mathrm{Poi}(\lambda)| \leq 20\epsilon$ provided $\epsilon \leq \frac{1}{30}$.*

See the appendix for the proof.

To complete our analysis of $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}|$ we derive bounds on the distance between the Poisson approximations of $D^k_{p_1^+,p_2^+}$ and $D^k_{p_1^-,p_2^-}$, and apply the triangle inequality. See the appendix for details. We note that the higher-order moments vanish rapidly, so under slightly modified conditions we can show that it is sufficient to bound only the sum of those moments of degree at most $\sqrt{\log n}$ see the appendix for details.

The Closeness Testing Lower Bound The proof of Theorem 1 is a realization of the outline that appeared in [5], making crucial use of the Wishful Thinking Theorem. See the appendix for details.

6 Step 2: The Matching Moments Theorem

In the previous section we showed that the moments capture essentially all the information that can be extracted from the low-frequency elements of a distribution. In short, moments are all that matter in the low-frequency setting. In this section we go further and show that even moments do not matter for the important special case of weakly-continuous properties. In essence, no useful information can be extracted from the low-frequency portion of a distribution. We will use these results in the next section to conclude that *distribution testing is possible if and only if it can be done solely using the high-frequency elements.*

The main result of this section is the Matching Moments Theorem which states that for any constants k, w we can modify the $\leq \frac{1}{k}$-frequency portion of any distribution pair p_1, p_2 by shifting at most w weight so as to rewrite the moments of p_1, p_2 —if we slightly relax the $\leq \frac{1}{k}$ frequency condition on the output distributions. Since any weakly-continuous property of p_1, p_2 will be preserved under such small changes, moments do not help test properties.

As in the previous section, for the sake of simplicity we present the main result for the case where there are no high-frequency elements. In the next section we show how to apply the Matching Moments Theorem to general distributions.

Theorem 5 (Matching Moments Theorem)**.** *There is a function M parameterized by $w \leq 1$ and $k > 1$ mapping distribution pairs p_1, p_2 whose frequencies all lie below $\frac{1}{k}$ to distribution pairs $(\bar{p}_1, \bar{p}_2) \leftarrow M_w^k(p_1, p_2)$ and a function f mapping such w, k to matrices of moments $\widetilde{m} \leftarrow f(w, k)$ such that, letting $\bar{k} = \frac{kw}{100 \cdot 2^{6\sqrt{\log n}}}$ we have*

- *For all $i \in [n]$, both $\bar{p}_1(i), \bar{p}_2(i) \leq \frac{1}{k}$;*
- *$|p_1 - \bar{p}_1| + |p_2 - \bar{p}_2| \leq w$*
- *The $\bar{k}$-based (a, b) moments of $(\bar{p}_1, \bar{p}_2)$, for $a + b \leq \sqrt{\log n}$ equal $\widetilde{m}$ to within $\frac{1}{10000 \log n}$.*

The key observation that lets us rewrite the moments of p_1, p_2 while changing the distributions only slightly is the following: the zeroth moment of (p_1, p_2) equals n and the first moments equal k times the sum of p_1 or p_2, namely just k so thus only the second and higher moments are relevant (with respect to an application of the Wishful Thinking Theorem); since the second and higher moments depend on high powers of the frequencies in (p_1, p_2), if we shift roughly w weight from elements with probabilities at most $\frac{1}{k}$ to frequencies more than $\frac{1}{w}$ factor higher, the new moments will dwarf the old moments, and thus be (roughly) independent of the moments of the original distribution.

A crucial step of our construction is to set up linear equations whose solution will tell us exactly how to shift the weight so as to cancel out the original moments. Because moments of a distribution are defined as linear combinations of the powers of the probabilities in the distribution, the coefficients of the linear transform we use will be a *Vandermonde* matrix.

Definition 13. *Given a vector z of length μ, the Vandermonde matrix generated by z is the $\mu \times \mu$ matrix with entries $z(i)^j$.*

Here we work with a particular special class of Vandermonde matrices.

Definition 14. *For positive integer μ let ℓ^μ be the $\mu \times \mu$ matrix with entries i^j for columns indexed by $1 \leq i \leq \mu$ and rows indexed by $0 \leq j \leq \mu - 1$.*

We use this matrix to compute moments *of single distribution* as shown by the following trivial lemma.

Lemma 6. *Let p be a probability distribution such that there exists a positive real number α and a vector of integers c of length μ such that for each $t \in [\mu]$, p contains $c(t)$ entries equal to $\alpha \cdot t$ and zeros elsewhere. Let $m = \ell^\mu \cdot c$. Then the t'th k-based moment of p equals $(k\alpha)^t m(t)$.*

In this work we deal with moments of distribution *pairs*, instead of single distributions, and for this reason we do not work with ℓ directly, but rather with its tensor product with itself.

For completeness' sake we include the following:

Definition 15. *Given a matrix X with rows and columns indexed respectively by a and u, and a matrix Y indexed by b and t, the tensor product $X \otimes Y$ is defined to be the matrix with rows indexed by pairs (a,b), columns indexed by pairs (t,u), and $((a,b),(t,u))$ entry defined by the product of the original entries from X and Y as $X(a,t) \cdot Y(b,u)$.*

Definition 16. *For positive integer μ let $L^{(\mu)} = \ell^\mu \otimes \ell^\mu$ be the $\mu^2 \times \mu^2$ matrix with entries $t^a u^b$ for columns indexed by pairs $1 \leq (t,u) \leq \mu$ and rows indexed by pairs $0 \leq (a,b) \leq \mu - 1$.*

The generalization of Lemma 6 is:

Lemma 7. *Let p_1, p_2 be a probability distribution pair such that there exists a positive real number α and a vector of integers c indexed by pairs $1 \leq (t,u) \leq \mu$ such that for each (t,u), there are $c((t,u))$ indices such that $p_1(i) = \alpha \cdot t$ and $p_2(i) = \alpha \cdot u$ and zeros elsewhere. Let $m = L^\mu \cdot c$. Then the (t,u)th k-based moment of (p_1,p_2) equals $(k\alpha)^{t+u} m((t,u))$.*

Now that we have expressed moments by linear equations, and aiming to solve these linear equations for the "target" moments $f(w,k)$ of the Matching Moments Theorem, it remains to bound the size of elements of the inverse of L^μ and then assemble the pieces. See the appendix.

7 Step 3: The Canonical Testing Theorem

In this section we prove the main results of this work. First we show how to combine the results of the previous three 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 distribution pairs (p_1^-, p_2^-) and (p_1^+, p_2^+) 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_1^-, p_2^-) and (p_1^+, p_2^+) then it is not testable.

Definition 17. *Given a distribution pair p_1, p_2 and positive integer k, the* k-high-frequency indices *are those $i \in [n]$ such that* $\max\{p_1(i), p_2(i)\} \geq \frac{1}{k}$.

Theorem 6 (Low Frequency Blindness)**.** *Given a property π on pairs of distributions on $[n]$ that is (ϵ,δ)-weakly-continuous and two pairs of distributions, (p_1^-, p_2^-) and (p_1^+, p_2^+) that are identical for any k-high-frequency index i but where $\pi(p_1^-, p_2^-) < a$ and $\pi(p_1^+, p_2^+) > b$, then no tester can distinguish between $\pi < a + \epsilon$ and $\pi > b - \epsilon$ in $\frac{k\delta}{100000 \cdot 2^{6\sqrt{\log n}}}$ samples.*

See the appendix for the proof.

Theorem 7 (Canonical Testing Theorem)**.** *Given a property π on pairs of distributions on $[n]$ that is (ϵ,δ)-weakly-continuous such that the Canonical Tester T_ϵ^c for $c = \frac{600 \log n}{\delta^2}$ fails to distinguish between $\pi < a$ and $\pi > b$ in k samples, then no tester can distinguish between $\pi < a + 2\epsilon$ and $\pi > b - 2\epsilon$ in $\frac{k\delta}{100000 \cdot 2^{6\sqrt{\log n}}}$ samples.*

Proof. Without loss of generality assume that the Canonical Tester fails by saying "$< a$" at least a third of the time when the correct answer is "$> b$". From the definition of the Canonical Tester, this occurs when there is a distribution pair (p_1, p_2) with $\pi(p_1, p_2) > b$ such that with probability greater than $\frac{1}{3}$ the c-high-frequency sampling approximation approximate has a low-frequency completion (p_1^-, p_2^-) with property $< a + \epsilon$. From Lemma 1, with probability at least $1 - \frac{4}{n}$ the sampling approximation has a low-frequency completion (p_1^+, p_2^+) within statistical distance δ from (p_1, p_2). Thus by the union bound there exists $p_1^-, p_2^-, p_1^+, p_2^+$ with the *same* c-high-frequency components. Since π is (ϵ,δ)-weakly-continuous, $\pi(p_1^+, p_2^+) > b - \epsilon$. Applying the Low Frequency Blindness Theorem yields the desired result. □

The Statistical Distance Approximation 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 2. Further, each of the frequencies in these distributions is at most $\frac{2}{n}$. We apply the Low Frequency Blindness Theorem with $\epsilon = \delta = \min\{\alpha, 2-\beta\}$ and $k = n^\theta$ for any $\theta < 1$ to yield the desired result. □

The Entropy Approximation Bound.

Lemma 8. *The entropy is* $(1, \frac{1}{2\log n})$*-weakly-continuous.*

See the appendix for the proof. We now prove a more formal statement of Theorem 3.

Lemma 9. *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 $\epsilon = 1$, using Lemma 8, to see that no tester can distinguish distributions with entropy at least $(\log n) - 1$ from those with entropy at most $\frac{\log n}{\gamma^2} - 1$ using fewer than $\frac{n^{1/\gamma^2}}{800000 \cdot 2^{6\sqrt{\log n}} \log n}$ queries. This implies the desired result. □

References

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

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

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

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

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

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

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

[8] Blum, M., Luby, M., and Rubinfeld, R. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences* 47:3 (Dec. 1993), 549595. (Preliminary Version in 22nd *STOC*, 1990).

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

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

[11] Cover, T. and Thomas, J. "Elements of Information Theory". 1991.

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

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

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

[15] Indyk, P. and Mcgregor, A. Declaring Independence via the Sketching of Sketches. *SODA* 2008.

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

[17] Mitzenmacher, M. Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[18] Raskhodnikova, S., Ron, D., Rubinfeld, R., Shpilka, A. and Smith, A. "Sublinear Algorithms for Approximating String Compressibility and the Distribution Support Size". *FOCS* 2007.

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

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

APPENDIX

Appendix to Section 3

Proof of Lemma 1. This proof is elementary, using Chernoff bounds for each i and then applying the union bound to combine these bounds. Explicitly, the Chernoff bounds we use have the following form: given k independent $\{0,1\}$ random variables where each one takes value 1 with probability σ, denoting their sum by S_σ the Chernoff bounds state that for any $\delta \in [0,1]$ we have $\Pr[S_\sigma > (1+\delta)k\sigma] \leq e^{-\delta^2 k\sigma/3}$ and $\Pr[S_\sigma < (1-\delta)k\sigma] \leq e^{-\delta^2 k\sigma/2}$.

For either distribution p_1, p_2 —we work with p_1 here for notational convenience— and any i, the probability that a sample draws i is $p_1(i)$ by definition, and all the samples are independent, so thus the Chernoff bounds state that if $S_{1,i}$ represents the random variable that counts the number of times i occurs in k samples then $\Pr[|\frac{S_{1,i}}{k} - p_1(i)| > \delta p_1(i)] \leq 2 \cdot e^{-\delta^2 k p_1(i)/2}$. Given k samples from each distribution we will show how, with probability $\frac{2}{3}$ we can construct p_1^*, p_2^* that satisfy all the desired conditions except the condition that their sums be 1; as a final step we correct their sums. The construction of p_1^*, p_2^* will be done separately for each i. Explicitly, for each i we will divide the possible outcomes into two classes: a success case where the number of times i is drawn from each distribution closely approximates $kp_1(i)$ and $kp_2(i)$ respectively and our approximation contributes at most e_i error, and a failure case which occurs with probability at most f_i. The analysis consists of two cases.

Case 1: $p_1(i), p_2(i) < \frac{c}{k}$. Suppose element i appears in s_1 samples from the first distribution and s_2 samples from the second distribution. If $s_1, s_2 \leq c$ then from Step 2 of the Canonical Tester, i is constrained to be "low-frequency", which means we may let $p_1^*(i) = p_1(i)$ and $p_2^*(i) = p_2(i)$, with no error and no failure probability. Otherwise, the constraints of Step 1 dictate that $p_1^*(i) = \frac{s_1}{k}$ and $p_2^*(i) = \frac{s_2}{k}$. Letting $\delta = (2\sqrt{\log n})\sqrt{\frac{1}{kp_1(i)}}$ we have $\Pr[|\frac{s_1}{k} - p_1(i)| > 2\sqrt{\log n}\frac{\sqrt{p_1(i)}}{\sqrt{k}}] \leq 2 \cdot e^{-2\log n}$, with a similar expression holding for the second distribution. We call the case when either of these conditions is violated the failure case, which occurs with probability at most $f_i = 4 \cdot e^{-2\log^2 n}$; otherwise, $p_1^*(i)$ approximates $p_1(i)$ to within $\frac{2\sqrt{\log n}\sqrt{p_1(i)}}{\sqrt{k}} \leq \frac{2\sqrt{\log n}\sqrt{c}}{k}$, and by a symmetric argument $p_2^*(i)$ approximates $p_2(i)$ to this bound too. Let $e(i) = \frac{4\sqrt{\log n}\sqrt{c}}{k}$.

Case 2: $p_1(i) > \frac{c}{k}$ or $p_1(i) > \frac{c}{k}$. As above, let s_1, s_2 denote the number of times i is sampled from each distribution. For the sake of simplicity we let $p_1^*(i) = \frac{s_1}{k}$ and $p_2^*(i) = \frac{s_2}{k}$ regardless of whether the constraint of Step 1 or Step 2 applies. Applying Chernoff bounds as above with the same $\delta = 2\sqrt{\log n}\sqrt{\frac{1}{kp_1(i)}}$ we have $\Pr[|\frac{s_1}{k} - p_1(i)| > 2\sqrt{\log n}\frac{\sqrt{p_1(i)}}{\sqrt{k}}] \leq 2 \cdot e^{-2\log n}$, with corresponding expression for p_2. Thus we let $f_i = 4 \cdot e^{-2\log^2 n}$ and $e_i = 2\sqrt{\log n}\frac{\sqrt{p_1(i)}+\sqrt{p_2(i)}}{\sqrt{k}}$.

Having analyzed these cases we now combine the errors and failure probabilities. Note that in every case the failure probability was bounded by $4 \cdot e^{-2\log n} = \frac{4}{n^2}$, so thus by the union bound the total failure probability is at most n times this, namely $\frac{4}{n}$. We now bound $\sum_i e_i$. We note that Case 1 consists of two sub-cases: if $s_1, s_2 \leq c$ then $e_i = 0$; otherwise $e_i = \frac{4\sqrt{\log n}\sqrt{c}}{k}$. Note that since the total number of samples is k, the situation $s_1 > c$ may occur at most $\frac{k}{c}$ times, and thus this subcase may occur at most $\frac{2k}{c}$ times yielding a contribution to $\sum_i e_i$ of at most $\frac{8\sqrt{\log n}}{\sqrt{c}}$. To bound the contribution from Case 2, let I denote the set of i that fall into case 2. We note that $|I| \leq \frac{2k}{c}$ since for each such i, $p_1(i) + p_2(i) \geq \frac{c}{k}$ and the total weight of p_1 and p_2 is 2. We apply Cauchy-Schwarz with one vector consisting of $\sqrt{p_1(i)}$ and $\sqrt{p_2(i)}$ for all $i \in I$, and the second vector the all-ones vector to yield $\sum_{i\in I}\sqrt{p_1(i)} + \sqrt{p_2(i)} \leq \sqrt{\sum_{i\in I} p_1(i) + p_2(i)}\sqrt{\sum_{i\in I} 1 + 1} \leq \sqrt{2}\sqrt{2|I|} \leq \sqrt{\frac{8k}{c}}$. Thus $\sum_{i\in I} e_i \leq 4\frac{\sqrt{\log n}}{\sqrt{c}}$ and in total we have $\sum_i e_i \leq 12\sqrt{\frac{\log n}{c}}$. Thus we have constructed p_1^* and p_2^* that satisfy all the desired properties with the possible exception that their sums may not be 1, and their total distance from p_1 and p_2 is at most $12\sqrt{\frac{\log n}{c}}$. We note that the total amount that these sum constraints is

violated equals $||p_1^*| - |p_1|| + ||p_2^*| - |p_2||$, which by the triangle inequality is at most $|p_1^* - p_1| + |p_2^* - p_2|$, which we just bounded as $12\sqrt{\frac{\log n}{c}}$. Thus if we define $\bar{p}_1, \bar{p}_2$ as the closest distributions to p_1^*, p_2^* that satisfy *all* the constraints, this rounding will change the distributions by at most $12\sqrt{\frac{\log n}{c}}$ in total. Thus by a final application of the triangle inequality, $|\bar{p}_1 - p_1| + |\bar{p}_2 - p_2| \leq 24\sqrt{\frac{\log n}{c}}$, as desired, and this is guaranteed as long as no failures occur, which happens with probability at least $1 - \frac{4}{n}$. □

Appendix to Section 4

The proof of the Positive-Negative Distance Lemma synthesizes three different techniques. These techniques address in turn three evident difficulties in deriving lower-bounds: (1) when taking k samples from a distribution, the number of times the first element is sampled is (anti-) correlated with the number of times the second, third, and other elements are sampled; (2) the complete record of each of the $2k$ samples is a lot of data to analyzes; and (3) we have no idea how the tester will make its decision from this data.

Poissonization In order to resolve (1), we follow [3] and apply a "Poissonization" technique. Recall the Poisson distribution, defined in Definition 8.

Given a tester T with sample complexity k on distributions p_1, p_2, we modify it in an essentially trivial way, but one which changes the analysis drastically. The following process is called *Poissonization*:

1. Draw two numbers k_1, k_2 from the Poisson process $Poi(k)$.
2. Draw k_1 samples from p_1 and k_2 samples from p_2.
3. If either $k_1 < k$ or $k_2 < k$, FAIL.
4. Otherwise, return the result of running T on the first k samples drawn from each distribution.

Lemma 10. *If T is a $(\frac{1}{3}, \frac{2}{3})$ tester, then the Poissonized T is a tester with soundness at least $\frac{1}{12}$.*

Proof. Note that when Step 3 does not fail, the procedure exactly simulates the original tester T, since the first k samples drawn from each of p_1, p_2 will be independent and identically distributed regardless of k_1, k_2.

We note the standard property of Poisson distributions that the median of $Poi(\lambda)$ is at least $\lfloor \lambda \rfloor$. Thus in our case the median of $Poi(k)$ is at least k, and thus the probability that $k_1 \geq k$ is at least $\frac{1}{2}$, and the probability that both $k_1, k_2 \geq k$ is at least $\frac{1}{4}$. Thus Step 3 fails with probability at most $\frac{3}{4}$, and thus the resulting tester has soundness at least $\frac{1}{12}$. □

The purpose of modifying T in this manner is revealed by the following fact, which is standard in balls-and-bins arguments:

Lemma 11. *The distribution of samples at Step 2 of a k-Poisson tester equals that generated by the following process:*

> *For each $i \in [n]$ draw $s_1(i) \leftarrow \mathrm{Poi}(k \cdot p_1(i))$, creating $s_1(i)$ samples "i" for the first distribution, and draw $s_2(i) \leftarrow \mathrm{Poi}(k \cdot p_2(i))$, creating $s_2(i)$ samples "i" for the second distribution.*

(For a proof of a variant of this see [17] page 100.)

Fingerprints We now make use of the fact that we test symmetric properties to greatly reduce the dimension and information content of the samples the tester must analyze. In essence, we claim that a set of samples is completely described by its collision statistics – for the purposes of symmetric property testing. For example, if p_1 and p_2 are distributions with support $\{a, b, c, d\}$ and we draw the 5 samples (a, a, a, b, c) from p_1, and the 5 samples (a, b, b, c, d) from p_2, then we could describe the samples by saying that there is a $(3, 1)$-way collision on the a's, a $(2, 1)$-way collision on the b's, a $(1, 1)$-way collision on the c's, and a $(0, 1)$-way collision on the d's. Because we consider symmetric properties, there is no real distinction between

a, b, c, or d, and thus we may simplify this description to: there is 1 $(3,1)$ collision, one $(1,2)$ collision, one $(1,1)$ collision, and one $(0,1)$ collision. We formalize this for the case where the number of samples drawn is a random variable, so that we may extend the results on Poissonization:

[Definition 9.] *Given two multisets of samples S_1, S_2 drawn from distributions with finite support set X, the* fingerprint *of S_1, S_2 is a function $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \to \mathbb{Z}^+$ such that $f(i,j)$ is number of elements of X that appear exactly i times in S_1 and j times in S_2.*

[Lemma 2.] *For any symmetric property U and random variable κ, if there exists a κ-sample tester T then there exists a κ-sample tester T' which takes as input only the fingerprint of κ samples drawn from each distribution.*

Proof. Given T and a fingerprint $f(\cdot,\cdot)$ of κ samples each from distributions p_1 and p_2 on $[n]$ we produce such a T' as follows:

1. Initialize empty lists s_1, s_2.
2. For each nonzero pair (i,j), pick $f(i,j)$ arbitrary new values in $[n]$ and append these i times to the list s_1 of "simulated samples for the first distribution", and j times to the list s_2.
3. Construct a random permutation π over $[n]$.
4. Return $T(\pi(s_1), \pi(s_2))$, namely, apply π to rename the elements of s_1, s_2, and run the original tester T on these simulated samples.

We note that the distribution of the lists we give to T is *identical* to that produced by the process of picking a random permutation γ on n elements and drawing κ samples each from the distributions $p_1 \circ \gamma$ and $p_2 \circ \gamma$. Furthermore, since T is a $(\frac{1}{3}, \frac{2}{3})$ tester, it will work for arbitrary input distributions, including $p_1 \circ \gamma$ and $p_2 \circ \gamma$ for any fixed γ. Thus T will also operate correctly when γ is drawn randomly, which implies that T' is a tester for U, as desired. □

For the rest of this paper when we refer to a tester we will generally consider its input to be in fingerprint form.

One of the principal objects of analysis of this paper is the fingerprint function applied to Poisson-distributed samples in the sense of the previous subsection. We define: **[Definition 10.]** *Given distributions p_1, p_2 with support $[n]$ and a positive integer k, define $D^k_{p_1,p_2}$ to be the distribution of fingerprints of the following sampling process:*

1. *Draw two numbers k_1, k_2 from the Poisson process $Poi(k)$.*
2. *Draw k_1 samples from p_1 and k_2 samples from p_2.*

We note that we may apply Lemma 11 to reexpress this distribution in a form that is often more easy to work with.

Lemma 12. *For any distributions p_1, p_2 with support $[n]$ and positive integer k, $D^k_{p_1,p_2}$ is identical to the following:*

1. *Initialize a fingerprint as the function mapping pairs of nonnegative integers to 0.*
2. *For each $i \in [n]$ draw $a \leftarrow Poi(k \cdot p_1(i))$ and $b \leftarrow Poi(k \cdot p_2(i))$ and increment the fingerprint's value at (a,b) by 1.*

The Statistical Distance Testing Bound. The third standard technique we apply consists of the following observation: if T is a k-sample tester for property U with soundness c then the *view* of T on any positive distribution pair p_1^+, p_2^+ and the view of T on any negative distribution p_1^-, p_2^- must have statistical distance at least c. Explicitly, the view of a tester consists of its input, which by the results of the previous section is a fingerprint. So we conclude that the statistical distance between the distributions of k-sample-fingerprints for any positive and any negative distribution pairs for U is at least c. For the case of Poisson testers we have

Lemma 13. *If π is a symmetric property testable by a k-Poisson tester with soundness γ then for any positive distribution pair (p_1^+, p_2^+) and any negative distribution pair (p_1^-, p_2^-), we have $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \geq \gamma$.*

Combining this with Lemma 10 yields:

[**The Positive-Negative Distance Lemma.**] *If π is a symmetric property testable in k samples then for any positive distribution pair p_1^+, p_2^+ and any negative distribution pair p_1^-, p_2^-, we have $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \geq \frac{1}{12}$.*

Appendix to Section 5

Proof of Lemma 5. For any $i \in [n]$ and fingerprint index (a,b) (note that this is a *single* index into λ, or columns of ρ) we have that

$$\rho_{i,(a,b)} = \text{poi}(a; k \cdot p_1(i))\text{poi}(b; k \cdot p_2(i)) = \frac{e^{-k(p_1(i)+p_2(i))}(k \cdot p_1(i))^a (k \cdot p_2(i))^b}{a!b!} \leq (k \cdot p_1(i))^a (k \cdot p_2(i))^b \leq \epsilon^{a+b}.$$

Thus for each (a,b) we have

$$\frac{\sum_i \rho(i,(a,b))^2}{\lambda((a,b))} \leq \frac{\sum_i \epsilon^{a+b}\rho(i,(a,b))}{\lambda((a,b))} = \epsilon^{a+b}\frac{\sum_i \rho(i,(a,b))}{\sum_i \rho(i,(a,b))} = \epsilon^{a+b}.$$

Thus by Roos's Theorem and Lemma 4 we have

$$|D^k_{p_1,p_2} - \text{Poi}(\lambda)| = |M^\rho - \text{Poi}(\lambda)| \leq 8.8 \sum_{(a,b)} \frac{\sum_i \rho(i,(a,b))^2}{\lambda(d)} \leq 8.8 \sum_{a+b\geq 1} \epsilon^{a+b}.$$

We recognize this last expression as a two-dimensional geometric series, which we can evaluate via the variable substitution $\gamma = a + b$ as

$$8.8 \sum_{\gamma \geq 1} (1+\gamma)\epsilon^\gamma \leq 8.8 \left[2\epsilon + \epsilon^2 \sum_{\gamma \geq 0} 3(1+\gamma)\epsilon^\gamma \right] = 8.8 \left[2\epsilon + \frac{3\epsilon^2}{(1-\epsilon)^2} \right] \leq 20\epsilon,$$

where the last inequality applies when $\epsilon \leq \frac{1}{30}$ since in this case $3\epsilon^2 \leq \frac{\epsilon}{10}$, and $(1-\epsilon)^2 \geq \frac{1}{2}$ so $\frac{3\epsilon^2}{(1-\epsilon)^2} \leq \frac{\epsilon}{5}$, implying the desired result. □

For the sake of completeness, we derive a bound for the statistical distance between two multivariate Poisson processes. (We are unaware of a similar derivation in the literature, but surely one exists. Please let us know if there is a suitable reference!)

We present the univariate case first.

Lemma 14. *The statistical distance between two univariate Poisson distributions with parameters λ, λ' is bounded as*

$$|\text{Poi}(\lambda) - \text{Poi}(\lambda')| \leq 2\frac{|\lambda - \lambda'|}{\sqrt{1 + \max\{\lambda, \lambda'\}}}.$$

Proof. Without loss of generality, assume $\lambda \leq \lambda'$. We have two cases.

Case 1: $\lambda' \geq 1$ We estimate the distance via the *relative entropy* of $\text{Poi}(\lambda)$ and $\text{Poi}(\lambda')$, defined for general distributions p, p' as

$$D(p||p') = \sum_i p(i) \log_e \frac{p(i)}{p'(i)}.$$

We compute the relative entropy of the Poisson processes as

$$D(\text{Poi}(\lambda)||\text{Poi}(\lambda')) = \sum_{c\geq 0} \text{poi}(c;\lambda) \log_e \frac{e^{-\lambda}\lambda^c}{e^{-\lambda'}\lambda'^c} = \sum_{c \geq 0} \text{poi}(c;\lambda) \left[\lambda' - \lambda + c \log_e \frac{\lambda}{\lambda'} \right] = \lambda' - \lambda + \lambda \log_e \frac{\lambda}{\lambda'},$$

where the last equality is because the Poisson distribution of parameter λ has total weight 1 and expected value λ. Further, since $\log_e x \leq x - 1$ for all x we have

$$\lambda' - \lambda + \lambda \log_e \frac{\lambda}{\lambda'} \leq \lambda' - \lambda + \lambda \log_e \frac{\lambda}{\lambda'} - \lambda(\log_e \frac{\lambda}{\lambda'} - \frac{\lambda}{\lambda'} + 1) = \frac{(\lambda' - \lambda)^2}{\lambda'}.$$

Thus $D(\text{Poi}(\lambda)||\text{Poi}(\lambda')) \leq \frac{(\lambda'-\lambda)^2}{\lambda'}$. We recall that statistical distance is related to the relative entropy as $|p - p'| \leq \sqrt{2D(p||p')}$ (see [11] p. 300), and thus we have $|\text{Poi}(\lambda) - \text{Poi}(\lambda')| \leq \frac{\sqrt{2}|\lambda-\lambda'|}{\sqrt{\lambda'}}$. Since $\lambda' \geq \frac{1}{2}(1 + \lambda')$ for $\lambda' \geq 1$ we conclude $|\text{Poi}(\lambda) - \text{Poi}(\lambda')| \leq 2\frac{|\lambda-\lambda'|}{\sqrt{1+\lambda'}}$, as desired.

Case 2: $\lambda' < 1$ We note that for $i \geq 1$ we have $\text{poi}(0; \lambda) - \text{poi}(0; \lambda') = e^{-\lambda} - e^{\lambda'} \leq \lambda' - \lambda$ where the last inequality is because the function e^x has derivative at most 1 for $x \in [\lambda, \lambda']$, since $\lambda \leq \lambda' < 1$. Further, we note that $\text{poi}(i; \lambda) - \text{poi}(i; \lambda') = \frac{1}{i!}[e^{-\lambda}\lambda^i - e^{\lambda'}\lambda^i < 0$ where the last inequality is because the function $f(x) = e^{-x}x^i$ has derivative $e^{-x}x^{i-1}(i - x)$ which is positive for $x \in [\lambda, \lambda']$ since both are less than 1. Since both Poisson processes have total weight 1, the negative difference between the $i \geq 1$ terms exactly balances the positive difference between the $i = 0$ terms, and thus the statistical difference equals this difference, which we bounded as $\lambda' - \lambda$.

Thus, $|\text{Poi}(\lambda) - \text{Poi}(\lambda')| \leq \lambda' - \lambda < 2\frac{|\lambda-\lambda'|}{\sqrt{1+\lambda'}}$ as desired, and we have proven the lemma for both cases. □

We generalize this lemma to the multivariate case by means of the following:

Lemma 15. *Statistical distance is subadditive on independent distributions: given multivariate distributions $p(i, j) = p_1(i) \cdot p_2(j)$ and $p'(i, j) = p'_1(i) \cdot p'_2(j)$ we have $|p - p'| \leq |p_1 - p'_1| + |p_2 - p'_2|$.*

Proof. We have

$$\begin{aligned}|p - p'| &= \sum_{i,j} |p_1(i)p_2(j) - p'_1(i)p'_2(j)| \\ &\leq \sum_{i,j} |p_1(i)p_2(j) - p'_1(i)p_2(j)| + \sum_{i,j} |p'_1(i)p_2(j) - p'_1(i)p'_2(j)| = |p_1 - p'_1| + |p_2 - p'_2|.\end{aligned}$$

□

Lemma 16. *The statistical distance between two multivariate Poisson distributions with parameters λ, λ' is bounded as*

$$|\text{Poi}(\lambda) - \text{Poi}(\lambda')| \leq 2 \sum_d \frac{|\lambda_d - \lambda'_d|}{\sqrt{1 + \max\{\lambda_d, \lambda'_d\}}}.$$

Proof. Immediate from repeated application of Lemmas 14 and 15. □

Combining Lemmas 5 and 16 almost achieves the Wishful Thinking Theorem.

Lemma 17. *Given two probability distribution pairs $p_1^+, p_2^+, p_1^-, p_2^-$, a positive number ϵ and integer k such that each frequency is at most $\frac{\epsilon}{k}$, let ρ^+, ρ^- be the matrices such that $D^k_{p_1^+,p_2^+} = M^{\rho^+}$ and $D^k_{p_1^-,p_2^-} = M^{\rho^-}$, and let $\lambda_d^+ = \sum_i \rho^+(i, d)$, $\lambda_d^- = \sum_i \rho^-(i, d)$. Then*

$$|D^k_{p_1,p_2} - D^k_{p_1,p_2}| \leq 40\epsilon + 2 \sum_d \frac{|\lambda^+(d) - \lambda^-(d)|}{\sqrt{1 + \max\{\lambda^+(d), \lambda^-(d)\}}}. \tag{1}$$

Proof. Immediate: if $\epsilon < \frac{1}{30}$ then apply Lemma 5 to both $D^k_{p_1^+,p_2^+}$ and $D^k_{p_1^-,p_2^-}$ to approximate them by Poisson processes $Poi(\lambda^+)$ and $Poi(\lambda^-)$ respectively, bound $|Poi(\lambda^+) - Poi(\lambda^-)|$ via Lemma 5, and combine these bounds via the triangle inequality; otherwise $40\epsilon > 1 \geq |D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}|$ trivially. □

To prove the Wishful Thinking Theorem it remains to reexpress this bound in terms of the moments m instead of the Poisson coefficients λ.

Proof of the Wishful Thinking Theorem. We note that if $\epsilon \geq \frac{1}{40}$ the theorem is trivially true. In what follows we assume the converse.

We start from Equation 1, expanding both the numerator and denominator of the last term via Taylor series expansions. Recall the definition $\lambda_{a,b} = \frac{1}{a!b!}\sum_i k^{a+b}e^{-k(p_1(i)+p_2(i))}p_1(i)^a p_2(i)^b$. For the numerator of the term $d = (a, b)$ we have from Taylor expansions and the triangle inequality that

$$\begin{aligned}
|\lambda_{a,b} - \lambda'_{a,b}| &= \frac{k^{(a+b)/2}}{\sqrt{a!b!}}\left|\sum_i \left[e^{-k(p_1(i)+p_2(i))/2}p_1(i)^a p_2(i)^b - e^{-k(p'_1(i)+p'_2(i))}p'_1(i)^a p'_2(i)^b\right]\right| \\
&= \frac{1}{a!b!}\left|\sum_j\sum_{\gamma,\delta}\frac{(-1)^{\gamma+\delta}}{\gamma!\delta!}k^{a+b+\gamma+\delta}\left[p_1(i)^{a+\gamma}p_2(i)^{b+\gamma} - p'_1(i)^{a+\gamma}p'_2(i)^{b+\gamma}\right]\right| \\
&= \frac{1}{a!b!}\left|\sum_{\gamma,\delta}\frac{(-1)^{\gamma+\delta}}{\gamma!\delta!}[m^+_{a+\gamma,b+\delta} - m^-_{a+\gamma,b+\delta}]\right| \\
&\leq \frac{1}{a!b!}\sum_{\gamma,\delta}\frac{1}{\gamma!\delta!}\left|m^+_{a+\gamma,b+\delta} - m^-_{a+\gamma,b+\delta}\right|.
\end{aligned}$$

We now bound terms in the denominator of Equation 1. Note that in the definition of $\lambda_{a,b}$ and the fact that $p_1(i), p_2(i) \leq \frac{1}{10k}$ we have $e^{-k(p_1(i)+p_2(i))} \geq e^{-\frac{2}{10}} > .9^2$. Thus $\lambda_{a,b} \geq \frac{.9^2}{a!b!}m^+_{a,b}$ by definition of m, with corresponding expression holding for λ' and m^-. Thus we bound terms in the denominator of Equation 1 as

$$\sqrt{1 + \max\{\lambda_{a,b}, \lambda'_{a,b}\}} \geq \frac{.9}{\sqrt{a!b!}}\sqrt{1 + \max\{m^+_{a,b}, m^-_{a,b}\}}.$$

Combining the bounds for the numerator and denominator, noting that (since $p_1(i), p_2(i) \leq \frac{1}{k}$) both m^+ and m^- are decreasing functions of a and b, and making the variable substitutions $\mu = a + \gamma$, and $\nu = b + \delta$ yields

$$\begin{aligned}
\sum_{a,b}\frac{|\lambda_{a,b} - \lambda'_{a,b}|}{\sqrt{1 + \max\{\lambda_{a,b}, \lambda'_{a,b}\}}} &\leq \sum_{a,b}\sum_{\gamma,\delta}\frac{\left|m^+_{a+\gamma,b+\delta} - m^-_{a+\gamma,b+\delta}\right|}{.9\gamma!\delta!\sqrt{a!b!}\sqrt{1 + \max\{m^+_{a,b}, m^-_{a,b}\}}} \\
&\leq \sum_{a,b}\sum_{\gamma,\delta}\frac{\left|m^+_{a+\gamma,b+\delta} - m^-_{a+\gamma,b+\delta}\right|}{.9\gamma!\delta!\sqrt{a!b!}\sqrt{1 + \max\{m^+_{a+\gamma,b+\delta}, m^-_{a+\gamma,b+\delta}\}}} \\
&= \sum_{\mu,\nu}\sum_{\substack{\gamma\leq\mu\\ \delta\leq\nu}}\frac{\left|m^+_{\mu,\nu} - m^-_{\mu,\nu}\right|}{.9\gamma!\delta!\sqrt{(\mu-\gamma)!(\nu-\delta)!}\sqrt{1 + \max\{m^+_{\mu,\nu}, m^-_{\mu,\nu}\}}} \\
&= \sum_{\mu,\nu}\frac{\left|m^+_{\mu,\nu} - m^-_{\mu,\nu}\right|}{\sqrt{1 + \max\{m^+_{\mu,\nu}, m^-_{\mu,\nu}\}}}\frac{1}{.9}\left(\sum_{\gamma\leq\mu}\frac{1}{\gamma!\sqrt{(\mu-\gamma)!}}\right)\left(\sum_{\delta\leq\nu}\frac{1}{\delta!\sqrt{(\nu-\delta)!}}\right).
\end{aligned}$$

We bound the expression $\sum_{\gamma\leq\mu}\frac{1}{\gamma!\sqrt{(\mu-\gamma)!}}$ as follows: note that the sum of the squares of the terms is bounded as $\sum_{\gamma\leq\mu}\frac{1}{\gamma!^2(\mu-\gamma)!} \leq \frac{\mu!}{\mu!}\sum_{\gamma\leq\mu}\frac{2}{2^\gamma\gamma!(\mu-\gamma)!} = 2\frac{1.5^\mu}{\mu!}$ by the binomial theorem. Having bounding the sum of the squares of the terms, Cauchy-Schwarz bounds the original sum as $\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}}$. We note that $\mu! \geq \lfloor\frac{\mu}{2}\rfloor!^2 \cdot 2^{n/2}$, as can be seen by comparing the elements of each factorial term-by-term. Thus

$\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}} \le \frac{1}{\lfloor \frac{\mu}{2} \rfloor !}$ for large enough μ; evaluating for small μ we see that in fact $\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}} \le \frac{3}{\lfloor \frac{\mu}{2} \rfloor !}$ for all μ, which is our bound on the γ sum; consequently the "$\delta \le \nu$" sum is bounded by $\frac{3}{\lfloor \frac{\nu}{2} \rfloor !}$, and since $\frac{1}{.9} 3 \cdot 3 = 10$ we have the theorem. □

We will find it convenient to work with a finite subset of the moments in the Section 6, 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 probability distribution pairs p_1^+, p_2^+ and p_1^-, p_2^- and positive number $\epsilon \le \frac{1}{10 \cdot 2^{\sqrt{\log n}}}$ and integer k such that each probability of any element in each distribution is bounded by $\frac{\epsilon}{k}$, define the (a,b)-order moments as $m_{a,b}^+ = k^{a+b} \sum_i p_1^+(i)^a \cdot p_2^+(i)^b$ and $m_{a,b}^- = k^{a+b} \sum_i p_1^-(i)^a \cdot p_2^-(i)^b$. Then*

$$|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \le .04 + 40\epsilon + 10 \sum_{2 \le a+b \le \sqrt{\log n}} \frac{|m_{a,b}^+ - m_{a,b}^-|}{\lfloor \frac{a}{2} \rfloor ! \lfloor \frac{b}{2} \rfloor ! \sqrt{1 + \max\{m_{a,b}^+, m_{a,b}^-\}}}.$$

Proof. We show that this follows from the bound of the Wishful Thinking Theorem. We note that for any distributions $p_1^+, p_2^+, p_1^-, p_2^-$, we have $m_{0,0}^+ = m_{0,0}^- = n$, and $m_{1,0}^+ = m_{0,1}^+ = m_{1,0}^- = m_{0,1}^- = k$, so thus the terms for $a+b < 2$ vanish. To bound the terms for $a+b > \max\{2, \sqrt{\log n}\}$ we note that for such an a, b we have $m_{a,b}^+ \le k^{a+b} n (\frac{\epsilon}{k})^{a+b} = n\epsilon^{a+b} \le .1^{a+b}$ Thus, since $\frac{|m_{a,b}^+ - m_{a,b}^-|}{\lfloor \frac{a}{2} \rfloor ! \lfloor \frac{b}{2} \rfloor ! \sqrt{1+\max\{m_{a,b}^+, m_{a,b}^-\}}} \le |m_{a,b}^+ - m_{a,b}^-|$, we can bound the contribution of the (a,b) term by $.1^{a+b}$, and the sum of these terms by $\sum_{a+b \ge 2} .1^{a+b} < .04$, yielding the corollary. □

Proof of Theorem 1. Let x, y be distributions on $[n]$ defined as follows: for $1 \le i \le n^{2/3}$ let $x_i = y_i = \frac{1}{2n^{2/3}}$; we refer to these elements as the "large" elements. For $n/2 < i \le 3/4n$ let $x_i = \frac{2}{n}$; and for $3n/4 < i \le n$ let $y_i = \frac{2}{n}$; we refer to these indices as the "small" elements. The remaining elements of x and y are zero.

Let $p_1^+ = p_2^+ = p_1^- = x$, and $p_2^- = y$ and $k = \frac{n^{2/3}}{800}$. We note that each frequency defined is at most $\frac{1}{1600k}$, and we let $\epsilon = \frac{1}{1600}$. Let $m_{a,b}^+$ and $m_{a,b}^-$ be defined as in the Wishful Thinking Theorem. 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

$$\sum_{a,b} \frac{|m_{a,b}^+ - m_{a,b}^-|}{\sqrt{1 + \max\{m_{a,b}^+, m_{a,b}^-\}}} \le \sum_{a,b \ge 1} \frac{\frac{n}{4(400n^{1/3})^{a+b}}}{\sqrt{\frac{n^{2/3}}{1600^{a+b}}}} = \sum_{a,b \ge 1} \frac{n^{2/3}}{4(10n^{1/3})^{a+b}} = \frac{1}{400} \sum_{a,b} \frac{1}{(10n^{1/3})^{a+b}}$$
$$\le \frac{1}{400} \sum_{a,b} \frac{1}{10^{a+b}} < .003.$$

Thus the Wishful Thinking Theorem yields $|D^k_{p_1^+,p_2^+} - D^k_{p_1^-,p_2^-}| \le 40\epsilon + 10 \cdot .003 = .055 < \frac{1}{12}$. From the Positive-Negative Distance Lemma we conclude that no tester can distinguish (p_1^+, p_2^+) from (p_1^-, p_2^-) in k samples, as desired. □

Appendix to Section 6

Proving the Matching Moments Theorem: Because matrix inversion commutes with the tensor product, to bound $inv(L^{\sqrt{\log n}})$ we need only bound elements of $inv(\ell^{\sqrt{\log n}})$ and square the answer. We make use of a standard (if slightly unwieldy) formula to compute the inverse of Vandermonde matrices:

Lemma 18 (From [14]). *The inverse of the Vandermonde matrix generated by z has (i,j)th entry*

$$\frac{(-1)^{i+1}\sum_{\substack{1\le s_1<s_2<\ldots<s_{\mu-i}\le\mu\\ \forall q, s_q\neq j}}\prod_{q=1}^{\mu-i} z_{s_q}}{\prod_{q\in\{1,\ldots,\mu\}-\{j\}}(z_q-z_j)}. \tag{2}$$

We apply this lemma to bound the inverse of ℓ^μ and from there, L^μ.

Lemma 19. *Each element of $inv(\ell^\mu)$ has magnitude at most $(2e)^\mu$.*

Proof. We bound the magnitudes of the numerator and denominator of Equation 2 when $z = \{1,\ldots,\mu\}$. Note that the magnitude of the denominator equals $(j-1)!(\mu-j)!$. We bound this using Stirling's approximation to the factorial function, $n! \ge S(n) \triangleq \sqrt{2\pi n}\frac{n^n}{e^n}$, which we note has convex logarithm. Thus

$$(j-1)!(\mu-j)! \ge \frac{1}{\mu}j!(\mu-j)! \ge \frac{1}{\mu}S(j)S(\mu-j) \ge \frac{1}{\mu}S(\frac{\mu}{2})^2 = \pi\frac{\mu^\mu}{(2e)^\mu} \ge \frac{\mu^\mu}{(2e)^\mu},$$

where the third inequality is Jensen's inequality, applied to the logarithm of S.

The sum in the numerator has at most $\binom{\mu}{\mu-i} = \binom{\mu}{i} \le \mu^i$ terms, where the summand is a product bounded by $\mu^{\mu-i}$, so the numerator has magnitude at most μ^μ. Comparing our bounds on the numerator and denominator yields the lemma. □

Lemma 20. *Each element of $inv(L^\mu)$ has magnitude at most 30^μ.*

Proof. Since matrix inversion and tensor products commute, $inv(L^\mu) = inv(\ell^\mu) \otimes inv(\ell^\mu)$, immediately yielding this lemma as a corollary of Lemma 19, since $(2\mu)^2 < 30$. □

Definition 18. *Define the function M mapping distribution pairs p_1, p_2 on n elements and real numbers $0 < w \le 1$ and $k < n$ to distribution pairs $(\bar{p}_1, \bar{p}_2) \leftarrow M_w^k(p_1, p_2)$ via the following sequence of modifications to p_1, p_2*

1. *Let $w' = \frac{w}{6}$; let I be the set of $\lfloor w'n \rfloor$ indices i such that $p_1(i) + p_2(i)$ is smallest. Change p_1, p_2 to the nearest distribution pair $\bar{p}_1, \bar{p}_2$ such that $\forall i \in I, \bar{p}_1(i) = \bar{p}_2(i) = 0$, $\forall i \notin I, \bar{p}_1(i), \bar{p}_2(i) \in [0, \frac{1}{k}]$, and $\sum_i \bar{p}_1(i) = \sum_i \bar{p}_2(i) = 1 - w'$.*
2. *Let $\mu = 1 + \lfloor\sqrt{\log n}\rfloor$, and let $\kappa = \frac{kw'}{\mu^2 30^\mu}$; for integers $0 \le a, b \le \mu - 1$ let $m_{a,b} = \sum_i \bar{p}_1(i)^a \cdot \bar{p}_2(i)^b \cdot \kappa^{a+b}$ be the κ-based moments of this modified vector, with $m_{0,0} = m_{1,0} = m_{0,1} = 0$ being defined separately. Let $c = inv(L^\mu) \cdot m$.*
3. *Let $\bar{m}_{a,b}$ be an upper-bound on m which has value 0 when $a + b < 2$ and value $\frac{\kappa^2}{k}$ otherwise. Let $\bar{L}^\mu$ be an element-by-element upper-bound on the magnitudes of the elements in $inv(L^\mu)$: a matrix of the same size as L^μ with entries 30^μ, and let $\bar{c} = \bar{L}^\mu \cdot \bar{m}$.*
4. *For each $0 \le \gamma, \delta < \mu$ choose $\lfloor \bar{c}(\gamma,\delta) - c(\gamma,\delta) \rfloor$ indices $i \in I$ with zero entries and let $\bar{p}_1(i) = \frac{\gamma}{\kappa}, \bar{p}_2(i) = \frac{\delta}{\kappa}$ for these indices.*
5. *Make $\sum \bar{p}_1(i) = 1$ by filling in the unassigned entries $\bar{p}_1(i)$ for $i \in I$ with identical constants, and perform the corresponding operation on $\bar{p}_2$.*

Define $\widetilde{m} \triangleq f(w,k)$ to be the (k-based) moments of the result of applying the above procedure to a pair of uniform distributions.

Proof of the Matching Moments Theorem. We show that M and f from Definition 18 satisfy the conditions of the theorem. We examine each stage of Definition 18 in turn, showing that certain invariants are satisfied.

1. Consider the set of indices I defined in Stage 1. We note that since $\sum_i p_1(i) + p_2(i) = 2$, the $\lfloor w'n \rfloor$ smallest indices, I must have total frequency at most $2\frac{\lfloor w'n \rfloor}{n}$. We note that if we ignore the condition that $\bar{p}_1(i), \bar{p}_2(i) \in [0, \frac{1}{k}]$, then clearly the conditions have a solution where $|p_1 - \bar{p}_1| + |p_2 - \bar{p}_2| \leq 4w'$ since zeroing out the entries $i \in I$ changes the distributions by at most $2w'$ (from above), each distribution now has total weight in $[1 - 2w', 1]$, so changing each distribution from here so as to have total weight $1 - w'$ changes each distribution by at most $\bar{w}$. After this process, the average frequency of each index not in I is at most $\frac{1}{n}$ in either distribution. Since $\frac{1}{n} \leq \frac{1}{k}$ we can find a solution with all the frequencies at most $\frac{1}{k}$ that is as close to p_1, p_2 as the above by shifting weight from those elements where $\bar{p}_*(i) > \frac{1}{n}$ to elements where $\bar{p}_*(i) < \frac{1}{n}$.

2+3. We note that $\bar{m}$ is indeed an upper-bound on m: the $a + b < 2$ cases are by definition; otherwise, without loss of generality assume $a \geq 1$, in which case $m_{a,b} \leq \sum_i \bar{p}_1(i)(\frac{1}{k})^{a+b-1} \cdot \kappa^{a+b} \leq \frac{\kappa^2}{k} \sum_i \bar{p}_1(i) \leq \frac{\kappa^2}{k}$, as desired. The fact that $\bar{L}^\mu$ bounds the magnitudes of the elements of $inv(L^\mu)$ is Lemma 20. Since $\bar{m}$ and $\bar{L}^\mu$ respectively bound the magnitudes of $inv(L^\mu)$ and m, their product $\bar{c}$ bounds the magnitudes of c.

4. Since $\bar{c}$ upper-bounds c, each of the expressions $\lfloor \bar{c}(\gamma, \delta) - c(\gamma, \delta) \rfloor$ is non-negative. To show that the new entries "fit in I", we bound the total frequency contribution of the new elements in the first distribution. Note that this equals $\frac{1}{\kappa}$ times the κ-based $(1, 0)$ moment of the portion of the distribution in I, which we bound via Lemma 7 as the $(1, 0)$ entry of $\frac{1}{\kappa} L^\mu \cdot \lfloor \bar{c} - c \rfloor \leq \frac{1}{\kappa}[L^\mu \cdot (\bar{c} - c)] = \frac{1}{\kappa} L^\mu \cdot \bar{c}$, all of whose entries equal $(\mu^2 - 3)30^\mu \frac{\kappa}{k} = \frac{(\mu^2-3)w'}{\mu^2} \leq w'$. Similarly, the total number of entries added to either distribution is at most $\sum_{\gamma,\delta} \bar{c}(\gamma, \delta) - c(\gamma, \delta)$, which is the $(0, 0)$ entry of $L^\mu \cdot (\bar{c} - c) = L^\mu \cdot \bar{c}$, all of whose entries are $(\mu^2 - 3)30^\mu \frac{\kappa^2}{k} \leq \frac{w'n}{30} \leq \lfloor w'n \rfloor = |I|$, so thus the new entries "fit".

We estimate the moments of the distribution again via another application of Lemma 7: note that (for $a + b \geq 2$) the moments of the portion of the distribution outside I are described by m, while the moments for the portion in I are described by Lemma 7 as $L^\mu \cdot \lfloor \bar{c} - c \rfloor$. Letting $L^\mu_{(a,b)}$ denote the (a, b) row of L^μ, we note that $L^\mu \cdot \lfloor \bar{c} - c \rfloor$ is at most $|L^\mu_{(a,b)}|$ less than $L^\mu \cdot (\bar{c} - c) = L^\mu \cdot \bar{c} - m$. Thus the κ-based moments for the entirety of $\bar{p}_1, \bar{p}_2$ are between $L^\mu \cdot \bar{c}$ and $|L^\mu_{(a,b)}|$ less than this (for $a + b \geq 2$).

5. We bound the change to the moments induced by the fifth step. We note that the sum of the row $(0, 0)$ of L^μ equals μ^2, and the sum of either the $(1, 0)$ or $(0, 1)$ rows of L^μ equals $\frac{\mu^3(\mu+1)}{2} \leq \mu^4$. Let $x = \frac{(\mu^2-3)w'}{\mu^2}$. A tighter analysis of the bounds found in the previous step yields that the number of entries added in Step 4 is between the $(0, 0)$ entries of $L^\mu \cdot (\bar{c} - c - 1)$ and $L^\mu \cdot (\bar{c} - c)$, namely in the range $[x\kappa - \mu^2, x\kappa]$, and the total weight added to $\bar{p}_1$ in Step 4 is similarly between $\frac{1}{\kappa}$ times the $(1, 0)$ entries of $L^\mu \cdot (\bar{c} - c - 1)$ and $L^\mu \cdot (\bar{c} - c)$, namely in the range $[x - \frac{\mu^4}{\kappa}, x]$. Thus the total number of entries of I unallocated until Step 5 is in the range $[\lfloor w'n \rfloor - x\kappa, \lfloor w'n \rfloor - x\kappa + \mu^2]$ and the amount of weight added to the first distribution in Step 5 is in the range $[\frac{3}{\mu^2}w', \frac{3}{\mu^2}w' + \frac{\mu^4}{\kappa}]$. We note that, if we represent these last two intervals as $[y, y']$ and $[z, z']$ respectively, then the κ-based (a, b) moment will be between $\frac{z^{a+b}}{y'^{a+b-1}}\kappa^{a+b}$ and $\frac{z'^{a+b}}{y^{a+b-1}}\kappa^{a+b}$, whose ratio is bounded as $(z/z')^{a+b}(y/y')^{a+b-1} \geq 1 - (a + b)[\frac{z'-z}{z'} + \frac{y'-y}{y'}]$. Here we have $y' = \lfloor w'n \rfloor - x\kappa \geq w'n - 1 - \frac{w'n}{30} \geq \frac{w'n}{2}$ and $z' = \frac{3}{\mu^2}w'$ we have that the ratio between the minimum and maximum contributions of Step 5 to this moment is at least $1-(a+b)[\frac{2\mu^2}{w'n} + \frac{\mu^6}{3w'\kappa}] \geq 1-(a+b)\frac{\mu^6}{w'\kappa}$. We note that since $2 \leq a+b \leq \mu$, the maximum contribution is at most $\frac{z'^{a+b}}{y^{a+b-1}}\kappa^{a+b} \leq \frac{(4w'/\mu^2)^{a+b}}{(w'n/2)^{a+b-1}}\kappa^{a+b} \leq \frac{8^{a+b}w'\kappa}{\mu^4}(\frac{\kappa}{n})^{a+b-1} \leq \frac{8^\mu w'\kappa}{\mu^4}\frac{1}{\mu^2 30^\mu}$, and thus the *difference* between the maximum and minimum contributions is at most $(a + b)\frac{\mu^6}{w'\kappa}\frac{8^\mu w'\kappa}{\mu^4}\frac{1}{\mu^2 30^\mu} \leq \frac{\mu^7}{w'\kappa}\frac{8^\mu w'\kappa}{\mu^6 8^\mu 2^\mu} \leq 1$.

Thus, for any fixed a, b such that $2 \leq a + b < \mu$ the difference between the maximum and minimum κ-based moments reached by this construction, from any starting distribution, is at most $1 + |L^\mu_{(a,b)}|$ (recall that for $a + b < 2$ the moments are invariant). Recall that the elements of the (a, b) row of L^μ are values $t^a \cdot u^b$ for $1 \leq t, u \leq \mu$, so thus there are μ^2 integer elements, all at most μ^{a+b} and some strictly less, so $1 + |L^\mu_{(a,b)}| \leq \mu^{a+b+2}$.

To convert this to a bound on the $\bar{k}$-based moments we multiply by $(\frac{\bar{k}}{\kappa})^{a+b}$. We have $\frac{\bar{k}}{\kappa} = \frac{kw\mu^2 30^\mu}{kw' \cdot 150 \cdot 2^{6\sqrt{\log n}}} \geq$

$\frac{\mu^2 30^\mu}{15 \cdot 64^\mu} \geq \frac{1}{100\mu^2}$, where the last equality follows by inspection for small integer values of μ. Thus the bound on the variation of the $\bar{k}$-based moments is $\mu^{a+b+2}(\frac{1}{100\mu^2})^{a+b} \leq \frac{1}{10000\mu^2}$ for $a+b \geq 2$, and 0 for $a+b < 2$, as desired.

We note that the other two results of the theorem follow easily from the construction: by construction, the maximum frequency in $\bar{p}_1, \bar{p}_2$ is at most $\frac{\mu}{\kappa} \leq \frac{1}{k}$, as desired. And we note that we showed that Step 1 modifies the distributions by at most $4w'$; since at the end of Step 1 the distributions have total weight $2 - 2w'$ and we only increase frequencies in the remaining steps, the total change is at most $6w' = w$, as desired. □

Appendix to Section 7

Proof of the Low Frequency Blindness Theorem. Assume for the sake of contradiction that there were a tester T that distinguishes between $\pi < a + \epsilon$ and $\pi > b - \epsilon$ in $\frac{k\delta}{100000 \cdot 2^{6\sqrt{\log n}}}$ samples.

Let H be the set of k-high-frequency indices of (p_1^-, p_2^-) (which are identical to those of (p_1^+, p_2^+) by definition), and let $L = [n] - H$. Let $\ell_1 = |p_1(L)|$, namely the probability that p_1 draws a low-frequency index, and let $\ell_2 = |p_2(L)|$. Consider the following property π' on distributions (p_1^L, p_2^L) with support L: construct the distribution p_1 such that $p_1(L) = \ell_1 p_1^L$ and $p_1(H) = p_1^-(H)$, with p_2 defined analogously, and return $\pi(p_1, p_2)$.

We construct a tester for π' based on T. From Lemma 10 we conclude that the Poissonized version of T is a tester with soundness $\frac{1}{12}$. We note that by Lemma 11, when the Poissonized T is applied to distributions (p_1, p_2) constructed as above, the distribution of samples from p_1 that lie in L is identical to that of drawing $t \leftarrow \text{Poi}(k \cdot \ell_1)$ and drawing t independent samples from p_1^L, with corresponding statements holding for p_2 and the H indices. Thus letting $\ell = \max\{\ell_1, \ell_2\}$ we may define a $k\ell$-Poissonized tester T' for π' as follows, assuming without loss of generality that $\ell_1 \geq \ell_2$:

- Draw integers $t_1^H \leftarrow \text{Poi}(k(1-\ell_1)), t_2^H \leftarrow \text{Poi}(k(1-\ell_2))$, and then simulate drawing t_1^H samples from (a rescaled) $p_1^-(H)$, and t_2^H samples from (a rescaled) $p_2^-(H)$.
- For each sample from p_2^L with probability $1 - \frac{\ell_2}{\ell}$ discard it.
- Run the Poissonized T on all the simulated samples, the remaining samples from p_2^L and the (unaltered) samples from p_1^L.

By construction this procedure exactly simulates how a k-Poissonized T would run on (p_1, p_2), so thus T' is a tester for π' with soundness at least $\frac{1}{12}$.

To reach the desired contradiction, we now show that in fact no such tester can exist. Note that by definition, $\pi'(p_1^-(L), p_2^-(L)) < a$ and $\pi'(p_1^+(L), p_2^+(L)) > b$. Consider the distributions obtained by applying the Moments Matching Theorem to each of these pairs. Explicitly, letting $w = \delta$ we define $(\bar{p}_1^-(L), \bar{p}_2^-(L)) \leftarrow M_\delta^{k\ell}(p_1^-(L), p_2^-(L))$ and $(\bar{p}_1^+(L), \bar{p}_2^+(L)) \leftarrow M_\delta^{k\ell}(p_1^+(L), p_2^+(L))$. From the Moments Matching Theorem's three conclusions we have (1) that the modified distributions are $\bar{k} = \frac{k\delta}{100 \cdot 2^{6\sqrt{\log n}}}$-low frequency; (2) that the statistical distance between each modified pair and the corresponding original pair is at most δ, which, since π is (ϵ, δ)-weakly-continuous implies that $\pi'(\bar{p}_1^-(L), \bar{p}_2^-(L)) < a + \epsilon$ and $\pi'(\bar{p}_1^+(L), \bar{p}_2^+(L)) > b - \epsilon$; and (3) that that the $\bar{k}$-based moments of $(\bar{p}_1^-(L), \bar{p}_2^-(L))$ and $(\bar{p}_1^+(L), \bar{p}_2^+(L))$ up to degree $\sqrt{\log n}$ are equal to within $\frac{2}{10000 \log n}$. Applying the corollary to the wishful thinking theorem for k equal to the number of samples T' takes, namely $\frac{\bar{k}}{1000}$, we have that the statistical distance between the distributions of samples returned when testing $(p_1^-(L), p_2^-(L))$ versus $(p_1^-(L), p_2^-(L))$ is at most $.04 + \frac{40}{1000} + 10 \sum_{a+b \leq \sqrt{\log n}} \frac{2}{10000} \log n \leq .081$, which by the Positive-Negative Distance Lemma implies that T' cannot exist, the desired contradiction. □

Proof of Lemma 8. Let p^+ and p^- be distributions at most $\frac{1}{2 \log n}$ far apart. Then the difference in their

entropies is bounded as

$$\begin{aligned}\left|\sum_i p^+(i)\log p^+(i) - p^-(i)\log p^-(i)\right| &\le \sum_i |p^+(i)\log p^+(i) - p^-(i)\log p^-(i)| \\ &\le \sum_i -|p^+(i)-p^-(i)|\log|p^+(i)-p^-(i)| \\ &\le -|p^+-p^-|\log\left[\frac{1}{n}|p^+-p^-|\right] \le (2\log n)|p^+-p^-| \le 1,\end{aligned}$$

where the first inequality is the triangle inequality, the second inequality results from the fact that the function $x\log x$ is convex, the third inequality is Jensen's inequality applied to the convex function $x\log x$, and the last inequality is from the fact that $|p^+-p^-| \ge \frac{1}{2\log n} \ge \frac{1}{n}$. □

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/