Testing Symmetric Properties of Distributions

by

Paul Valiant

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Author . . . .
Department of Electrical Engineering and Computer Science
May 23, 2008

Certified by . . . .
Silvio Micali
Professor of Computer Science
Thesis Supervisor

Accepted by . . . .
Arthur Smith
Chairman, EECS Committee on Graduate Students

Testing Symmetric Properties of Distributions

by

Paul Valiant

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

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 properties of one or two distributions that are symmetric and satisfy 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.

Thesis Supervisor: Silvio Micali
Title: Professor of Computer Science

Acknowledgments

I am indebted to each member of my thesis committee – Silvio Micali, Ronitt Rubinfeld and Madhu Sudan – for their longstanding guidance, support, and friendship.

Silvio Micali, my advisor, has been a tireless source of inspiration, ideas, and raw energy since I first arrived at MIT. Our interactions are distinguished by his knack for recasting seemingly intractable problems into ones whose solutions are simple, natural, and deep. Where once I might have seen an insurmountable problem, I now realize that there are always many more ways around an obstacle than over it; nothing has broadened my view of the research landscape more. And despite his protestations that "perfection is the enemy of the good", he has kept my eyes firmly directed towards that elusive goal, despite all cross-currents, obstacles, and distractions in the way. His dedication to speaking and writing with flair serves as a continuing example by which to improve my own work.

Ronitt Rubinfeld is responsible for directing me to the beautiful problems with which this thesis is concerned. For this, and her inexhaustible enthusiasm about both the field of property testing and my efforts in it, I am deeply grateful.

Madhu Sudan was my first advisor at MIT, and to him go my thanks for a smooth and effective introduction to the art of computer science research. To him I also owe the occasional temptation to regard all of computer science as an adjunct field of coding theory, since, after all, "the mere realization that a problem can be related to coding theory [is often] the solution." For many problems solved, I thank him.

I must also thank all my coauthors, collaborators, and office mates for much time well spent together (with sincere apologies to the people I'm assuredly forgetting) Tim Abbott, Adi Akavia, Arnab Bhattacharyya, Xi Chen, Costis Daskalakas, Brendan Juba, Daniel Kane, Swastik Kopparty, Kevin Matulef, Rafael Pass, Chris Peikert, Alon Rosen, Ben Rossman, Guy Rothblum, Amit Sahai, Grant Schoenbeck, abhi shelat, Tasos Sidiropoulos, Adam Smith, Shang-Hua Teng, and Ning Xie.

I am also grateful to Be Blackburn and Joanne Hanley for insulating me from departmental bureaucracy with cheerfulness, efficiency, and large amounts of candy.

Finally, I would like to thank my parents and brother Greg for everything else.

Paul Valiant

May, 2008

Contents

Chapter 1

Introduction

Computer hardware and software has advanced to the point where for almost any feasible test of speed or memory in which computers can reasonably compete, they may be set up to outperform people – and thus with respect to the touchstones of time and space complexity, computers can be said to have beaten the benchmark of human-level performance. One area, however, in which our abilities vastly exceed anything currently attainable algorithmically is that of *data complexity*: how much data does one need to classify a phenomenon? Human ability to make accurate decisions on apparently little data is prodigious and sometimes manifests itself as "one-shot learning".

It is with this goal of data efficiency that the developing field of *property testing* is concerned. Explicitly, property testing asks what is the minimum amount of data needed about an object to probably return a correct decision on whether it possesses a certain property. Property testing has been extensively investigated in a variety of settings, in particular, graph testing (e.g. [12]), testing of algebraic properties (e.g. [9, 20]), and the related area of program checking (e.g. [8, 9]). In particular, we draw the reader's attention to the recent 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 a collection of beautiful but specific results, without a common framework. In this thesis we aim to remedy this.

1.0.1 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 over each element of the absolute value of the differences between the probability of this element under the two distributions). Again, it is clear that this property does not depend on the specific naming scheme for the elements.

1.0.2 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 thesis —see Theorems 1.1.1, 1.1.2, and 1.1.3 respectively, and Chapter 2 for definitions. The prior state of the art for the three problems we consider is:

Closeness Testing Distinguishing two identical distributions 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:

1.1.1 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 symmetric properties satisfying this condition, we essentially prove that *there is no difference between proving an upper bound and proving a lower bound.* We formalize this with the notion of a *Canonical Tester.*

[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.

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 *a fortiori* the property is testable with $f(n)$ samples; if the Canonical Tester does not test the property, then (as we show) 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]

1.1.2 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 the main results of [18]; we sketch this construction at the end of Section 3.3.[3]

Theorem 1.1.1. *Distinguishing two identical distributions from two distributions with statistical distance at least* $\frac{1}{2}$ *requires* $\Omega(n^{2/3})$ *samples.*

Theorem 1.1.2. *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 1.1.3. *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 1.1.2 and 1.1.3 result directly from the Canonical Tester; along the path to deriving the properties of the Canonical Tester, we prove a structural theorem that

[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]We note that in all the new results of this thesis the Canonical Tester improves a lower bound to match the performance of a known algorithm. It might be interesting 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.

may be of independent interest, the Wishful Thinking Theorem (Theorem 4.5.6), and it is from this result that we derive Theorem 1.1.1.

1.2 Our Techniques

The properties of the Canonical Tester, as described above, are encapsulated in the Canonical Testing Theorem (Theorem 3.1.2). This theorem is essentially equivalent to a result that at first glance looks to be very different: the Low-Frequency Blindness Theorem (Theorem 3.1.3) states that testers for symmetric weakly-continuous properties are by necessity "blind" to the low-weight portion of the distribution. That is, if we wish to test a property by taking k samples from a distribution, then the tester will be essentially blind to any element occurring with probability much less than $\frac{1}{k}$, even despite the possibility that the low-frequency elements may *collectively* constitute most of the probability mass of the distribution. This result immediately leads to applications such as: any symmetric weakly-continuous property which is true for the uniform distribution on n elements and false for the uniform distribution on $\frac{n}{2}$ elements needs roughly n samples to test, for otherwise all the elements from both distributions would fall in the tester's "blind spot".

We emphasize that the Low-Frequency Blindness Theorem is *not* saying that distinguishing the uniform distribution on n elements from the uniform distribution on $\frac{n}{2}$ elements takes n samples; rather, the theorem rests delicately on its assumptions that we are testing a symmetric weakly-continuous property. To illustrate this, we note first that the problem of distinguishing the uniform distribution on n elements from the uniform distribution on the first half of these elements takes $O(1)$ samples: answer according to whether or not any samples appear from the second half of the elements. However, if we add the condition of symmetry (but not yet continuity!) then we know that a property which is true on the uniform distribution over n elements and false on the uniform distribution over the first $\frac{n}{2}$ elements must also be false on uniform distributions over *any* $\frac{n}{2}$ elements, and thus to test this property we must be able to distinguish the uniform distribution on n elements from the uniform distribution on an *arbitrary* $\frac{n}{2}$-sized subset of these elements. Birthday paradox arguments show that one can do this with $\theta(\sqrt{n})$ samples, by counting how many elements are sampled twice; standard arguments show this is tight.

To summarize: if all we know about a property is that it is true for a uniform distribution on n elements and false for a uniform distribution on $\frac{n}{2}$ elements, then all one can show is $\Omega(1)$ sample complexity; if in addition, however, we know that the property is symmetric, then we have $\Omega(\sqrt{n})$ sample complexity; the main contribution of this thesis is techniques to see that if we additionally know that the property is weakly-continuous then we have $n^{1-o(1)}$ sample complexity.

1.2.1 Wishful Thinking

The first step in the derivation of the main results of this thesis is a general and precise characterization of the role of symmetry in testing properties of low-frequency distributions – that is, before we can add the crucial ingredient of continuity, we must first shore up our understanding of symmetric properties.

This result, which we call the Wishful Thinking Theorem (Theorem 4.5.6) formalizes and proves an intuition which has appeared implicitly in [6] and (in slightly different form) in [5]. The theorem states essentially, that when testing symmetric properties of the low-frequency portion of distributions, the only relevant information about the distribution consists of its moments, informally, “moments describe all”.

We can thus use this theorem to prove lowerbounds for testing a symmetric property given any construction of a positive and a negative example of the property whose moments almost match. Indeed, we take such a construction from [6] to immediately prove our Theorem 1.1.1.

1.2.2 Matching Moments

The final piece needed for low-frequency blindness, complementing the “moments describe all” result for symmetric properties, is a “moments describe nothing” result for weakly-continuous properties, roughly, that moments fail to determine whether a weakly-continuous property is true or false (see Theorem 5.2.5). Combined, we see that if moments describe all the useful information that a tester can extract from low-frequency elements, but that moments fail to determine the property, then testers are low-frequency blind.

The technical portion of this theorem concerns an analysis of the solution of linear equations with a Vandermonde matrix for coefficients. This idea comes from [18].

Chapter 2

Definitions

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. Subscripts are used almost exclusively to index one of the two elements of a pair, as in p_1, p_2, for those contexts where we analyze properties of pairs of distribution.

Definition 2.0.1. *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 $[n]$, and $\mathcal{D}_n^2$ to denote the set of all pairs of distributions.*

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

Definition 2.0.2. *A* property *of a (single) distribution is a function $\pi : \mathcal{D}_n \to \mathbb{R}$. A property of a pair of distributions is a function $\pi : \mathcal{D}_n^2 \to \mathbb{R}$. A* binary property *of a distribution (respectively, distribution pair) is a function $\beta : \mathcal{D}_n \to \{$"yes", "no", $\emptyset\}$ (respectively, $\beta : \mathcal{D}_n^2 \to \{$"yes", "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 2.0.3. *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, the 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 $\pi_a^b(p) = \emptyset$.*

Definition 2.0.4. *Given a binary property π_a^b on distribution* pairs *and functions $k_1, k_2 : \mathbb{Z}^+ \to \mathbb{Z}^+$, an algorithm T is a "π_a^b-tester with sample complexity $(k_1(\cdot), k_2(\cdot))$" if, for any distribution pair p_1, p_2, algorithm T on input $k_1(n)$ random samples from p_1 and $k_2(n)$ random samples from p_2 will accept with probability greater than $\frac{2}{3}$ if $\pi_a^b(p_1, p_2) =$ "yes", and accept with probability less than $\frac{1}{3}$ if $\pi_a^b(p_1, p_2) =$ "no". The behavior is unspecified when $\pi_a^b(p_1, p_2) = \emptyset$.*

If we refer to a distribution pair tester as having "sample complexity $k(\cdot)$" we intend this as shorthand for having "sample complexity $(k(\cdot), k(\cdot))$".

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 2.0.5. *A property π 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 of distribution pairs π 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 2.0.6. *A property π is symmetric if for all distributions p and all permutations $\sigma \in S_n$, the symmetric group on $[n]$, we have $\pi(p) = \pi(p \circ \sigma)$. A property of distribution pairs π is symmetric if for all distributions p_1, p_2 and all permutations $\sigma \in S_n$ we have $\pi(p_1, p_2) = \pi(p_1 \circ \sigma, p_2 \circ \sigma)$.*

We note that this definition of symmetry for properties of distribution pairs is more permissive than a natural variant which would insist that the property be invariant for all *pairs* of permutations σ_1, σ_2, that is, $\pi(p_1, p_2) = \pi(p_1 \circ \sigma_1, p_2 \circ \sigma_2)$. This stronger notion of symmetry would disallow any notion of correlating between the two distributions, and specifically does not include the property that measures statistical distance $|p_1 - p_2|$. All results in this thesis are for the more general notion of symmetry, as stated in Definition 2.0.6, so that we may work with statistical distance and related properties.

Chapter 3

The Canonical Tester and Applications

3.1 The Single Distribution Case

To motivate the rest of this thesis 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 $\mathcal{C}^\theta$ tester for π_a^b* returns an answer “yes” or “no” according to the following steps.

Definition 3.1.1 (Canonical Tester C^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 thesis.

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 3.1.2 (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 k-sample Canonical Tester C^θ for $\theta = \frac{600 \log n}{\delta^2}$ on π_a^b fails to distinguish between $\pi > b + \epsilon$ and $\pi < a - \epsilon$, then no tester can distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$ in $\frac{\delta}{1000 \cdot 2^{4\sqrt{\log n}}} k$ 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.

3.1.1 Discussion

While it will take us the rest of the work 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 it 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". This is 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 3.1.3 (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 $\frac{\delta}{1000 \cdot 2^{4\sqrt{\log n}}} k$ 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

4.5.6), 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 5.2.5), 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 then apply the Wishful Thinking Theorem.

3.2 The Two Distribution Case

Given a binary property on two distributions $\pi_a^b : \mathcal{D}_n^2 \to \{\text{“yes”}, \text{“no”}, \emptyset\}$, two sets of samples from $[n]$ of sizes k_1, k_2 respectively represented as a pair of histograms $s_1, s_2 : [n] \to \mathbb{Z}^+$ counting the number of times each element has been sampled in each of the two distributions, and a threshold $\theta \in \mathbb{Z}^+$, then the (k_1, k_2)-*sample* $\mathcal{C}^\theta$ *tester* for π_a^b returns an answer “yes” or “no” according to the following steps.

Definition 3.2.1 (2-Distribution Canonical Tester C^θ for π_a^b)**.**

1. *For each i such that $s_1(i) > \theta$ or $s_2(i) > \theta$ insert the pair of constraints $p_1(i) = \frac{s_1(i)}{k_1}$ and*[1] *$p_2(i) = \frac{s_2(i)}{k_2}$, otherwise insert the pair of constraints $p_1(i) \in [0, \frac{\theta}{k_1}]$ and $p_2(i) \in [0, \frac{\theta}{k_2}]$.*

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

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.*

The corresponding theorem is almost exactly the one of the single distribution case, with the constants slightly modified.

Theorem 3.2.2 (2-Distribution Canonical Testing Theorem)**.** *Given a symmetric (ϵ, δ)-weakly-continuous property on distribution pairs $\pi : \mathcal{D}_n^2 \to \mathbb{R}$ and two thresholds*

[1]The “and” here is in crucial contrast to the “or” of the previous line —see the discussion below.

$a < b$, such that the $(k_1, k_2)-$sample Canonical Tester C^θ for $\theta = \frac{600 \log n}{\delta^2}$ on π_a^b fails to distinguish between $\pi > b + \epsilon$ and $\pi < a - \epsilon$, then no tester can distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$ in $\frac{\delta}{640000 \cdot 2^{7\sqrt{\log n}}}(k_1, k_2)$ samples.

3.2.1 Discussion

As noted above, the one surprise in the generalization of the Canonical Tester is the "and" in Step (1) of Definition 3.2.1 where it might perhaps be more intuitive to expect an "or". Explicitly, if we observe many samples of a certain index i from the first distribution and few samples from the other distribution, then, while it might be a more natural generalization of Definition 3.1.1 if we were to insert a equality constraint for the first distribution only, this intuition is misleading and we must in fact use equality constraints for both distributions. We defer a rigorous explanation to the final chapter, but mention a few partial justifications here. First, we do not aim to test two separate properties of two distribution, but rather a joint property of two distributions, so it is natural for our tester to process the samples in joint fashion, with samples from one distribution affecting the analysis of samples from the other. Second, we put forward the notion that those indices i which do not receive a statistically significant number of samples may be said to be "invisible" to a property tester; conversely, if an index i receives a large number of samples from either distribution, it suddenly becomes "visible", and we must pay special attention to this index, each time it is sampled from either distribution. Finally, we note that this choice to use a stronger constraint leads to a smaller set P of feasible distribution pairs, and thus can only shrink the set $\pi_a^b(P)$, which will only make Step (4) of the algorithm more likely to return a definite answer.

As in the single distribution case, a fundamental ingredient of the proof of the 2-distribution Canonical Testing Theorem is a "low-frequency blindness" result:

Theorem 3.2.3 (2-Distribution Low Frequency Blindness). *Given a symmetric property π on distributions pairs on $[n]$ that is (ϵ, δ)-weakly-continuous, numbers $k_1, k_2 \in \mathbb{R}^+$, and two distribution pairs, $p_1^+, p_2^+, p_1^-, p_2^-$ that are identical for any index occur-*

ring with probability at least $\frac{1}{k_1}$ *in* p_1^+ *or* p_1^- *or with probability at least* $\frac{1}{k_2}$ *in* p_2^+ *or* p_2^-*, but where* $\pi(p_1^+, p_2^+) > b$ *and* $\pi(p_1^-, p_2^-) < a$*, then no tester can distinguish between* $\pi > b - \epsilon$ *and* $\pi < a + \epsilon$ *in* $\frac{\delta}{640000 \cdot 2^{7\sqrt{\log n}}}(k_1, k_2)$ *samples.*

3.3 Applications

We prove Theorems 1.1.2 and 1.1.3 here, and further, outline how to reproduce the results of [18] on estimating the distribution support size. (Theorem 1.1.1 is shown at the end of Chapter 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 —in the examples below they are often in fact quite easy to distinguish. We remind the reader of the discussion in the second and third paragraphs of Section 1.2.

In practice, it may be quite hard to come up with indistinguishable distributions satisfying certain other properties, and for this reason we have set up the machinery of this thesis to save the property testing community from this step: 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.

3.3.1 The Entropy Approximation Bound

As a technical but straightforward preliminary we show that entropy is weakly continuous:

Lemma 3.3.1. *The entropy function of distributions in $\mathcal{D}_n$ is $(1, \frac{1}{2\log n})$-weakly-continuous.*

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

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

where the first inequality is the triangle inequality, the second inequality holds term-by-term as can be easily checked, the third inequality is Jensen's inequality applied to the convex function $x\log x$; the last inequality can be seen by letting $x = \frac{1}{n}|p^+ - p^-| \leq \frac{1}{2n\log n}$, from which we can easily see that $-nx\log x \leq 1$, as desired. $\square$

We now prove our bound on entropy approximation —a more precise form of Theorem 1.1.3.

Lemma 3.3.2. *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. $\square$

We compare this to the best previous result of [18], which applies for θ less than $\frac{2}{3\gamma^2}$, a factor of $\frac{2}{3}$ off in the exponent.

We also note the significance of the bound $\frac{\log n}{\gamma^2} - 2$ in that if we were guaranteed that the distribution has entropy at least $\frac{\log n}{\gamma^2}$ then a γ approximation is obtained by the *constant* guess of $\frac{\log n}{\gamma}$. Our result shows surprisingly that if we enlarge this range by only 2, then we get (essentially) linear time inapproximability.

3.3.2 The Statistical Distance Bound

Proof of Theorem 1.1.2. 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. □

3.3.3 The Distribution Support Size Bound

We sketch how to reproduce from our techniques a $n^{1-o(1)}$ lowerbound on testing Distribution Support Size, which is the main result of [18]. 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, each element with nonzero probability must have 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 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. Essentially, there is only one point in the proof of the Low Frequency Blindness Theorem where we modify the input distributions p^+, p^-, and that is in the construction of Definition 5.2.4; in this construction, the distributions are modified three times, none of which will introduce weights in the interval $(0, \frac{1}{n})$: in Step 1, some probabilities are made 0; in Step 4 some probabilities much larger than $\frac{1}{k} > \frac{1}{n}$ are introduced; and in Step 5 some probabilities are modified to be uniformly at least $\frac{1}{n}$ as noted by the footnote in the proof of Theorem 5.2.5.

3.4 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.[2] Indeed we suspect that this tester —or something very similar— may be shown optimal for more general properties. However, neither the symmetry nor the continuity condition can be relaxed entirely:

- Consider the problem of determining whether a (single) 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, if the number of samples —k— 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 1.1.1, determining whether a pair of distributions are

[2] 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 "very discontinuous" properties. The situation in between remains open.

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 is easy to see that the Canonical Tester for $\pi_{-1/2}^{1/2}$ requires $\widetilde{\Theta}(n)$ samples, which is $\sim n^{1/3}$ worse than the bound of $\widetilde{O}(n^{2/3})$ provided by [6] (and proven optimal by our Theorem 1.1.1).

Chapter 4

The Wishful Thinking Theorem

4.1 Histograms and Fingerprints

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].)

4.2 Intuition

Our goal in this chapter 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 chapter, 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 chapter.

> 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 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.

4.3 Poissonization

Definition 4.3.1. *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 from the univariate $Poi(\vec{\lambda}(i))$ for each i.

Definition 4.3.2. *A k-*Poissonized tester *T (for properties of a single distribution) 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 4.3.3. *If there exists a k-sample tester T for a binary property π, then there exists a k-Poissonized tester T' for π.*

Proof. With probability at least $\frac{1}{2}$, independent of $\pi(p)$, k' drawn from $Poi(k)$ will have value at least k. Let T' simulate T when given at least k samples, and return a random answer otherwise. Thus with probability at least $\frac{1}{2}$ T' will simulate T, which returns a correct answer with probability at least $\frac{2}{3}$, and the remainder of the time T' will guess with 50% success, yielding a total success rate at least $\frac{1}{2}\frac{2}{3} + \frac{1}{2}\frac{1}{2} = \frac{7}{12}$. □

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.

4.4 Roos's Theorem and Multinomial Distributions

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 4.4.1. *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 4.4.2. *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 ith 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.

4.5 Assembling the Pieces

We begin by analyzing the approximation error of Roos's Theorem in the case that concerns us here: when the multinomial distribution models the distribution of fingerprints of Poissonized samples from a low-frequency distribution.

Lemma 4.5.1. *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 provide such a comparison, we first derive the statistical distance between univariate Poisson distributions.

Lemma 4.5.2. *The statistical distance between two univariate Poisson distributions with parameters λ, λ' is bounded as*

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

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 $\mathrm{Poi}(\lambda)$ and $\mathrm{Poi}(\lambda')$, 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

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

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 [?] 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 $0 \leq \lambda \leq \lambda'$. Further, we note that $\text{poi}(i; \lambda) - \text{poi}(i; \lambda') = \frac{1}{i!}[e^{-\lambda}\lambda^i - e^{\lambda'}\lambda^i] \leq 0$ 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 nonnegative for $x \in [0, 1] \supset [\lambda, \lambda']$. 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. □

The corresponding multivariate bound is as follows:

Lemma 4.5.3. *The statistical distance between two multivariate Poisson distributions with parameters $\vec{\lambda}^+, \vec{\lambda}^-$ is bounded as*

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

Proof. We prove this as a direct consequence of Lemma 4.5.2 and the fact that the statistical distance of multivariate distributions with independent marginals is at most the sum of the corresponding distances between the marginals, which we prove here.

Suppose we have bivariate distributions $p(i, j) = p_1(i) \cdot p_2(j)$ and $p'(i, j) = p_1'(i) \cdot$

$p_2'(j)$ then

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

Induction yields the subadditivity claim for arbitrary multivariate distributions, and thus we conclude this lemma from Lemma 4.5.2. $\square$

Combining results yields:

Lemma 4.5.4. *Given a positive integer k and two distributions p^+, p^- all of whose frequencies are at most $\frac{1}{500k}$, then, letting $\vec{\lambda}^+(a) = \sum_i \mathrm{poi}(a; k \cdot p^+(i))$ and $\vec{\lambda}^-(a) = \sum_i \mathrm{poi}(a; k \cdot p^-(i))$ for $a > 0$, if it is the case that*

$$\sum_{a>0} \frac{|\vec{\lambda}^+(a) - \vec{\lambda}^-(a)|}{\sqrt{1 + \max\{\vec{\lambda}^+(a), \vec{\lambda}^-(a)\}}} < \frac{1}{25}. \tag{4.1}$$

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

Proof. Combining Lemma 4.5.1 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 4.5.3, which (from Equation 4.1) 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 4.3.3 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 somewhat technical.

Definition 4.5.5. *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 4.5.6 (Wishful Thinking)**.** *Given an integer $k > 0$ 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*

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

then it is impossible to test any symmetric property that is true for p^+ and false for p^- in k samples.[1]

Proof. We derive the theorem as a consequence of Lemma 4.5.4. We start from Equation 4.1, (recall the definition $\vec{\lambda}^+(a) = \sum_i \text{poi}(a; k \cdot p^+(i)) = \frac{k^a}{a!} \sum_i e^{-k \cdot p^+(i)} p^+(i)^a$ and the corresponding one for $\vec{\lambda}^-(a)$) and expand both the numerator and denominator of each fraction via Taylor series expansions.

For the numerator of the a term we have from Taylor expansions and the triangle inequality that

$$\begin{aligned}
|\lambda^+(a) - \lambda^-(a)| &= \frac{k^a}{a!} \left| \sum_i \left[e^{-k \cdot p^+(i)} p^+(i)^a - e^{-k \cdot p^-(i)} p^-(i)^a \right] \right| \\
&= \frac{1}{a!} \left| \sum_i \sum_\gamma \frac{(-1)^\gamma}{\gamma!} k^{a+\gamma} \left[p^+(i)^{a+\gamma} - p^-(i)^{a+\gamma} \right] \right| \\
&= \frac{1}{a!} \left| \sum_\gamma \frac{(-1)^\gamma}{\gamma!} [m^+(a+\gamma) - m^-(a+\gamma)] \right| \\
&\le \frac{1}{a!} \sum_\gamma \frac{1}{\gamma!} \left| m^+(a+\gamma) - m^-(a+\gamma) \right|.
\end{aligned}$$

[1] We note that we may strengthen the lemma by inserting a term of $\lfloor \frac{a}{2} \rfloor !$ in the denominator of the summand; for simplicity of presentation, and since we never make use of this stronger form, we prove the simpler version. See Section 4.6 for a version of the lemma with this term.

We now bound terms in the denominator of Equation 4.1. Since $p^+(i), p^-(i) \leq \frac{1}{500k}$ by assumption, we have $e^{k\cdot p^+(i)}, e^{k\cdot p^-(i)} > 0.9$, which implies that $\vec{\lambda}^+(a) > \frac{0.9}{a!}m^+(a)$ by definition of m^+, with corresponding expression holding for $\vec{\lambda}^-$ and m^-. Thus we bound terms in the denominator of Equation 4.1 as

$$\sqrt{1+\max\{\vec{\lambda}^+(a), \vec{\lambda}^-(a)\}} \geq \frac{0.9}{\sqrt{a!}}\sqrt{1+\max\{m^+(a), m^-(a)\}}.$$

Combining the bounds for the numerator and denominator, where in the second line we make use of the fact that (since $p^+(i), p^-(i) \leq \frac{1}{k}$) both m^+ and m^- are decreasing functions of their index, and where we make the variable substitution $\mu = a+\gamma$ in the third line, yields

$$\begin{aligned}
\sum_{a>0} \frac{|\vec{\lambda}^+(a) - \vec{\lambda}^-(a)|}{\sqrt{1+\max\{\vec{\lambda}^+(a), \vec{\lambda}^-(a)\}}} &\leq \sum_{a>0}\sum_{\gamma} \frac{|m^+(a+\gamma) - m^-(a+\gamma)|}{0.9\gamma!\sqrt{a!}\sqrt{1+\max\{m^+(a), m^-(a)\}}} \\
&\leq \sum_{a>0}\sum_{\gamma} \frac{|m^+(a+\gamma) - m^-(a+\gamma)|}{0.9\gamma!\sqrt{a!}\sqrt{1+\max\{m^+(a+\gamma), m^-(a+\gamma)\}}} \\
&= \sum_{\mu}\sum_{\gamma<\mu} \frac{|m^+(\mu) - m^-(\mu)|}{0.9\gamma!\sqrt{(\mu-\gamma)!}\sqrt{1+\max\{m^+(\mu), m^-(\mu)\}}} \\
&= \sum_{\mu} \frac{|m^+(\mu) - m^-(\mu)|}{\sqrt{1+\max\{m^+(\mu), m^-(\mu)\}}} \frac{1}{0.9} \sum_{\gamma<\mu} \frac{1}{\gamma!\sqrt{(\mu-\gamma)!}}.
\end{aligned}$$

We note that the expression $\sum_{\gamma<\mu} \frac{1}{\gamma!\sqrt{(\mu-\gamma)!}}$ clearly tends to 0 for large μ, as each of the μ terms is at most $\frac{1}{\lfloor \mu/2 \rfloor!}$; evaluating for small μ we see that this expression attains its maximum value of $1+\frac{\sqrt{2}}{2}$ at $\mu = 2$. Thus $\frac{1}{0.9}\sum_{\gamma<\mu} \frac{1}{\gamma!\sqrt{(\mu-\gamma)!}} \leq 2$, from which we conclude that $\sum_{a>0} \frac{|\vec{\lambda}^+(a)-\vec{\lambda}^-(a)|}{\sqrt{1+\max\{\vec{\lambda}^+(a),\vec{\lambda}^-(a)\}}} < 2\sum_{a\geq 0} \frac{|m^+(a)-m^-(a)|}{\sqrt{1+\max\{m^+(a),m^-(a)\}}}$. Finally, we note that $m^+(0) = \sum_{i=1}^n p^+(i)^0 = n$ and $m^+(1) = k\sum_i p^+(i) = k$, regardless of p^+, and thus by symmetry, $m^+(0) = m^-(0)$ and $m^+(1) = m^-(1)$. Thus this last sum equals $\sum_{a>1} \frac{|m^+(a)-m^-(a)|}{\sqrt{1+\max\{m^+(a),m^-(a)\}}}$, which by hypothesis is less than $\frac{1}{50}$, from which we conclude that $\sum_{a>0} \frac{|\vec{\lambda}^+(a)-\vec{\lambda}^-(a)|}{\sqrt{1+\max\{\vec{\lambda}^+(a),\vec{\lambda}^-(a)\}}} < \frac{1}{25}$. We invoke Lemma 4.5.4 to finish. □

We will find it convenient to work with a finite subset of the moments in Chapter 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 4.5.7. *Given an integer $k > 0$, real number $\epsilon \leq \frac{1}{10 \cdot 2^{\sqrt{\log n}}}$ 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 $m^+(a) \leq k^a n (\frac{\epsilon}{k})^a = n\epsilon^a \leq .1^a$ Thus, since $\frac{|m^+(a)-m^-(a)|}{\sqrt{1+\max\{m^+(a), m^-(a)\}}} \leq m^+(a)$, we can bound these terms by $\sum_{a \geq 2} .1^{a+b} < \frac{1}{50} - \frac{1}{120}$, yielding the corollary. □

4.6 The Two Distribution Case

We follow the same outline as for the single distribution case.

The first step is to define the fingerprint of samples from a pair of distributions. As above, it is defined as the histogram of the histogram of the samples, but because we have two distributions instead of one, the form of the fingerprint is a bit more intricate. Let us introduce this by way of an example. Suppose we draw 7 samples from each of two distributions, with the sequence $(3, 1, 2, 2, 5, 1, 2)$ being drawn from the first distribution, and $(4, 3, 1, 2, 3, 5, 5)$ being drawn from the second distribution. A single application of the histogram function returns a sequence of pairs $((2, 1), (3, 1), (1, 2), (0, 1), (1, 2))$ indicating that 1 was seen twice from the first distribution and once from the second distribution; 2 was seen three times from the first distribution and once from the second; 3 was seen once from the first distribution and twice from the second distribution, etc. The second application of the histogram now takes as input these five pairs, and thus returns a table counting how many times each *pair* was seen. That is, the fingerprint of these samples is the matrix

$$\left(\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 \end{array}\right)$$

which indicates that the pair $(0, 1)$ occurs once in the histogram, the pair $(1, 2)$ occurs twice, the pair $(2, 1)$ occurs once, and the pair $(3, 1)$ occurs once. Or, stretching the language slightly, we have one "(0,1)-way collision", two "(1,2)-way collisions", one "(2,1)-way collision", and one "(3,1)-way collision".

We give a formal definition and prove the fact that the fingerprint captures all the useful information about the samples.

Definition 4.6.1. *Given two sequences 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 the number of elements of X that appear exactly i times in S_1 and j times in S_2.*

Lemma 4.6.2. *For any symmetric property π of distribution pairs and random variables κ_1, κ_2, if there exists a tester T taking as input κ_1 samples from the first distribution and κ_2 samples from the second distribution, then there exists a tester T' which takes as input only the fingerprint of κ_1 samples drawn from the first distribution and κ_2 samples drawn from the second distribution.*

Proof. Given T and a fingerprint $f(\cdot, \cdot)$ of κ_1, κ_2 samples respectively from distributions p_1 and p_2 on $[n]$ we let T' run 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_1, κ_2 samples respectively from the distributions $p_1 \circ \gamma$ and $p_2 \circ \gamma$. Furthermore, since T is a tester for a symmetric property, it has the same performance guarantees for (p_1, p_2) as for $(p_1 \circ \gamma, p_2 \circ \gamma)$ for any permutation γ. Thus T will also operate correctly when γ is drawn randomly, which implies that T' is a tester for π, as desired. $\square$

Following the outline from above, we next consider Poissonized testers of distribution pairs. Akin to Definition 4.3.2 and Lemma 4.3.3 we have (note the slight change in constants):

Definition 4.6.3. *A* (k_1, k_2)-Poissonized tester T *(for properties of two distributions) is a function that correctly classifies a property on a distribution pair* p_1, p_2 *with probability* $\frac{13}{24}$ *on input samples generated in the following way:*

- *Draw* $k_1' \leftarrow Poi(k_1)$ *and* $k_2' \leftarrow Poi(k_2)$.
- *Return* k_1' *samples from* p_1 *and* k_2' *samples from* p_2.

We have the following standard lemma:

Lemma 4.6.4. *If there exists a* (k_1, k_2)*-sample tester* T *for a 2-distribution binary property* π*, then there exists a* (k_1, k_2)*-Poissonized tester* T' *for* π*.*

Proof. With probability at least $\frac{1}{4}$, independent of $\pi(p_1, p_2)$, we will have both $k_1' \geq k_1$ and $k_2' \geq k_2$. Let T' simulate T when given at least k_1, k_2 samples respectively from the distributions, and return a random answer otherwise. Thus with probability at least $\frac{1}{4}$ T' will simulate T, which returns a correct answer with probability at least $\frac{2}{3}$, and the remainder of the time T' will guess with 50% success, yielding a total success rate at least $\frac{1}{4}\frac{2}{3} + \frac{3}{4}\frac{1}{2} = \frac{13}{24}$. $\square$

The next step is to express the distribution of fingerprints of (k_1, k_2)-Poissonized samples as a multinomial distribution. As above, we create a matrix ρ with rows corresponding to elements of distributions' domain, and columns corresponding to histogram entries. We note that in this case, however, the histogram is not indexed by a single index (a) as it was above, but instead by a pair of indices, which we take to be a, b. Thus ρ is indexed as $\rho(i, (a, b))$.

Akin to Lemma 4.4.2 we have:

Lemma 4.6.5. *For any pair of distributions p_1, p_2 with support $[n]$ and positive integers k_1, k_2, the distribution of fingerprints of $Poi(k_1)$ samples from p_1 and $Poi(k_2)$ samples from p_2 is the generalized multinomial distribution M^ρ where matrix ρ has n rows, columns indexed by fingerprint indices a, b, and $(i, (a, b))$ entry equal to* $\mathrm{poi}(a; k_1 \cdot p_1(i))\mathrm{poi}(b; k_2 \cdot p_2(i))$, *that is, the ith row of ρ expresses the bivariate distribution $Poi([k_1 \cdot p_1(i)\,,\, k_2 \cdot p_2(i)])$ over the values (a, b).*

Proof. From basic properties of the Poisson distribution, the ith element of the histogram of [drawing a $Poi(k_1)$-distributed number of samples p_1 and a $Poi(k_2)$-distributed number of samples from p_2] is a pair with the first element drawn (independently) from the distribution $Poi(k_1 \cdot p_1(i))$ and the second element drawn (independently) from the distribution $Poi(k_2 \cdot p_2(i))$. The generalized multinomial distribution M^ρ, by definition, simply draws these samples for each i and returns the histogram, which is distributed as the histogram of the histogram of the original distribution of samples, as desired. □

Roos's Theorem we invoke as is, via a generalization of Lemma 4.5.1

Lemma 4.6.6. *Given a pair of distributions p_1, p_2, integers k_1, k_2, and a real number $0 < \epsilon \leq \frac{1}{2}$ such that $\forall i, p_1(i) \leq \frac{\epsilon}{k_1}$ and $p_2(i) \leq \frac{\epsilon}{k_2}$, if ρ is the matrix with $(i, (a, b))$ entry* $\mathrm{poi}(a; k_1 \cdot p_1(i))\mathrm{poi}(b; k_2 \cdot p_2(i))$ *then* $\sum_{a+b>0} \frac{\sum_i \rho(i,(a,b))^2}{\sum_i \rho(i,(a,b))} \leq 4\epsilon$.

Proof. We note that $\mathrm{poi}(a; k_1 \cdot p_1(i)) = \frac{e^{-k_1 \cdot p_1(i)}(k_1 \cdot p_1(i))^a}{a!} \leq (k_1 \cdot p(i))^a \leq \epsilon^a$, and correspondingly $\mathrm{poi}(b; k_2 \cdot p_2(i)) \leq \epsilon^b$, so thus

$$\sum_{a+b>0} \frac{\sum_i \rho(i, (a, b))^2}{\sum_i \rho(i, (a, b))} \leq \sum_{a+b>0} \max_i \rho(i, (a, b)) \leq \sum_{a+b>0} \epsilon^{a+b} \leq 4\epsilon.$$

$\square$

We thus have the following generalization of Lemma 4.5.4

Lemma 4.6.7. *Given positive integers k_1, k_2 and two distribution pairs $p_1^+, p_2^+, p_1^-, p_2^-$ where p_1^+, p_1^- have frequencies at most $\frac{1}{2000k_1}$ and p_2^+, p_2^- have frequencies at most $\frac{1}{2000k_2}$, then, letting $\vec{\lambda}^+(a,b) = \sum_i \text{poi}(a; k_1 \cdot p_1^+(i))\text{poi}(b; k_2 \cdot p_2^+(i))$ and $\vec{\lambda}^-(a,b) = \sum_i \text{poi}(a; k_1 \cdot p_1^-(i))\text{poi}(b; k_2 \cdot p_2^-(i))$ for $a+b>0$, if it is the case that*

$$\sum_{a+b>0} \frac{|\vec{\lambda}^+(a,b) - \vec{\lambda}^-(a,b)|}{\sqrt{1+\max\{\vec{\lambda}^+(a,b), \vec{\lambda}^-(a,b)\}}} < \frac{1}{50}. \tag{4.2}$$

then it is impossible to test any symmetric property that is true for (p_1^+, p_2^+) and false for (p_1^-, p_2^-) in (k_1, k_2) samples.

Proof. Combining Lemma 4.6.6 with Roos's Theorem we have that for each of (p_1^+, p_2^+) and (p_1^-, p_2^-) the distance of the Poisson approximation from the distribution of fingerprints of (k_1, k_2)-Poissonized samples is at most $\frac{4 \cdot 8.8}{2000} < \frac{1}{50}$. Thus, by the triangle inequality, the distance between the distribution of fingerprints of (k_1, k_2)-Poissonized samples from each of p_1^+, p_2^+ versus each of p_1^-, p_2^- is at most $\frac{2}{50}$ plus the bound from Lemma 4.5.3, which (from Equation 4.2) is also $\frac{2}{50}$, yielding a total distance of at most $\frac{4}{50}$, which is less than $\frac{1}{12}$. Assume for the sake of contradiction that there is a (k_1, k_2)-sample tester that distinguishes between (p_1^+, p_2^+) and (p_1^-, p_2^-). By Lemma 4.6.4 there must thus exist a corresponding (k_1, k_2)-Poissonized tester. However, the definition of a Poissonized tester requires that the tester succeed with probability at least $\frac{13}{24}$ on (p_1^+, p_2^+) and succeed with probability at most $\frac{11}{24}$ on (p_1^-, p_2^-), which contradicts the fact that their input distributions have statistical distance strictly less than $\frac{1}{12}$. Thus no such tester can exist. $\square$

We now reexpress this lemma in terms of the "moments of the distribution pairs" —which we define now. As promised above, we prove a version that is slightly tighter than the single-distribution version in that the condition of the theorem (Equation 4.6.9) now has factorials in the denominator.

Definition 4.6.8. *For integers* (k_1, k_2) *and distribution pair* p_1, p_2*, the* (k_1, k_2)-based moments *of* (p_1, p_2) *are the values* $k_1^a k_2^b \sum_i p_1(i)^a p_2(i)^b$ *for* $a, b \in \mathbb{Z}^+$.

Theorem 4.6.9 (Wishful Thinking for Two Distributions)**.** *Given integers* $k_1, k_2 > 0$ *and two distribution pairs* $(p_1^+, p_2^+), (p_1^-, p_2^-)$ *where* p_1^+, p_1^- *have frequencies at most* $\frac{1}{2000k_1}$ *and* p_2^+, p_2^- *have frequencies at most* $\frac{1}{2000k_2}$*, then, letting* m^+, m^- *be the* (k_1, k_2)*-based moments of* $(p_1^+, p_2^+), (p_1^-, p_2^-)$ *respectively, if it is the case that*

$$\sum_{a>1} \frac{|m^+(a) - m^-(a)|}{\lfloor \frac{\mu}{2} \rfloor ! \lfloor \frac{\nu}{2} \rfloor ! \sqrt{1 + \max\{m^+(a), m^-(a)\}}} < \frac{1}{500}.$$

then it is impossible to test any symmetric property that is true for (p_1^+, p_2^+) *and false for* (p_1^-, p_2^-) *in* (k_1, k_2) *samples.*

Proof. As in the proof of the original Wishful Thinking Theorem, we derive the theorem as a consequence of Lemma 4.6.7. We start from Equation 4.2, and expand both the numerator and denominator of each fraction via Taylor series expansions.

For the numerator of the (a, b) term we have from Taylor expansions and the triangle inequality that

$$\begin{aligned}
|\lambda^+(a,b) - \lambda^-(a,b)| &= \frac{k_1^a k_2^b}{a!b!} \left| \sum_i \left[e^{-k_1 p_1^+(i) - k_2 p_2^+(i)} p_1^+(i)^a p_2^+(i)^b - e^{-k_1 p_1^-(i) - k_2 p_2^-(i)} p_1^-(i)^a p_2^-(i)^b \right] \right| \\
&= \frac{1}{a!b!} \left| \sum_i \sum_{\gamma,\delta} \frac{(-1)^{\gamma+\delta}}{\gamma!\delta!} k_1^{a+\gamma} k_2^{b+\delta} \left[p_1^+(i)^{a+\gamma} p_2^+(i)^{b+\delta} - p_1^-(i)^{a+\gamma} p_2^-(i)^{b+\delta} \right] \right| \\
&= \frac{1}{a!b!} \left| \sum_{\gamma,\delta} \frac{(-1)^{\gamma+\delta}}{\gamma!\delta!} [m^+(a+\gamma, b+\delta) - m^-(a+\gamma, b+\delta)] \right| \\
&\le \frac{1}{a!b!} \sum_{\gamma,\delta} \frac{1}{\gamma!\delta!} \left| m^+(a+\gamma, b+\delta) - m^-(a+\gamma, b+\delta) \right|.
\end{aligned}$$

We now bound terms in the denominator of Equation 4.2. Since $p_1^+(i), p_2^+(i) \le \frac{1}{2000k_1}$ by assumption, we have $e^{k_1 \cdot p_1^+(i)}, e^{k_1 \cdot p_2^+(i)} > 0.9$ and correspondingly, $e^{k_2 \cdot p_1^-(i)}, e^{k_2 \cdot p_2^-(i)} > 0.9$, which imply that $\vec{\lambda}^+(a,b) > \frac{0.9^2}{a!b!} m^+(a,b)$ by definition of m^+, with corresponding expression holding for $\vec{\lambda}^-$ and m^-. Thus we bound terms in the denominator of Equation 4.1 as

$$\sqrt{1 + \max\{\vec{\lambda}^+(a,b), \vec{\lambda}^-(a,b)\}} \ge \frac{0.9}{\sqrt{a!b!}} \sqrt{1 + \max\{m^+(a,b), m^-(a,b)\}}.$$

Combining the bounds for the numerator and denominator, where in the second line we make use of the fact that (since $k_1 \cdot p_1^+(i)$, $k_1 \cdot p_1^-(i)$, $k_2 \cdot p_2^+(i)$, $k_2 \cdot p_2^-(i) < 1$) both m^+ and m^- are decreasing functions of their index, and where we make the variable substitutions $\mu = a + \gamma$, and $\nu = b + \delta$ in the third line, yields

$$\begin{aligned}
&\sum_{a+b>0} \frac{|\vec{\lambda}^+(a,b) - \vec{\lambda}^-(a,b)|}{\sqrt{1 + \max\{\vec{\lambda}^+(a,b), \vec{\lambda}^-(a,b)\}}} \\
&\quad \leq \sum_{a,b} \sum_{\gamma,\delta} \frac{|m^+(a+\gamma, b+\delta) - m^-(a+\gamma, b+\delta)|}{0.9\gamma!\delta!\sqrt{a!b!}\sqrt{1 + \max\{m^+(a,b), m^-(a,b)\}}} \\
&\quad \leq \sum_{a,b} \sum_{\gamma,\delta} \frac{|m^+(a+\gamma, b+\delta) - m^-(a+\gamma, b+\delta)|}{0.9\gamma!\delta!\sqrt{a!b!}\sqrt{1 + \max\{m^+(a+\gamma, b+\delta), m^-(a+\gamma, b+\delta)\}}} \\
&\quad = \sum_{\mu,\nu} \sum_{\substack{\gamma \leq \mu \\ \delta \leq \nu}} \frac{|m^+(\mu,\nu) - m^-(\mu,\nu)|}{0.9\gamma!\delta!\sqrt{(\mu-\gamma)!(\nu-\delta)!}\sqrt{1 + \max\{m^+(\mu,\nu), m^-(\mu,\nu)\}}} \\
&\quad = \sum_{\mu,\nu} \frac{|m^+(\mu,\nu) - m^-(\mu,\nu)|}{\sqrt{1 + \max\{m^+(\mu,\nu), m^-(\mu,\nu)\}}} \frac{1}{0.9} \sum_{\gamma \leq \mu} \frac{1}{\gamma!\sqrt{(\mu-\gamma)!}} \sum_{\delta \leq \nu} \frac{1}{\delta!\sqrt{(\nu-\delta)!}}.
\end{aligned}$$

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 bounded the sum of the squares of the terms, Cauchy-Schwarz bounds the original sum of these $\mu + 1$ terms as $\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}}$. We note that $\frac{\mu!}{\lfloor\frac{\mu}{2}\rfloor!^2}$ grows asymptotically as 2^μ by Sterling's formula and thus $\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}} \leq \frac{1}{\lfloor\frac{\mu}{2}\rfloor!}$ for large enough μ; evaluating for small μ we see that in fact $\sqrt{2(\mu+1)\frac{1.5^\mu}{\mu!}} \leq \frac{3}{\lfloor\frac{\mu}{2}\rfloor!}$ for all μ, which is our bound on the γ sum; consequently the "$\delta \leq \nu$" sum is bounded by $\frac{3}{\lfloor\frac{\nu}{2}\rfloor!}$, and since $\frac{1}{.9}3 \cdot 3 = 10$, the theorem follows from Lemma 4.6.7. $\square$

Corollary 4.6.10. *Given integers $k_1, k_2 > 0$, real number $\epsilon \leq \frac{1}{64 \cdot 2^{\sqrt{\log n}}}$ and two distribution pairs $(p_1^+, p_2^+), (p_1^-, p_2^-)$ where p_1^+, p_1^- have frequencies at most $\frac{\epsilon}{k_1}$ and p_2^+, p_2^- have frequencies at most $\frac{\epsilon}{k_2}$, then, letting m^+, m^- be the (k_1,k_2)-based moments of p^+, p^- respectively, if it is the case that*

$$\sum_{a,b \leq \sqrt{\log n}} \frac{|m^+(a,b) - m^-(a,b)|}{\sqrt{1 + \max\{m^+(a,b), m^-(a,b)\}}} < \frac{1}{1000}.$$

then it is impossible to test any symmetric property that is true for (p_1^+, p_2^+) and false for (p_1^-, p_2^-) in (k_1, k_2) samples.

Proof. We derive this from the bound of the 2-distribution Wishful Thinking Theorem. We note that for any distribution pairs $(p_1^+, p_2^+), (p_1^-, p_2^-)$, we have $m^+(0,0) = m^-(0,0) = n$, and $m^+(1,0) = m^-(1,0) = k_1$, $m^+(0,1) = m^-(0,1) = k_2$, 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 a, b we have $m^+(a,b) \leq k_1^a k_2^b n(\frac{\epsilon}{k_1})^a(\frac{\epsilon}{k_2})^b = n\epsilon^{a+b} \leq \frac{1}{64^{a+b}}$ Thus, since $\frac{|m^+(a,b)-m^-(a,b)|}{\sqrt{1+\max\{m^+(a,b),m^-(a,b)\}}} \leq m^+(a,b)$, we can bound these terms by $\sum_{a+b\geq 2} \frac{1}{64^{a+b}} < \frac{1}{1000}$, yielding the corollary as a consequence of Theorem 4.6.9. □

4.6.1 The Closeness Testing Lower Bound

We are now in a position to prove Theorem 1.1.1, 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.

Proof of Theorem 1.1.1. Let x, y be distributions on $[n]$ defined as follows: for $1 \leq i \leq n^{2/3}$ let $x(i) = y(i) = \frac{1}{2n^{2/3}}$. For $n/2 < i \leq 3/4n$ let $x(i) = \frac{2}{n}$; and for $3n/4 < i \leq n$ 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 $k = \frac{n^{2/3}}{1800}$. We note that each frequency defined is at most $\frac{1}{3600k}$. Let $m_{a,b}^+$ and $m_{a,b}^-$ be the (k,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, $a, b \geq 1$ and we explicitly compute $m_{a,b}^- = \frac{n^{2/3}}{3600^{a+b}}$ and $m_{a,b}^+ = \frac{n^{2/3}}{3600^{a+b}} + \frac{n}{4(900n^{1/3})^{a+b}}$, so thus

$$\sum_{a,b} \frac{|m_{a,b}^+ - m_{a,b}^-|}{\sqrt{1+\max\{m_{a,b}^+, m_{a,b}^-\}}} \leq \sum a, b \frac{|m_{a,b}^+ - m_{a,b}^-|}{\sqrt{m^- a, b}}$$

$$\leq \sum_{a,b\geq 1} \frac{\frac{n}{4(900n^{1/3})^{a+b}}}{\sqrt{\frac{n^{2/3}}{3600^{a+b}}}} = \sum_{a,b\geq 1} \frac{n^{2/3}}{4(15n^{1/3})^{a+b}}$$

$$= \frac{1}{900} \sum_{a,b\geq 0} \frac{1}{(15n^{1/3})^{a+b}}$$

$$\leq \frac{1}{900} \sum_{a,b} \frac{1}{15^{a+b}} < \frac{1}{500}.$$

Invoking the Wishful Thinking Theorem (two-distribution version) yields the desired result. □

Chapter 5

The Matching Moments Theorem

5.1 Intuition

In the previous chapter we showed essentially that moments are all that matter in the low-frequency setting. In this chapter 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 chapter.

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 (say, at least $b + \epsilon$) and distribution p^- with small value of π (say, at most $a - \epsilon$), 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 chapter 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 somewhat-low frequency distribution (we relax the constraint to frequency at most $\frac{k\delta}{n^{o(1)}}$), so that we can conclude the argument as follows: there exists $\hat{p}^+$ near p^+ and there exists $\hat{p}^-$ near p^- such that the moments of $\hat{p}^+$ and $\hat{p}^-$ are close to each other and such that both $\hat{p}^+$ and $\hat{p}^-$ have frequencies below $\frac{k\delta}{n^{o(1)}}$; 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 lies in or close to the image of each of these spheres under the moments map.

In other words, the plan for this chapter is to show how we can modify low-frequency distributions (1) slightly, (2) into somewhat-low-frequency distributions so that (3) their moments almost match $\hat{m}$. We address the single-distribution case first.

5.2 The Single Distribution Case

Recall from Chapter 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 5.2.4 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.)

5.2.1 Properties of Vandermonde Matrices

We define the particular Vandermonde matrices we use:

Definition 5.2.1. *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 5.2.2 (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

$$\frac{(-1)^{i+1}\displaystyle\sum_{\substack{1\le s_1<s_2<\ldots<s_{\mu-i}\le\mu\\ \forall q, s_q\neq j}}\prod_{q=1}^{\mu-i} z_{s_q}}{\displaystyle\prod_{q\in\{1,\ldots,\mu\}-\{j\}}(z_q-z_j)}. \tag{5.1}$$

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

Lemma 5.2.3. *Each element of $inv(V^\mu)$ has magnitude at most 6^μ.*

Proof. We bound the magnitudes of the numerator and denominator of Equation 5.1 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. □

5.2.2 Construction and Proof

We now present the construction for "matching moments".

Definition 5.2.4. *Define the function M mapping distributions p on $[n]$, positive integer $k\le n$, and real number $0<\delta\le 1$ to distribution $\hat{p}\leftarrow M_\delta^k(p)$ 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 $\sum_{i\in I}p(i)\le\delta'$. 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 $2\le a\le\mu$ let $m(a)$ be the κ-based moments of this modified distribution, 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 $\overline{c} = \overline{V}^{\mu I} \cdot \overline{m}$.*

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 5.2.5 (Matching Moments Theorem)**.** *For integers k, n and real number δ, the vector $\hat{m}^k_\delta$ and the function M of Definition 5.2.4 are such that for any distribution p for which $\forall i, p(i) \le \frac{1}{k}$, 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 $i \in [n]$, $\hat{p}(i) \le 1/\hat{k}$;*
- $|p - \hat{p}| \le \delta$
- *The $\hat{k}$-based ath moment of $\hat{p}$, for $a \le \sqrt{\log n}$ equals $\hat{m}$ to within $\frac{1}{10000 \log n}$.*

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

We note that $\overline{m}$ is indeed an upper-bound on m: we have $m(1) = \overline{m}(1) = 0$; otherwise, since $p(i) \le \frac{1}{k}$ for each i, the κ-based moments are bounded as $m(a) \le \sum_i \hat{p}(i)(\frac{1}{k})^{a-1} \cdot \kappa^a \le \frac{\kappa^2}{k} \sum_i \hat{p}(i) \le \frac{\kappa^2}{k}$, as desired. The fact that $\overline{V}^{\mu I}$ bounds the magnitudes of the elements of $inv(V^\mu)$ is Lemma 5.2.3. 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 $V^\mu_1 \cdot c$, where V^μ_1 denotes the first row of V^μ. We note that V^μ_1 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 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.[1] 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 5.2.4) this expression is at most by 1.

[1] And thus each of the new weights added in Step 5 is at least $\frac{1}{n}$, which is what we needed in Section 3.3.3.

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. □

5.3 The Two Distribution Case

5.3.1 Preliminaries

The 2-distribution case is analogous to the single distribution case, but the number of indices needed to describe each of the various objects constructed in the argument increases somewhat. As above, we start simply, by considering how the (κ_1, κ_2)-based moments (for arbitrary κ_1, κ_2) of a distribution pair p_1, p_2 change when, for arbitrary c, t, u we add c new entries to the distribution pair with value pairs $(\frac{t}{\kappa_1}, \frac{u}{\kappa_2})$, again, ignoring as above the constraint that p_1 and p_2 each sum to 1. By trivial application of the definition, we see that the (a,b) moment increases simply by $ct^a u^b$. We note that, as above, these moments depend linearly on c, so that if we wish to fix the (a,b) moments for all $a, b < \mu \equiv \sqrt{\log n}$ we need set up and solve μ^2 linear equations. The equations will specify μ^2 parameters $c_{t,u}$ where $t, u \in [\mu]$ and $c_{t,u}$ counts the number of different i for which the pair $(\frac{t}{\kappa_1}, \frac{u}{\kappa_2})$ occurs in the distribution pair as $(p_1(i), p_2(i))$.

We note that the constants $t^a u^b$ no longer constitute a Vandermonde matrix; however, we can treat them as the *tensor product* of two Vandermonde matrices. For completeness' sake we define:

Definition 5.3.1. *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 $X(a, t) \cdot Y(b, u)$.

Thus if we consider the constants $t^a u^b$ as forming a matrix with rows indexed by pairs (a, b) and columns indexed by pairs (t, u) then this matrix is exactly the tensor product of Vandermonde matrices $V^\mu \otimes V^\mu$. We invoke the standard fact that matrix inversion distributes over the tensor product to see the generalization of Lemma 5.2.3:

Lemma 5.3.2. *Each element of $inv(V^\mu \otimes V^\mu)$ has magnitude at most 36^μ.*

Proof. We have $inv(V^\mu \otimes V^\mu) = inv(V^\mu) \otimes inv(V^\mu)$. From Lemma 5.2.3 each entry of $inv(V^\mu)$ has magnitude at most 6^μ; thus the tensor product of this matrix with itself has entries bounded by the square of this, namely 36^μ. □

5.3.2 Construction

Definition 5.3.3. *Define the function M mapping distribution pairs p_1, p_2 on $[n]$, positive integers $k_1 \leq k_2 \leq n$, and real number $0 < \delta \leq 1$ to distribution pairs $\hat{p}_1, \hat{p}_2 \leftarrow M_\delta^{k_1,k_2}(p_1, p_2)$ via the following sequence of modifications to p_1, p_2:*

1. *Let $\delta' = \frac{\delta}{6}$; let I be the set of $\lfloor \delta' n \rfloor$ indices i such that $p_1(i)+p_2(i)$ is smallest. Set $\hat{p}_1, \hat{p}_2$ to the those distributions nearest to p_1, p_2 respectively such that $\forall i \in I$, $\hat{p}_1(i) = \hat{p}_2(i) = 0$, $\forall i \notin I$, $\hat{p}_1(i) \in [0, \frac{1}{k_1}]$ and $\hat{p}_2(i) \in [0, \frac{1}{k_2}]$, and $\sum_i \hat{p}_1(i) = \sum_i \hat{p}_2(i) = 1 - \delta'$.*
2. *Let $\mu = \lfloor \sqrt{\log n} \rfloor$, and let $(\kappa_1, \kappa_2) = \frac{\delta'}{12\mu^5 36^\mu}(k_1, k_2)$; for integers $2 \leq a, b \leq \mu$ let $m(a, b)$ be the (κ_1, κ_2)-based moments of the modified distribution pair, with $m(0, 0) = m(1, 0) = m(0, 1) = 0$ defined separately. Let $\hat{c} = inv(V^\mu \otimes V^\mu) \cdot m$.*
3. *Let $\overline{m}(a, b)$ be an upper-bound on m which has value 0 for (a, b) equal to $(0, 0)$, $(0, 1)$ or $(1, 0)$, value $\frac{\kappa_2^2}{k_2}$ when $a = 0, b \geq 2$ and value $\frac{\kappa_1^2}{k_1}$ otherwise. Let $\overline{V}^{\mu I}$ be a $\mu^2 \times \mu^2$ matrix with entries 36^μ, and let $\overline{c} = \overline{V}^{\mu I} \cdot \overline{m}$.*

4. *For each $t, u < \mu$ choose $c(t,u) = \lfloor \bar{c}(t,u) - \hat{c}(t,u) \rfloor$ indices $i \in I$ with $\hat{p}_1(i) = \hat{p}_2(i) = 0$ and set $\hat{p}_1(i) = \frac{t}{\kappa_1}, \hat{p}_2(i) = \frac{u}{\kappa_2}$ for these indices.*

5. *Make $\sum \hat{p}_1(i) = \sum \hat{p}_2(i) = 1$ by choosing $n\frac{\delta'}{2}$ of the unassigned indices from I and filling in those entries from $\hat{p}_1$ and $\hat{p}_2$ uniformly.*

Let $\hat{m}_\delta^{k_1,k_2}$ be the moments produced by applying this procedure to the uniform distribution.

For these $\hat{m}, M$ we have the following theorem. The proof is omitted as it contains no essentially new ideas not found in the proof of its single distribution form.

Theorem 5.3.4 (Matching Moments Theorem for Two Distributions)**.** *For integers k_1, k_2, n and real number δ, the vector $\hat{m}_\delta^{k_1,k_2}$ and the function M of Definition 5.2.4 are such that for any distribution pair p_1, p_2 for which $\forall i, p_1(i) \le \frac{1}{k_1}$ and $p_2(i) \le \frac{1}{k_2}$, letting $\hat{p}_1, \hat{p}_2 \leftarrow M_\delta^{k_1,k_2}(p_1,p_2)$ and $(\hat{k}_1, \hat{k}_2) = \frac{\delta}{10000 \cdot 2^{6\sqrt{\log n}}}(k_1, k_2)$ we have*

- *For all $i \in [n]$, $\hat{p}_1(i) \le 1/\hat{k}_1$ and $\hat{p}_2(i) \le 1/\hat{k}_2$;*
- $|p_1 - \hat{p}_1| + |p_2 - \hat{p}_2| \le \delta$
- *The $(\hat{k}_1, \hat{k}_2)$-based (a,b)th moment of the pair $(\hat{p}_1, \hat{p}_2)$, for $a, b < \sqrt{\log n}$ equals $\hat{m}$ to within $\frac{1}{10000 \log n}$.*

Chapter 6

The Canonical Testing Theorem

In this chapter we prove the main results of this thesis, the Low Frequency Blindness and Canonical Testing theorems (Theorems 3.1.2 and 3.1.3 for single distributions and 3.2.2 and 3.2.3 for distribution pairs). First we show how to combine the results of the previous two chapters 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.

6.1 The Single Distribution Case

The lower-bound we present completes the argument we have been making in the last few chapters 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 6.1.1. *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* $k \cdot \frac{\delta}{1000\cdot 2^{4\sqrt{\log n}}}$ *samples.*

Proof. Consider the distributions obtained by applying the Matching Moments Theorem (Theorem 5.2.5) 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 $\hat{k} = \frac{k\delta}{100\cdot 2^{3\sqrt{\log n}}}$; (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 4.5.7) 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 4.5.7 as

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

and thus Corollary 4.5.7 yields the desired conclusion. □

We now easily derive the full Low Frequency Blindness Theorem (Theorem 3.1.3).

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 6.1.1.

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 6.1.1 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 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 $p^-|L$ consist of frequencies at most $\frac{1}{\ell \cdot k}$, where by definition, $\pi'(p^+|L) > b$ and $\pi'(p^-|L) < a$. We thus invoke Lemma 6.1.1 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, which we carry out for the 2-distribution case in Section 6.2. $\square$

We conclude with a proof of the Canonical Testing Theorem (Theorem 3.1.2), making use of the following lemma:

Lemma 6.1.2. *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 6.1.2, 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* high-frequency elements. Since π is (ϵ, δ)-weakly-continuous, $\pi(p^+) > b$. Applying the Low Frequency Blindness Theorem to p^+, p^- yields the desired result. $\square$

6.2 The Two Distribution Case

We first generalize Lemma 6.1.1 to the case of low-frequency distribution pairs.

Lemma 6.2.1. *Given a symmetric property π on distribution pairs on $[n]$ that is (ϵ, δ)-weakly-continuous and two distribution pairs, $(p_1^+, p_2^+), (p_1^-, p_2^-)$ where p_1^+, p_1^- have frequencies at most $\frac{1}{k_1}$ and p_2^+, p_2^- have frequencies at most $\frac{1}{k_2}$ but where $\pi(p_1^+, p_2^+) > b$ and $\pi(p_1^-, p_2^-) < a$, then no tester can distinguish between $\pi > b-\epsilon$ and $\pi < a+\epsilon$ in $\frac{\delta}{640000 \cdot 2^{7\sqrt{\log n}}}(k_1, k_2)$ samples.*

Proof. Consider the distributions obtained by applying the 2-distribution Matching Moments Theorem to (p_1^+, p_2^+) and p_1^-, p_2^-: let $\hat{p}_1^+, \hat{p}_2^+ = M_\delta^{k_1,k_2}(p_1^+, p_2^+)$ and $\hat{p}_1^-, \hat{p}_2^- = M_\delta^{k_1,k_2}(p_1^-, p_2^-)$. From the Matching Moments Theorem's three conclusions we have that (1) the modified distributions have frequencies at most $(\hat{k}_1, \hat{k}_2) =$

$\frac{\delta}{10000 \cdot 2^{7\sqrt{\log n}}}(k_1, k_2)$ respectively; (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 $\pi(\hat{p}_1^+, \hat{p}_2^+) > b - \epsilon$ and $\pi(\hat{p}_1^-, \hat{p}_2^-) < a + \epsilon$; and (3) the $(\hat{k}_1, \hat{k}_2)$-based moments of $(\hat{p}_1^+, \hat{p}_2^+)$ and $(\hat{p}_1^-, \hat{p}_2^-)$ up to degree $\sqrt{\log n}$ are equal to within $\frac{2}{10000 \log n}$.

We then apply the corollary to the 2-distribution Wishful Thinking Theorem (Corollary 4.6.10) for $(k_1, k_2) = \frac{1}{64 \cdot 2^{\sqrt{\log n}}}(\hat{k}_1, \hat{k}_2)$. (The k_1, k_2 we use for the Wishful Thinking Theorem is different from the k_1, k_2 used in the previous paragraph for the Matching Moments Theorem; however, we retain $\hat{k}_1, \hat{k}_2$ from the previous paragraph.) We note that the (a, b)th (k_1, k_2)-based moment is proportional to $k_1^a k_2^b$, so since the $(\hat{k}_1, \hat{k}_2)$-based moments of $(\hat{p}_1^+, \hat{p}_2^+)$ and $(\hat{p}_1^-, \hat{p}_2^-)$ match to within $\frac{2}{10000 \log n}$ and since $k_1 < \hat{k}_1$,and $k_2 \leq \hat{k}_2$, the (k_1, k_2)-based moments also match to within this bound. We may thus evaluate the condition of Corollary 4.6.10 as

$$\begin{aligned}
\sum_{a=2}^{\sqrt{\log n}} \frac{|m^+(a,b) - m^-(a,b)|}{\sqrt{1 + \max\{m^+(a,b), m^-(a,b)\}}} &\leq \sum_{a+b \leq \sqrt{\log n}} |m^+(a,b) - m^-(a,b)| \\
&\leq \frac{2\sqrt{\log n}}{10000 \log n} < \frac{1}{1000},
\end{aligned}$$

and thus Corollary 4.6.10 yields the desired conclusion. □

We now derive the full 2-distribution Low Frequency Blindness Theorem (Theorem 3.2.3).

Proof of the Two Distribution Low Frequency Blindness Theorem. We follow the outline of the proof of the single distribution version of this theorem, as found in the previous section.

Let H be the set of indices occurring either with frequency at least $\frac{1}{k_1}$ in p_1^+ or p_1^- or with frequency at least $\frac{1}{k_2}$ in p_1^+ or p_1^-. Let $p_{1H} = p_1^-|H (= p_1^+|H)$, namely the high-frequency portion of p_1^- and p_1^+, and correspondingly let $p_{2H} = p_2^-|H$. Let $L = [n] - H$, and let $\ell_1 = |p_1^+(L)|$, namely the probability that p_1^+ (or p_1^-) draws a low-frequency index, with $\ell_2 = |p_2^+(L)| = |p_2^-(L)|$ defined correspondingly.

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_1^+, p_2^+) and (p_1^-, p_2^-). We show how a tester for π would imply a tester for π', and conclude by invoking Lemma 6.2.1 to see that neither tester can exist.

Consider the following property π' on arbitrary distribution pairs (p_{1L}, p_{2L}) with support L: define the function f mapping (p_{1L}, p_{2L}) to the distribution pair (p_1, p_2) on $[n]$ such that $p_1|H = p_{1H}$, $p_1|L = p_{1L}$, and the probability of p_1 being in L, namely $|p_1(L)|$, equals ℓ_1, with the corresponding properties holding for the second element of the pair, $p_2|H = p_{2H},\ \ p_2|L = p_{2L}$, and $|p_2(L)| = \ell_2$. Let $\pi'(p_{1L}, p_{2L}) = \pi(f(p_{1L}, p_{2L}))$.

Assume for the sake of contradiction that there exists a $(\bar{k}_1, \bar{k}_2)$-sample tester T for $\pi_{a+\epsilon}^{b-\epsilon}$ (for some $\bar{k}_1, \bar{k}_2$). By Lemma 4.6.4 we may construct the corresponding $(\bar{k}_1, \bar{k}_2)$-Poissonized tester T^p. We construct a $(\ell_1\bar{k}_1, \ell_2\bar{k}_2)$-Poissonized tester T' for $\pi'^{b-\epsilon}_{a+\epsilon}$ that processes samples from p_{1L}, p_{2L} as follows:

1. Draw integers $t_1^H \leftarrow \text{Poi}(k_1(1-\ell_1)), t_2^H \leftarrow \text{Poi}(k_2(1-\ell_2))$, and then *simulate* drawing t_1^H samples from p_{1H}, and t_2^H samples from p_{2H}.

2. Run the (Poissonized) tester T^p on all the simulated samples, plus all the authentic samples from p_{1L} and p_{2L}.

By construction, the distribution of samples input to T^p is exactly that of drawing respectively $Poi(\bar{k}_1)$ and $Poi(\bar{k}_2)$-distributed samples from the distribution pair $f(p_{1L}, p_{2L})$. Thus running T' exactly simulates running the tester T^p on the pair $f(p_{1L}, p_{2L})$, and thus since 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 $\ell_1, \ell_2 \le 1$, the (ϵ,δ)-weak-continuity of π implies the (ϵ,δ)-weak-continuity of π'. Further, we have that the two distributions $p_1^+|L,\ p_1^-|L$ all have frequencies below $\frac{1}{\ell_1 k_1}$ and the two distributions $p_2^+|L,\ p_2^-|L$ all have frequencies below $\frac{1}{\ell_2 k_2}$, where by definition, $\pi'(p^+|L) > b$ and $\pi'(p^-|L) < a$. We thus invoke Lemma 6.2.1 on $\pi', p^+|L, p^-|L$, and $(\ell_1 k_1, \ell_2 k_2)$ to conclude that no Poissonized tester can distinguish $\pi' > b-\epsilon$ from $\pi' < a+\epsilon$ in $\frac{\delta}{640000\cdot 2^{7\sqrt{\log n}}}(\ell_1 k_1, \ell_2 k_2)$ samples (from the proof of Lemma 4.6.7 we see that the lower bounds of Section 4.6 apply to Poissonized testers exactly as they do to regular testers). Since we showed in the previous paragraph that a $(\bar{k}_1, \bar{k}_2)$-sample tester for π implies an $(\ell_1\bar{k}_1, \ell_2\bar{k}_2)$-Poissonized tester for π', we conclude that no tester can distinguish $\pi > b - \epsilon$ from $\pi < a + \epsilon$ in $\frac{\delta}{640000 \cdot 2^{7\sqrt{\log n}}}(k_1, k_2)$ samples, as desired. $\square$

We conclude with a proof of the 2-distribution Canonical Testing Theorem (Theorem 3.2.2), making use of the following lemma which generalizes Lemma 6.1.2:

Lemma 6.2.2. *Given a distribution pair (p_1, p_2) and parameter θ, if we draw k_1 random samples p_1 and k_2 random samples from k_2 then with probability at least $1 - \frac{4}{n}$ the set P constructed by the Canonical Tester will include a distribution pair $(\hat{p}_1, \hat{p}_2)$ such that $|p_1 - \hat{p}_1| + |p_2 - \hat{p}_2| \leq 24\sqrt{\frac{\log n}{\theta}}$.*

Proof of the Two Distribution 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 pair (p_1, p_2) when in fact $\pi_a^b(p_1, p_2) > b + \epsilon$. From the definition of the Canonical Tester this occurs when, with probability at least $\frac{1}{3}$, the set P constructed contains a distribution pair (p_1^-, p_2^-) such that $\pi(p_1^-, p_2^-) < a$. From Lemma 6.2.2, P contains some pair (p_1^+, p_2^+) within statistical distance δ from (p_1, p_2) 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 $(p_1^-, p_2^-), (p_1^+, p_2^+)$ lying in the same P, and thus having the *same* high-frequency elements. Since π is (ϵ, δ)-weakly-continuous, $\pi(p_1^+, p_2^+) > b$. Applying the Low Frequency Blindness Theorem to (p_1^+, p_2^+) and (p_1^-, p_2^-) yields the desired result. $\square$

Bibliography

[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):549595, 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.