

אוניברסיטת תל-אביב
הפקולטה למדעים מדויקים ע"ש ריימונד וברלי סאקלר
בית הספר למדעי המחשב ע"ש בלבטניק

בדיקת תכונות של קבוצת התפלגויות

חיבור זה מוגש כחלק ממילוי הדרישות לקבלת התואר "מוסמך למדעים" (M. Sc.)
בבית הספר למדעי המחשב באוניברסיטת תל-אביב

ע"י

רעות לוי

העבודה הוכנה באוניברסיטת תל-אביב
בהדרכתן של פרופ' רונית רובינפלד ופרופ' דנה רון

תשרי ה'תשע"א

תקציר

אנו מציעים מסגרת לחקירת בדיקות תכונות של קבוצות של התפלגויות, שבה מספר ההתפלגויות בקבוצה הוא פרמטר של הבעיה. עבודות קודמות על בדיקת תכונות של התפלגויות חקרו התפלגויות בודדות או זוגות של התפלגויות. אנו מציעים שני מודלים שונים בדרך שבה לאלגוריתם ניתנת גישה לדגימות מההתפלגויות. במודל אחד האלגוריתם עשוי לבקש מדגם מהתפלגות כלשהי לבחירתו, במודל השני הבחירה של ההתפלגות היא אקראית.

ההתמקדות העיקרית שלנו היא על הבעיה הבסיסית של הבחנה בין מקרה שבו כל ההתפלגויות בקבוצה זהות (או דומות מאוד), לבין המקרה כי יש צורך לשנות את התפלגויות בקבוצה במידה שאינה זניחה על מנת להשיג את תכונה הנבדקת. אנחנו נותנים חסם עליון ותחתון צמודים כמעט לבעיה זו, וכמו כן חוקרים את ההרחבה לתכונות הקיבוציות. אחד מהחסמים התחתונים שלנו גורר ישירות חסם תחתון על בדיקת אי-תלות של התפלגות משותפת, תוצאה אשר הושארה פתוחה על ידי עבודה קודמת.

TEL-AVIV UNIVERSITY
RAYMOND AND BEVERLY SACKLER
FACULTY OF EXACT SCIENCES
BLAVATNIK SCHOOL OF COMPUTER SCIENCE

Testing Properties of Collections of Distributions

Thesis submitted in partial fulfillment of the requirements for the M.Sc. degree in the School of
Computer Science, Tel-Aviv University

by

Reut Levi

The research work for this thesis has been carried out at Tel-Aviv University
under the supervision of Prof. Dana Ron and Prof. Ronitt Rubinfeld

Oct 2010

Acknowledgements

Abstract

We propose a framework for studying property testing of collections of distributions, where the number of distributions in the collection is a parameter of the problem. Previous work on property testing of distributions considered single distributions or pairs of distributions. We suggest two models that differ in the way the algorithm is given access to samples from the distributions. In one model the algorithm may ask for a sample from any distribution of its choice, and in the other the choice of the distribution is random.

Our main focus is on the basic problem of distinguishing between the case that all the distributions in the collection are the same (or very similar), and the case that it is necessary to modify the distributions in the collection in a non-negligible manner so as to obtain this property. We give almost tight upper and lower bounds for this testing problem, as well as study an extension to a clusterability property. One of our lower bounds directly implies a lower bound on testing independence of a joint distribution, a result which was left open by previous work.

Contents

1	Introduction and Results	1
1.1	Introduction	1
1.1.1	Our Contributions	2
	The Models	2
	Testing Equivalence in the sampling model	2
	Testing Clusterability in the query model	3
	Implications of our results	4
1.1.2	Related Work	4
1.1.3	Other related work	8
1.1.4	Open Problems and Further Research	8
1.2	Preliminaries	8
2	Equivalence Testing	10
2.1	A Lower Bound of $\Omega(\mathbf{n}^{2/3}\mathbf{m}^{1/3})$ for Testing Equivalence in the Uniform Sampling Model when $\mathbf{n} = \Omega(\mathbf{m} \log \mathbf{m})$	10
2.1.1	Preliminaries concerning Poisson distributions	10
2.1.2	Testability of symmetric properties of lists of distributions	13
2.1.3	The proof of Theorem 1	16
2.1.4	A lower bound for testing Independence	24
2.2	Algorithms for Testing Equivalence in the Sampling Model	24
2.2.1	Proof of Theorem 5	25
2.2.2	Proof of Theorem 6	29
2.3	Algorithms for Testing Tolerant Equivalence in the Sampling Model	29
2.3.1	Algorithm for Testing Tolerant Identity in the Sampling Model	29
2.3.2	Algorithm for Testing Tolerant Equivalence in the Known-Weights Sampling Model	33
2.3.3	Algorithm for Testing Tolerant Equivalence in the Unknown-Weights Sampling Model	35
2.4	A Lower Bound of $\Omega(\mathbf{n}^{1/2}\mathbf{m}^{1/2})$ for Testing Equivalence in the Uniform Sampling Model	37

3 Clusterability Testing	43
3.1 Testing (k, β) -Clusterability in the Query Model	43

Chapter 1

Introduction and Results

1.1 Introduction

In recent years, several works have investigated the problem of testing various properties of data that is most naturally thought of as samples of an unknown distribution. More specifically, the goal in testing a specific property is to distinguish the case that the samples come from a distribution that has the property from the case that the samples come from a distribution that is far (usually in terms of ℓ_1 norm, but other norms have been studied as well) from any distribution that has the property. To give just a few examples, such tasks include testing whether a distribution is uniform [GR00, Pan08] or similar to another known distribution [BFR⁺], and testing whether a joint distribution is independent [BFF⁺01]. Related tasks concern sublinear estimation of various measures of a distribution, such as its entropy [BDKR05, GMV09] or its support size [RRSS09]. Recently, general techniques have been designed to obtain nearly tight lower bounds on such testing and estimation problems [Val08a, Val08b].

These types of questions have arisen in several disparate areas, including physics [Ma81, SKSB98, NBS04], cryptography and pseudorandom number generation [Knu69], statistics [Csi67, Har75, WW95, Pan04, Pan08, Pan03], learning theory [Yam95], property testing of graphs and sequences (e.g., [GR00, CS07, KS08, NS07, RRRS07, FM08]) and streaming algorithms (e.g., [AMS99, FKS99, FS00, GMV09, CMIM03, CK04, BYJK⁺02, IM08, BO10a, BO10b, BO08, IKOS09]). In these works, there has been significant focus on properties of distributions over very large domains, where standard statistical techniques based on learning an approximation of the distribution may be very inefficient.

In this work we consider the setting in which one receives data which is most naturally thought of as samples of *several* distributions, for example, when studying purchase patterns in several geographic locations, or the behavior of linguistic data among varied text sources. Such data could also be generated when samples of the distributions come from various sensors that are each part of a large sensor-net. In these examples, it may be reasonable to assume that the number of such distributions might be quite large, even on the order of a thousand or more. However, for the most part, previous research has considered properties

of at most two distributions [BFR⁺10, Val08a]. We propose new models of property testing that apply to properties of several distributions. We then consider the complexity of testing properties within these models, beginning with properties that we view as basic and expect to be useful in constructing building blocks for future work. We focus on quantifying the dependence of the sample complexities of the testing algorithms in terms of the number of distributions that are being considered, as well as the size of the domain of the distributions.

1.1.1 Our Contributions

The Models

We begin by proposing two models that describe possible access patterns to multiple distributions D_1, \dots, D_m over the same domain $[n]$. In these models there is no explicit description of the distribution – the algorithm is only given access to the distributions via samples. In the first model, referred to as the *sampling model*, at each time step, the algorithm receives a pair of the form (i, j) where $i \in [n]$ is distributed according to D_j and j is selected uniformly in $[m]$. In the second model, referred to as the *query model*, at each time step, the algorithm is allowed to specify $j \in [m]$ and receives i that is distributed according to D_j . It is immediate that any algorithm in the sampling model can also be used in the query model. On the other hand, as is implied by our results, there are property testing problems which have a significantly larger sample complexity in the sampling model than in the query model.

In both models the task is to distinguish between the case that the tested distributions have the property and the case that they are ϵ -far from having the property, for a given distance parameter ϵ . Distance to the property is measured in terms of the average ℓ_1 -distance between the tested distributions and the closest collection of distributions that have the property. In all of our results, the dependence of the algorithms on the distance parameter ϵ is (inverse) polynomial. Hence, for the sake of succinctness, in all that follows we do not mention this dependence explicitly. We note that the sampling model can be extended to allow the choice of the distribution (that is, the index j) to be non-uniform (i.e., be determined by a weight w_j) and the distance measure is adapted accordingly.

Testing Equivalence in the sampling model

One of the first properties of distributions studied in the property testing model is that of determining whether two distributions over domain $[n]$ are identical (alternatively, very close) or far (according to the ℓ_1 -distance). In [BFR⁺10], an algorithm is given that uses $\tilde{O}(n^{2/3})$ samples and distinguishes between the case that the two distributions are ϵ -far and the case that they are $O(\epsilon/\sqrt{n})$ -close. This algorithm has been shown to be nearly tight (in terms of the dependence on n) by Valiant [Val08b]. Valiant also shows that in order to distinguish between the case that the distributions are ϵ -far and the case that they are β -close, for two constants ϵ and β , requires almost linear dependence on n .

Our main focus is on a natural generalization, which we refer to as the *equivalence property* of distributions D_1, \dots, D_m , in which the goal of the tester is to distinguish the case in which there is a distribution

D^* for which $\frac{1}{m} \sum_{i=1}^m \|D_i - D^*\|_1 \leq \text{poly}(\epsilon)/\sqrt{n}$, from the case in which there is no distribution D^* for which $\frac{1}{m} \sum_{i=1}^m \|D_i - D^*\|_1 \leq \epsilon$. To solve this problem in the (uniform) sampling model with sample complexity $\tilde{O}(n^{2/3}m)$ (which ensures with high probability that each distribution is sampled $\tilde{\Omega}(n^{2/3} \log m)$ times), one can make $m - 1$ calls to the algorithm of [BFR⁺10] to check that every distribution is close to D_1 .

OUR ALGORITHMS. We show that one can get a better sample complexity dependence on m . Specifically, we give two algorithms, one with sample complexity $\tilde{O}(n^{2/3}m^{1/3} + m)$ and the other with sample complexity $\tilde{O}(n^{1/2}m^{1/2} + n)$. The first result in fact holds for the case that for each sample pair (i, j) , the distribution D_j (which generated i) is not selected necessarily uniformly, and furthermore, it is unknown according to what weight it is selected. The second result holds for the case where the selection is non-uniform, but the weights are known. Moreover, the second result extends to the case in which it is desired that the tester pass distributions that are close for each element, to within a multiplicative factor of $(1 \pm \epsilon/c)$ for some constant $c > 1$, and for sufficiently large frequencies. Thus, starting from the known result for $m = 2$, as long as $n \geq m$, the complexity grows as $\tilde{O}(n^{2/3}m^{1/3} + m) = \tilde{O}(n^{2/3}m^{1/3})$, and once $m \geq n$, the complexity is $\tilde{O}(n^{1/2}m^{1/2} + n) = \tilde{O}(n^{1/2}m^{1/2})$ (which is lower than the former expression when $m \geq n$).

Both of our algorithms build on the close relation between testing equivalence and testing independence of a joint distribution over $[n] \times [m]$ which was studied in [BFF⁺01]. The $\tilde{O}(n^{2/3}m^{1/3} + m)$ algorithm follows from [BFF⁺01] after we fill in a certain gap in the analysis of their algorithm due to an imprecision of a claim given in [BFR⁺10]. The $\tilde{O}(n^{1/2}m^{1/2} + n)$ algorithm exploits the fact that j is selected uniformly (or, more generally, according to a known weight w_j) to improve on the $\tilde{O}(n^{2/3}m^{1/3} + m)$ algorithm (in the case that $m \geq n$).

ALMOST MATCHING LOWER BOUNDS. We show that the behavior of the upper bound on the sample complexity of the problem is not just an artifact of our algorithms, but rather (almost) captures the complexity of the problem. Namely, we give almost matching lower bounds of $\Omega(n^{2/3}m^{1/3})$ for $n = \Omega(m \log m)$ and $\Omega(n^{1/2}m^{1/2})$ (for every n and m). The latter lower bound can be viewed as a generalization of a lower bound given in [BFR⁺10], but the analysis is somewhat more subtle.

Our lower bound of $\Omega(n^{2/3}m^{1/3})$ consists of two parts. The first is a general theorem concerning testing symmetric properties of collections of distributions. This theorem extends a central lemma of Valiant [Val08b] on which he builds his lower bounds, and in particular the lower bound of $\Omega(n^{2/3})$ for testing whether two distributions are identical or far from each other (i.e., the case of equivalence for $m = 2$). The second part is a construction of two collections of distributions to which the theorem is applied (where the construction is based on the one proposed in [BFF⁺01] for testing independence). As in [Val08b], the lower bound is shown by focusing on the similarity between the typical collision statistics of a family of collections of distributions that have the property and a family of collections of distributions that are far from having the property. However, since many more types of collisions are expected to occur in the case of collections of distributions, our proof outline is more intricate and requires new ways of upper bounding

the probabilities of certain types of events.

Testing Clusterability in the query model

The second property that we consider is a natural generalization of the equivalence property. Namely, we ask whether the distributions can be partitioned into at most k subsets (clusters), such that within in cluster the distance between every two distributions is (very) small. We study this property in the query model, and give an algorithm whose complexity does not depend on the number of distributions and for which the dependence on n is $\tilde{O}(n^{2/3})$. The dependence on k is almost linear. The algorithm works by combining the diameter clustering algorithm of [ADPR03] (for points in a general metric space where the algorithm has access to the corresponding distance matrix) with the closeness of distributions tester of [BFR⁺10]. Note that the results of [Val08b] imply that this is tight to within polylogarithmic factors in n .

Implications of our results

As noted previously, in the course of proving the lower bound of $\Omega(n^{2/3}m^{1/3})$ for the equivalence property, we prove a general theorem concerning testability of symmetric properties of collections of distributions (which extends a lemma in [Val08b]). This theorem may have applications to proving other lower bounds on collections of distributions. Further byproducts of our research regard the sample complexity of testing whether a joint distribution is independent. More precisely, the following question is considered in [BFR⁺10]: Let Q be a distribution over pairs of elements drawn from $[n] \times [m]$ (without loss of generality, assume $n \geq m$); what is the sample complexity in terms of m and n required to distinguish independent joint distributions, from those that are far from the nearest independent joint distribution (in term of ℓ_1 distance)? The lower bound claimed in [BFF⁺01], contains a known gap in the proof. Similar gaps in the lower bounds of [BFR⁺10] for testing the closeness of distributions and of [BDKR05] for estimating the entropy of a distribution were settled by the work of [Val08b], which applies to symmetric properties. Since independence is not a symmetric property, the work of [Val08b] cannot be directly applied here. In this work, we show that the lower bound of $\Omega(n^{2/3}m^{1/3})$ indeed holds. Furthermore, by the aforementioned correction of the upper bound of $\tilde{O}(n^{2/3}m^{1/3})$ from [BFF⁺01], we get nearly tight bounds on the complexity of testing independence.

1.1.2 Related Work

In all the following algorithms, unless otherwise stated, the distance measure is the ℓ_1 -norm, the dependence on $1/\epsilon$, where ϵ is a distance parameter, is polynomial and all the lower bounds are for constant ϵ .

Distance Approximation

There are two models which have been studied. In both models the algorithm is given a pair of distributions \mathbf{p} and \mathbf{q} and distance parameters $0 \leq \alpha < \beta$. The algorithm should accept (with high probability) if \mathbf{p} and

\mathbf{q} are α -close, and should reject (with high probability) if the \mathbf{p} and \mathbf{q} are β -far (where the distance measure is part of the problem definition). When $\alpha > 0$ we refer to the tester as a *tolerant tester*.

In the setting of Identity Testing, the algorithm is given sample access to \mathbf{p} and an explicit description of \mathbf{q} . A special case of Identity testing is Uniformity Testing, where the fixed distribution, \mathbf{q} , is the uniform distribution. Goldreich and Ron [GR00] studies Uniformity Testing in the context of approximating graph expansion. They show that counting pairwise collisions in a sample is closely related to approximating the ℓ_2 -norm of the probability distribution from which the sample was drawn from. They observe that though the pairwise collisions are not entirely pairwise independent, there is still enough “independence” in the sample to give a good bound on the variance of the variable which counts the number of collisions. Using Chebyshev’s inequality they prove that their algorithm, which takes a sublinear sample of size $\tilde{O}(n^{1/2+\gamma})$ from \mathbf{p} , approximates $\|\mathbf{p}\|_2^2$ within a factor of $1 \pm (n^{-\gamma/2}/4)$. Approximating $\|\mathbf{p}\|_2^2$ allows them to test Uniformity (for $\alpha = 0$), since $\|\mathbf{p}\|_2^2$ is minimal when \mathbf{p} is uniform. Batu et al. [BFR⁺] note that running the [GR00] algorithm with $\tilde{O}(\sqrt{n})$ samples yields an algorithm for uniformity testing in the ℓ_1 -norm for $\alpha = 0$. Paninski [Pan08] gives an optimal algorithm in this setting that takes a sample of size $O(\sqrt{n}/\beta^2)$ and proves a matching lower bound of $\Omega(\sqrt{n}/\beta^2)$. Paninski proves that the expected number of elements with only one occurrence in the sample is much higher when \mathbf{p} is uniform than in any β -far from uniform distribution. Bounding the corresponding variance via the Efron-Stein inequality, allows him to apply Chebyshev’s inequality and to obtain the desired result. Valiant [Val08a] shows that a tolerant tester for uniformity (for constant α) would require $n^{1-o(1)}$ samples.

A general algorithm for Identity Testing was developed by Batu et al. [BFF⁺01] (for the case that $\alpha = \tilde{\Omega}(\beta^3/\sqrt{n})$) as a tool for testing independence of two random variables. They show an algorithm that takes only $\tilde{O}(\sqrt{n})$ samples and that (almost) matches the lower bound of $\Omega(\sqrt{n})$ (the lower bound is obtained by the fact that Uniformity Testing is a special case of Identity Testing). Their algorithm partitions the domain into subsets in which \mathbf{q} is almost uniform on, and then rejects if the ℓ_2 -norm of \mathbf{p} restricted to each such subset is too big.

In the setting of Closeness Testing, the algorithm is given sample access to \mathbf{p} and \mathbf{q} . In this setting Batu et al. [BFR⁺] give an algorithm that tests (for $\alpha = \beta/2$) whether a pair of distributions, \mathbf{p} and \mathbf{q} , is close in the ℓ_2 -norm by counting the number of collisions between a set of samples from \mathbf{p} and a set of samples from \mathbf{q} . The sample complexity of their algorithm depends only on $1/\beta$. For the ℓ_1 -norm distance measure, they give an algorithm for $\alpha = \beta/(4\sqrt{n})$. They achieve sample complexity of $\tilde{O}(n^{2/3})$ by handling heavy (with high frequency) and light (with low frequency) elements separately. Valiant [Val08a] shows that their algorithm is essentially optimal. Specifically, he proves a lower bound of $\Omega(n^{2/3})$ based on two families of pairs of distributions suggested in Batu et al. [BFR⁺]. He shows that any algorithm that takes $o(n^{2/3})$ samples can’t distinguish between the two families. In the first family all pairs of distributions are identical. In each pair of distributions there are heavy elements and light elements. In the second family, in each pair of distributions, the distributions are identical on the heavy elements and different on the light elements. The fact that it is much more likely for a heavy element to be sampled more than once (than a light element)

makes it hard to distinguish between the two families.

Independence Testing

In the setting of Independence Testing, the algorithm is given samples drawn from a joint distribution over $[n] \times [m]$ of two random variables. The algorithm should accept (with high probability) if the random variables are independent, and should reject (with high probability) if the random variables are far from being independent. Batu et al. [BFF⁺01] studies the problem of Independence Testing. They observe that Independence Testing is equivalent to Closeness Testing between the joint distribution and the product distribution of the marginal distributions. This alone provides them with an upper bound of $\tilde{O}(n^{2/3}m^{2/3})$. In order to reduce the complexity to $\tilde{O}(n^{2/3}m^{1/3})$ for $n > m$ (which is essentially optimal, as we show in this thesis), they obtain a multiplicative approximation of the marginal distributions and get a good approximation for the heavy elements in $[n]$ and for all the elements in $[m]$. The domain restricted to the heavy elements is divided to almost uniform subsets of elements. Every subset of $[n]$ and subset of $[m]$ correspond to a rectangle in $[n] \times [m]$. Each rectangle is tested by uniformity testing using a filtering scheme to correct for the fact that the matching product distribution is not entirely uniform. The closeness between the distribution restricted to the light elements and the product distribution is tested by a Closeness tester that achieves better complexity as a function of the bound on the L_∞ -norm of the tested distributions.

Entropy Approximation

Batu et al. [BDKR05] studies Entropy Approximation. They give an algorithm that for every $\gamma > 1$ and every distribution that has $\Omega(\gamma/\eta)$ entropy, takes $O(n^{1+\eta/\gamma^2} \log n)$ samples, where $\eta > 0$ is an arbitrarily small constant, and returns a γ -multiplicative approximation of the entropy, i.e. if x is the entropy and y is the output of the algorithm then: $x/\gamma \leq y \leq \gamma x$. They prove a lower bound of $\Omega(n^{1/2\gamma^2})$ samples for a class of distributions that their algorithm applies to. This proves that accepting distributions with entropy less than α and rejecting distributions with entropy greater than β , for sufficiently large $0 < \alpha < \beta < 1$, can be done in $n^{\alpha/\beta+o(1)}$ samples. Valiant [Val08a] shows a nearly matching lower bound of $n^{\alpha/\beta-o(1)}$ for this problem. Batu et al. [BDKR05] also show that it is impossible to approximate small entropy via sampling. This leads to the “combined oracle” where both samples and direct access to the probability function are allowed. In the combined oracle model they give an algorithm that for every $\gamma > 1$ and every distribution that has $\Omega(h)$ entropy, takes $O(\frac{\gamma^2 \log^2 n}{h^2(\gamma-1)^2})$ queries and obtains a γ -multiplicative approximation of the entropy and a lower bound of $\Omega(\frac{\log n}{h(\gamma^2-1)})$ queries. Guha et al. [GMV09] apply the same concept introduced in [BDKR05] and give an improved algorithm in the combined oracle model that matches the lower bound in [BDKR05].

Monotonicity and Unimodality Testing

Batu et al. [BKR04] studies Monotonicity and Unimodality Testing of a distribution over a totally ordered set. They observe that a monotone distribution can't be divided into many intervals all which are far from

uniform. Their algorithm takes sample of size $\tilde{O}(\sqrt{n})$ and bisects the domain recursively until all intervals are close to uniform or until there are too many intervals. They prove a lower bound of $\Omega(\sqrt{n})$ by reducing Testing Uniformity of a distribution to testing both increasing-monotonicity and decreasing-monotonicity of the distribution. Bhattacharyya et al. [BFRV10] studied Monotonicity Testing of a partially ordered set (poset). They prove a lower bound of $n^{1-o(1)}$ on a specific family of posets and prove that this result extends to other families of posets. Though monotonicity is a non-symmetric property, in the proof of their lower bound they use techniques developed in [Val08a] for symmetric properties of distributions.

Testing Properties of Monotone Distributions

Some properties can be tested more efficiently when the underlying distribution(s) is (are) monotone. Rubinfeld and Servedio [RS04] show that approximating the entropy for monotone distributions over partially ordered set can be done by taking $\text{poly}(\log n)$ samples, which is exponentially fewer than in the case that the distribution is not known to be monotone. They give an algorithm for testing whether an unknown monotone distribution over n -dimensional Boolean cube is uniform or far from uniform that uses $\tilde{O}(n)$ samples, which is logarithmic in the size of the domain, and prove that is optimal up to $\text{poly}(\log n)$ factors. They define the bias of an element of the Boolean cube to be the sum of its bits. The expected bias of elements drawn from the uniform distribution is 0 while the bias of elements drawn from a distribution that is ϵ -far from the uniform distribution is at least ϵ . Thus approximating the bias allows them to test Uniformity. Batu et al. [BKR04] give exponentially faster (than in the case that monotonicity is not guaranteed) algorithms for Closeness Testing, Independence Testing and entropy approximation of monotone distributions over a totally ordered set that take only $\text{poly}(\log n)$ samples.

Distribution Support Size Approximation

Raskhodnikova et al. [RRSS09] studies Distribution Support Size Approximation. When each element in the distribution appears with probability at least $1/n$ they prove a lower bound of $n^{1-o(1)}$ samples for additive error approximation of the distribution support size. They prove that the lower bound can be derived from a lower bound for the “Distinct Elements” problem. In the proof of the lower bound for “Distinct Elements” they define “frequency variables” to be the number of times a color appears in the string when the color is chosen uniformly. They prove that if they come up with two frequency variables that have ℓ proportional moments then any tester that doesn’t look at the statistics of k -collisions for $k > \ell$ can’t distinguish between the two strings that corresponds to the two frequency variables. They prove their lower bound by constructing such frequency variables that have very different expectations which translate to different number of distinct elements. The lower bound on the distribution support size has implications also for the problem of approximating the compressibility of a string according to the Lempel-Zip Compression scheme. This is due to Raskhodnikova et al. [RRRS07] that prove that Distinct Elements approximation is reducible to Lempel-Zip compressibility approximation.

Symmetric Properties

Valiant [Val08a] studies the general problem of testing any symmetric property. Valiant gives a characterization of the optimal tester for a large class of properties, namely the symmetric properties that satisfy a continuity condition. This tester is referred to as the “Canonical Tester”. The sample complexity depends on the specific property being tested. The optimality of the Canonical Tester in terms of sample complexity allows one to prove general lower bounds (as mentioned above) by proving lower bounds on the sample complexity of the Canonical Tester.

1.1.3 Other related work

Other works on testing and estimating properties of (single or pairs of) distributions include [Bat01, AAK⁺07, RX10, BNNR09, ACS10, AIOR09].

1.1.4 Open Problems and Further Research

There are many interesting directions to pursue concerning the testing of properties of collections of distributions, and because of the applicability of the model to a wide range of circumstances, we expect that new directions will present themselves. Here we give a few examples: One natural extension of our results is to give algorithms for testing the property of clusterability for $k > 1$ in the sampling model. One may also consider testing properties of collections of distributions that are defined by certain measures of distributions, and may be less sensitive to the exact form of the distributions. For example, a very basic measure is the mean (expected value) of the distribution, when we view the domain $[n]$ as integers instead of element names, or when we consider other domains. Given this measure, we may consider testing whether the distributions all have similar means (or whether they should be modified significantly so that this holds). It is not hard to verify that this property can be quite easily tested in the query model by selecting $\Theta(1/\epsilon)$ distributions uniformly and estimating the mean of each. On the other hand, in the sampling model an $\Omega(\sqrt{m})$ lower bound is quite immediate even for $n = 2$ (and a constant ϵ). We are currently investigating whether the complexity of this problem (in the sampling model) is in fact higher, and it would be interesting to consider other measures as well.

1.2 Preliminaries

Let $[n] \stackrel{\text{def}}{=} \{1, \dots, n\}$, and let $\mathcal{D} = (D_1, \dots, D_m)$ be a list of m distributions, where $D_j : [n] \rightarrow [0, 1]$ and $\sum_{i=1}^n D_j(i) = 1$ for every $1 \leq j \leq m$. For a vector $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$, let $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$ denote the ℓ_1 norm of the vector v .

For a property \mathcal{P} of lists of distributions and $0 \leq \epsilon \leq 1$, we say that \mathcal{D} is ϵ -far from (having) \mathcal{P} if $\frac{1}{m} \sum_{j=1}^m \|D_j - D_j^*\|_1 > \epsilon$ for every list $\mathcal{D}^* = (D_1^*, \dots, D_m^*)$ that has the property \mathcal{P} (note that $\|D_j - D_j^*\|_1$ is twice the the statistical distance between the two distributions).

Given a distance parameter ϵ , a *testing algorithm* for a property \mathcal{P} should distinguish between the case that \mathcal{D} has the property \mathcal{P} and the case that it is ϵ -far from \mathcal{P} . We consider two models within which this task is performed.

1. **The Query Model.** In this model the testing algorithm may indicate an index $1 \leq j \leq m$ of its choice and it gets a sample i distributed according to $D_j(i)$.
2. **The Sampling Model.** In this model the algorithm cannot select (“query”) a distribution of its choice. Rather, it may obtain a pair (i, j) where j is selected uniformly (we refer to this as the *Uniform* sampling model) and i is distributed according to $D_j(i)$.

We also consider a generalization in which there is an underlying weight vector $\mathbf{w} = (w_1, \dots, w_m)$ (where $\sum_{j=1}^m w_j = 1$), and the distribution D_j is selected according to \mathbf{w} . In this case the notion of ϵ -far needs to be modified accordingly. Namely, we say that \mathcal{D} is ϵ -far from \mathcal{P} *with respect to* \mathbf{w} if $\sum_{j=1}^m w_j \cdot \|D_j - D_j^*\|_1 > \epsilon$ for every list $\mathcal{D}^* = (D_1^*, \dots, D_m^*)$ that has the property \mathcal{P} .

We consider two variants of this non-uniform model: The *Known-Weights* sampling model, in which \mathbf{w} is known to the algorithm, and the *Unknown-Weights* sampling model in which \mathbf{w} is known.

A main focus of this work is on the following property. We shall say that a list $\mathcal{D} = (D_1 \dots D_m)$ of m distributions over $[n]$ belongs to $\mathcal{P}_{m,n}^{\text{eq}}$ (or has the property $\mathcal{P}_{m,n}^{\text{eq}}$) if $D_j = D_{j'}$ for all $1 \leq j, j' \leq m$.

Chapter 2

Equivalence Testing

2.1 A Lower Bound of $\Omega(n^{2/3}m^{1/3})$ for Testing Equivalence in the Uniform Sampling Model when $n = \Omega(m \log m)$

In this section we prove the following theorem:

Theorem 1 *Any testing algorithm for the property $\mathcal{P}_{m,n}^{\text{eq}}$ in the uniform sampling model for every $\epsilon \leq 1/20$ and for $n > cm \log m$ where c is some sufficiently large constant, requires $\Omega(n^{2/3}m^{1/3})$ samples.*

The proof of Theorem 1 consists of two parts. The first is a general theorem (Theorem 2) concerning testing symmetric properties of lists of distributions. This theorem extends a lemma of Valiant [Val08b, Lem. 4.5.4] (which leads to what Valiant refers to as the “Wishful Thinking Theorem”). The second part is a construction of two lists of distributions to which Theorem 2 is applied. Our analysis uses a technique called *Poissonization* [Szp01] (which was used in the past in the context of lower bounds for testing and estimating properties of distributions in [RRSS09, Val08a, Val08b]), and hence we first introduce some preliminaries concerning Poisson distributions.

2.1.1 Preliminaries concerning Poisson distributions

For a positive real number λ , the Poisson distribution $\text{poi}(\lambda)$ takes the value $x \in \mathbb{N}$ (where $\mathbb{N} = \{0, 1, 2, \dots\}$) with probability $\text{poi}(x; \lambda) = e^{-\lambda} \lambda^x / x!$. The expectation and variance of $\text{poi}(\lambda)$ are both λ . For λ_1 and λ_2 we shall use the following bound on the ℓ_1 distance between the corresponding Poisson distributions (for a proof see for example [RRSS09, Claim A.2]):

$$\|\text{poi}(\lambda_1) - \text{poi}(\lambda_2)\|_1 \leq 2|\lambda_1 - \lambda_2|. \quad (2.1)$$

For a vector $\vec{\lambda} = (\lambda_1, \dots, \lambda_d)$ of positive real numbers, the corresponding *multivariate* Poisson distribution $\text{poi}(\vec{\lambda})$ is the product distribution $\text{poi}(\lambda_1) \times \dots \times \text{poi}(\lambda_d)$. That is, $\text{poi}(\vec{\lambda})$ assigns each vector $\vec{x} = x_1 \dots, x_d \in \mathbb{N}^d$ the probability $\prod_{i=1}^d \text{poi}(x_i; \lambda_i)$.

We shall sometimes consider vectors $\vec{\lambda}$ whose coordinates are indexed by vectors $\vec{a} = (a_1, \dots, a_m) \in \mathbb{N}^m$, and will use $\vec{\lambda}(\vec{a})$ to denote the coordinate of $\vec{\lambda}$ that corresponds to \vec{a} . Thus, $\text{poi}(\vec{\lambda}(\vec{a}))$ is a univariate Poisson distribution. With a slight abuse of notation, for a subset $I \subseteq [d]$ (or $I \subseteq \mathbb{N}^m$), we let $\text{poi}(\vec{\lambda}(I))$ denote the multivariate Poisson distributions restricted to the coordinates of $\vec{\lambda}$ in I .

For any two d -dimensional vectors $\vec{\lambda}^+ = (\lambda_1^+, \dots, \lambda_d^+)$ and $\vec{\lambda}^- = (\lambda_1^-, \dots, \lambda_d^-)$ of positive real values, we get from the proof of [Val08b, Lemma 4.5.3] that,

$$\left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 \leq \sum_{j=1}^d \left\| \text{poi}(\lambda_j^+) - \text{poi}(\lambda_j^-) \right\|_1, \quad (2.2)$$

for our purposes we shall use the following generalized lemma.

Lemma 1 *For any two d -dimensional vectors $\vec{\lambda}^+ = (\lambda_1^+, \dots, \lambda_d^+)$ and $\vec{\lambda}^- = (\lambda_1^-, \dots, \lambda_d^-)$ of positive real values, and for any partition $\{I_i\}_{i=1}^\ell$ of $[d]$,*

$$\left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 \leq \sum_{i=1}^\ell \left\| \text{poi}(\vec{\lambda}^+(I_i)) - \text{poi}(\vec{\lambda}^-(I_i)) \right\|_1. \quad (2.3)$$

Proof: Let $\{I_i\}_{i=1}^\ell$ be a partition of $[d]$, by the triangle inequality we have that for every $k \in [\ell]$,

$$\begin{aligned} & \left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 \\ &= \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left| \text{poi}(i_1, \dots, i_d; \vec{\lambda}^+) - \text{poi}(i_1, \dots, i_d; \vec{\lambda}^-) \right| \end{aligned} \quad (2.4)$$

$$= \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left| \prod_{j \in [d]} \text{poi}(i_j; \lambda_j^+) - \prod_{j \in [d]} \text{poi}(i_j; \lambda_j^-) \right| \quad (2.5)$$

$$\leq \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left| \prod_{j \in [d]} \text{poi}(i_j; \lambda_j^+) - \prod_{j \in [d] \setminus I_k} \text{poi}(i_j; \lambda_j^+) \cdot \prod_{j \in I_k} \text{poi}(i_j; \lambda_j^-) \right| \quad (2.6)$$

$$+ \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left| \prod_{j \in [d] \setminus I_k} \text{poi}(i_j; \lambda_j^+) \cdot \prod_{j \in I_k} \text{poi}(i_j; \lambda_j^-) - \prod_{j \in [d]} \text{poi}(i_j; \lambda_j^-) \right| \quad (2.7)$$

$$= \left\| \text{poi}(\vec{\lambda}^+(I_k)) - \text{poi}(\vec{\lambda}^-(I_k)) \right\|_1 + \left\| \text{poi}(\vec{\lambda}^+([d] \setminus I_k)) - \text{poi}(\vec{\lambda}^-([d] \setminus I_k)) \right\|_1. \quad (2.8)$$

Thus, the lemma follows by induction on ℓ . ■

Lemma 2 *For any two d -dimensional vectors $\vec{\lambda}^+ = (\lambda_1^+, \dots, \lambda_d^+)$ and $\vec{\lambda}^- = (\lambda_1^-, \dots, \lambda_d^-)$ of positive real values,*

$$\left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 \leq 2 \sqrt{2 \sum_{j=1}^d \frac{(\lambda_j^- - \lambda_j^+)^2}{\lambda_j^-}}. \quad (2.9)$$

Proof: In order to prove the lemma we shall use the *KL-divergence* between distributions. Namely, for two distributions p_1 and p_2 over a domain X , $D_{\text{KL}}(p_1 \| p_2) \stackrel{\text{def}}{=} \sum_{x \in X} p_1(x) \cdot \ln \frac{p_1(x)}{p_2(x)}$. Let $\vec{\lambda}^+ = (\lambda_1^+ \dots, \lambda_d^+)$ and $\vec{\lambda}^- = (\lambda_1^- \dots, \lambda_d^-)$. Then,

$$D_{\text{KL}}(\text{poi}(\vec{\lambda}^+) \| \text{poi}(\vec{\lambda}^-)) = \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \text{poi}(i_1, \dots, i_d; \vec{\lambda}^+) \cdot \ln \frac{\text{poi}(i_1, \dots, i_d; \vec{\lambda}^+)}{\text{poi}(i_1, \dots, i_d; \vec{\lambda}^-)} \quad (2.10)$$

$$= \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left(\text{poi}(i_1, \dots, i_d; \vec{\lambda}^+) \cdot \sum_{j=1}^d \ln \left(e^{\lambda_j^- - \lambda_j^+} (\lambda_j^+ / \lambda_j^-)^{i_j} \right) \right) \quad (2.11)$$

$$= \sum_{j=1}^d \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \left(\text{poi}(i_1, \dots, i_d; \vec{\lambda}^+) \cdot \left((\lambda_j^- - \lambda_j^+) + i_j \cdot \ln(\lambda_j^+ / \lambda_j^-) \right) \right) \quad (2.12)$$

$$= \sum_{j=1}^d \sum_{i_1, \dots, i_d \in \mathbb{N}^d} \prod_{k \in [d] \setminus \{j\}} \text{poi}(i_k; \lambda_k^+) \cdot \left(\text{poi}(i_j; \lambda_j^+) \left((\lambda_j^- - \lambda_j^+) + i_j \cdot \ln(\lambda_j^+ / \lambda_j^-) \right) \right) \quad (2.13)$$

$$= \sum_{j=1}^d \sum_{i_j \in \mathbb{N}} \left(\text{poi}(i_j; \lambda_j^+) \cdot \left((\lambda_j^- - \lambda_j^+) + i_j \cdot \ln(\lambda_j^+ / \lambda_j^-) \right) \right) \quad (2.14)$$

$$= \sum_{j=1}^d \left((\lambda_j^- - \lambda_j^+) + \lambda_j^+ \cdot \ln(\lambda_j^+ / \lambda_j^-) \right) \quad (2.15)$$

$$\leq \sum_{j=1}^d \left((\lambda_j^- - \lambda_j^+) + \lambda_j^+ \cdot (\lambda_j^+ / \lambda_j^- - 1) \right) \quad (2.16)$$

$$= \sum_{j=1}^d \frac{(\lambda_j^- - \lambda_j^+)^2}{\lambda_j^-}, \quad (2.17)$$

where in Equation (2.16) we used the fact $\ln x \leq x - 1$ for every $x > 0$, and in Equations (2.13) and (2.14) we used the facts that $\sum_{i \in \mathbb{N}} \text{poi}(i; \lambda) = 1$ and $\sum_{i \in \mathbb{N}} \text{poi}(i; \lambda) \cdot i = \lambda$. The ℓ_1 distance is related to the KL-divergence by $\|D - D'\|_1 \leq 2\sqrt{2D_{\text{KL}}(D \| D')}$ and thus we obtain the lemma. ■

Lemma 3 *Let $X \sim \text{poi}(\lambda)$, then, $\Pr[X < \lambda/2] < (3/4)^{\lambda/4}$.*

Proof: Consider the matching between j and $j + \lambda/2$ for every $j = 0, \dots, \lambda/2 - 1$. We consider the ratio

between $\text{poi}(j; \lambda)$ and $\text{poi}(j + \lambda/2; \lambda)$:

$$\frac{\text{poi}(j + \lambda/2; \lambda)}{\text{poi}(j; \lambda)} = \frac{e^{-\lambda} \cdot \lambda^{j+\lambda/2}/(j + \lambda/2)!}{e^{-\lambda} \cdot \lambda^j/j!} \quad (2.18)$$

$$= \frac{\lambda^{\lambda/2}}{(j + \lambda/2)(j + \lambda/2 - 1) \cdots (j + 1)} \quad (2.19)$$

$$= \frac{\lambda}{j + \lambda/2} \cdot \frac{\lambda}{j + \lambda/2 - 1} \cdots \frac{\lambda}{j + 1} \quad (2.20)$$

$$\geq \frac{\lambda}{\lambda - 1} \cdot \frac{\lambda}{\lambda - 2} \cdots \frac{\lambda}{\lambda/2} \quad (2.21)$$

$$> \left(\frac{\lambda}{(3/4)\lambda} \right)^{\lambda/4} \quad (2.22)$$

$$= (4/3)^{\lambda/4} \quad (2.23)$$

This implies that

$$\Pr[X < \lambda/2] = \frac{\Pr[X < \lambda/2]}{\Pr[\lambda/2 \leq X < \lambda]} \cdot \Pr[\lambda/2 \leq X < \lambda] < \frac{\Pr[X < \lambda/2]}{\Pr[\lambda/2 \leq X < \lambda]} < (3/4)^{\lambda/4}, \quad (2.24)$$

and the proof is completed. ■

The next two notations will play an important technical role in our analysis. For a list of distributions $\mathcal{D} = (D_1 \dots D_m)$, an integer κ and a vector $\vec{a} = (a_1, \dots, a_m) \in \mathbb{N}^m$, let

$$p^{\mathcal{D}, \kappa}(i; \vec{a}) \stackrel{\text{def}}{=} \prod_{j=1}^m \text{poi}(a_j; \kappa \cdot D_j(i)). \quad (2.25)$$

That is, for a fixed choice of a domain element $i \in [n]$, consider performing m independent trials, one for each distribution D_j , where in trial j we select a non-negative integer according to the Poisson distribution $\text{poi}(\lambda)$ for $\lambda = \kappa \cdot D_j(i)$. Then $p^{\mathcal{D}, \kappa}(i; \vec{a})$ is the probability of the joint event that we get an outcome of a_j in trial j , for each $j \in [m]$. Let $\vec{\lambda}^{\mathcal{D}, \kappa}$ be a vector whose coordinates are indexed by all $\vec{a} \in \mathbb{N}^m$, such that

$$\vec{\lambda}^{\mathcal{D}, \kappa}(\vec{a}) = \sum_{i=1}^n p^{\mathcal{D}, \kappa}(i; \vec{a}). \quad (2.26)$$

That is, $\vec{\lambda}^{\mathcal{D}, \kappa}(\vec{a})$ is the expected number of times we get the joint outcome (a_1, \dots, a_m) if we perform the probabilistic process defined above independently for every $i \in [n]$.

2.1.2 Testability of symmetric properties of lists of distributions

In this subsection we prove the following theorem (which is used to prove Theorem 1).

Theorem 2 *Let \mathcal{D}^+ and \mathcal{D}^- be two lists of m distributions over $[n]$, all of whose frequencies are at most $\frac{\delta}{\kappa \cdot m}$ where κ is some positive integer and $0 < \delta < 1$. If*

$$\left\| \text{poi}(\vec{\lambda}^{\mathcal{D}^+, \kappa}) - \text{poi}(\vec{\lambda}^{\mathcal{D}^-, \kappa}) \right\|_1 < \frac{16}{30} - \frac{352\delta}{5}, \quad (2.27)$$

then testing in the uniform sampling model any symmetric property of distributions such that \mathcal{D}^+ has the property, while \mathcal{D}^- is $\Omega(1)$ -far from having the property requires $\Omega(\kappa \cdot m)$ samples.

For an element $i \in [n]$ and a distribution $D_j, j \in [m]$, let $\alpha_{i,j}$ be the number of times the pair (i, j) appears in the sample (when the sample is selected in the uniform sampling model). Thus $(\alpha_{i,1}, \dots, \alpha_{i,m})$ is the *sample histogram* of the element i . The histogram of the elements' histograms is called the *fingerprint* of the sample. That is, the fingerprint indicates, for every $\vec{a} \in \mathbb{N}^m$, the number of elements i such that $(\alpha_{i,1}, \dots, \alpha_{i,m}) = \vec{a}$. As shown in [BFR⁺], when testing symmetric properties of distributions, it can be assumed without loss of generality that the testing algorithm is provided only with the fingerprint of the sample. Furthermore, since the number, n , of elements is fixed, it suffices to give the tester the fingerprint of the sample without the $\vec{0} = (0, \dots, 0)$ entry.

In order to prove Theorem 2, we would like to show that the distributions of the fingerprints when the sample is generated according to \mathcal{D}^+ and when it is generated according to \mathcal{D}^- are similar (for a sample size that is below the lower bound stated in the theorem). To this end we first define a slightly different process for generating the samples that involves *Poissonization* [Szp01]. The process is such that if we are able to prove a lower bound for algorithms that receive samples generated by this process, then we obtain a related lower bound for algorithms that work in the uniform sampling model. The benefit of this process is that it breaks certain dependencies (among the $a_{i,j}$'s defined above), and hence is easier to analyze. Details follow.

Definition 1 *In the Poissonized uniform sampling model with parameter κ (which we'll refer to as the κ -Poissonized model), given a list $\mathcal{D} = (D_1, \dots, D_m)$ of m distributions, a sample is generated as follows:*

- Draw $\kappa_1, \dots, \kappa_m \leftarrow \text{poi}(\kappa)$
- Return κ_j samples distributed according to D_j for each $j \in [m]$.

Lemma 4 *Assume that there exists a tester T in the uniform sampling model for a property \mathcal{P} of lists of m distributions, that takes a sample of size $s = \kappa m$ where $\kappa \geq c \log m$ for some sufficiently large constant c , and works for every $\epsilon \geq \epsilon_0$ where ϵ_0 is a constant (and whose success probability is at least $2/3$). Then there exists a tester T' for \mathcal{P} in the Poissonized uniform sampling model with parameter 4κ , that works for every $\epsilon \geq \epsilon_0$ and whose success probability is at least $\frac{19}{30}$.*

Proof: Roughly speaking, the tester T' tries to simulate T if it has a sufficiently large sample, and otherwise it guesses the answer. More precisely, let $\mathcal{D} = (D_1, \dots, D_m)$ be a list of m distributions. For each $j \in [m]$ let κ_j denote the random variable that equals the number of samples that are selected according to D_j in the uniform sampling model, when the total number of samples is κm . Thus, $\kappa_j \sim \text{Bin}(\kappa m, \frac{1}{m})$. By [AS92, Thm. A.12], for each $j \in [m]$,

$$\Pr[\kappa_j \geq 2\kappa] < (e/4)^\kappa. \quad (2.28)$$

Now consider a tester T' that receives κ'_j samples from each D_j where $\kappa'_j \sim \text{poi}(4\kappa)$. By Lemma (3), for each j we have that,

$$\Pr [\kappa'_i < 2\kappa] \leq (3/4)^\kappa \quad (2.29)$$

Suppose T' also selects $\kappa_1, \dots, \kappa_m$ as in the distribution induced by the uniform sampling model. If $\kappa'_j \geq \kappa_j$ for each j , then T' simulates T on the union of the first κ_j samples that it got for each j . Otherwise it outputs “accept” or “reject” with equal probability.

By taking a union bound over all $j \in [m]$ we get that the probability that for every $j \in [m]$ it holds that both $\kappa_j \leq 2\kappa$ and $\kappa'_j \geq 2\kappa$ (so that $\kappa'_j \geq \kappa_j$), is at least $1 - m((e/4)^\kappa + (3/4)^\kappa)$, which is greater than $\frac{4}{5}$ for $\kappa > c \log m$ and a sufficiently large constant c . Therefore, the success probability of T' is at least $\frac{4}{5} \cdot \frac{2}{3} + \frac{1}{5} \cdot \frac{1}{2} = \frac{19}{30}$, as desired. ■

Given Lemma 4 it suffices to consider samples that are generated in the Poissonized uniform sampling model. The process for generating a sample $\{\alpha_{i,1}, \dots, \alpha_{i,m}\}_{i \in [n]}$ (recall that $\alpha_{i,j}$ is the number of times that element i was selected by distribution D_j) in the κ -Poissonized model is equivalent to the following process: For each $i \in [n]$ and $j \in [m]$, independently select $\alpha_{i,j}$ according to $\text{poi}(\kappa \cdot D_j(i))$ (see [Fel67], p. 216). Thus the probability of getting a particular histogram $\vec{a}_i = (a_{i,1}, \dots, a_{i,m})$ for element i is $p^{\mathcal{D}, \kappa}(i; \vec{a}_i)$ (as defined in Equation (2.25)). We can represent the event that the histogram of element i is \vec{a}_i by a Bernoulli random vector \vec{b}_i that is indexed by all $\vec{a} \in \mathbb{N}^m$, is 1 in the coordinate corresponding to \vec{a}_i , and is 0 elsewhere. Given this representation, the fingerprint of the sample corresponds to $\sum_{i=1}^n \vec{b}_i$. In fact, we would like \vec{b}_i to be of finite dimension, so we have to consider only a finite number (sufficiently large) of possible histograms. Under this relaxation, $\vec{b}_i = (0, \dots, 0)$ would correspond to the case that the sample histogram of element i is not in the set of histograms we consider (which would be a very rare event). Roos’s theorem, stated next, shows that the distribution of the fingerprints can be approximated by a multivariate Poisson distribution. (For simplicity, the theorem is stated for vectors \vec{b}_i that are indexed directly, that is $\vec{b}_i = (b_{i,1}, \dots, b_{i,h})$.)

Theorem 3 ([Roo99]) *Let D^{S_n} be the distribution of the sum S_n of n independent Bernoulli random vectors $\vec{b}_1, \dots, \vec{b}_n$ in \mathbb{R}^h where $\Pr [\vec{b}_i = \vec{e}_\ell] = p_{i,\ell}$ and $\Pr [\vec{b}_i = (0, \dots, 0)] = 1 - \sum_{\ell=1}^h p_{i,\ell}$ (here \vec{e}_ℓ satisfies $e_{i,\ell} = 1$ and $e_{i,\ell'} = 0$ for every $\ell' \neq \ell$). Suppose we define an h -dimensional vector $\vec{\lambda} = (\lambda_1, \dots, \lambda_h)$ as follows: $\lambda_\ell = \sum_{i=1}^n p_{i,\ell}$. Then*

$$\left\| D^{S_n} - \text{poi}(\vec{\lambda}) \right\|_1 \leq \frac{88}{5} \sum_{\ell=1}^h \frac{\sum_{i=1}^n p_{i,\ell}^2}{\sum_{i=1}^n p_{i,\ell}}. \quad (2.30)$$

We next show how to obtain a bound on sums of the form given in Equation (2.30) under appropriate conditions.

Lemma 5 *Given a list $\mathcal{D} = (D_1, \dots, D_m)$ of m distributions over $[n]$ and a real number $0 < \delta \leq 1/2$*

such that for all $i \in [n]$ and for all $j \in [m]$, $D_j(i) \leq \frac{\delta}{m \cdot \kappa}$ for some integer κ , we have that

$$\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{i=1}^n p^{\mathcal{D}, \kappa}(i; \vec{a})^2}{\sum_{i=1}^n p^{\mathcal{D}, \kappa}(i; \vec{a})} \leq 2\delta. \quad (2.31)$$

Proof:

$$\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{i=1}^n p^{\mathcal{D}, \kappa}(i; \vec{a})^2}{\sum_{i=1}^n p^{\mathcal{D}, \kappa}(i; \vec{a})} \leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_i (p^{\mathcal{D}}(i; \vec{a})) \quad (2.32)$$

$$= \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_i \left(\prod_{j=1}^m \text{poi}(a_j; \kappa \cdot D_j(i)) \right) \quad (2.33)$$

$$\leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \left(\frac{\delta}{m} \right)^{a_1 + \dots + a_m} \quad (2.34)$$

$$\leq \sum_{a=1}^{\infty} m^a \left(\frac{\delta}{m} \right)^a \quad (2.35)$$

$$\leq 2\delta, \quad (2.36)$$

where the inequality in Equation (2.36) holds for $\delta \leq 1/2$ and the inequality in Equation (2.34) follows from:

$$\text{poi}(a; \kappa \cdot D_j(i)) = \frac{e^{-\kappa \cdot D_j(i)} (\kappa \cdot D_j(i))^a}{a!} \quad (2.37)$$

$$\leq (\kappa \cdot D_j(i))^a \quad (2.38)$$

$$\leq \left(\frac{\delta}{m} \right)^a, \quad (2.39)$$

and the proof is completed. ■

Proof of Theorem 2: By the first premise of the theorem, $D_j^+(i), D_j^-(i) \leq \frac{\delta}{\kappa m}$ for every $i \in [n]$ and $j \in [m]$. By Lemma 5 this implies that Equation (2.31) holds both for $\mathcal{D} = \mathcal{D}^+$ and for $\mathcal{D} = \mathcal{D}^-$. Combining this with Theorem 3 we get that the ℓ_1 distance between the fingerprint distribution when the sample is generated according to \mathcal{D}^+ (in the κ -Poissonized model, see Definition 1) and the distribution $\text{poi}(\vec{\lambda}^{\mathcal{D}^+, \kappa})$ is at most $\frac{88}{5} \cdot 2\delta = \frac{176}{5}\delta$, and an analogous statement holds for \mathcal{D}^- . By applying the premise in Equation (2.27) (concerning the ℓ_1 distance between $\text{poi}(\vec{\lambda}^{\mathcal{D}^+, \kappa})$ and $\text{poi}(\vec{\lambda}^{\mathcal{D}^-, \kappa})$) and the triangle inequality, we get that the ℓ_1 distance between the two fingerprint distributions is smaller than $2 \cdot \frac{176}{5}\delta + \frac{16}{30} - \frac{352\delta}{5} = \frac{16}{30}$, which implies that the statistical difference is smaller than $\frac{8}{30}$, and thus it is not possible to distinguish between \mathcal{D}^+ and \mathcal{D}^- in the κ -Poissonized model with success probability at least $\frac{19}{30}$. By Lemma 4 we get the desired result. ■

2.1.3 The proof of Theorem 1

In this subsection we show how to apply Theorem 2 to two lists of distributions, that we will shortly define, \mathcal{D}^+ and \mathcal{D}^- , where $\mathcal{D}^+ \in \mathcal{P}^{\text{eq}} = \mathcal{P}_{m,n}^{\text{eq}}$ while \mathcal{D}^- is $(1/20)$ -far from \mathcal{P}^{eq} . Recall that by the premise of

Theorem 1, $n \geq cm \log m$ for some sufficiently large constant $c > 1$. In the proof it will be convenient to assume that m is even and that n is divisible by 4. It is not hard to verify that it is possible to reduce the general case to this case. In order to define \mathcal{D}^- , we shall need the next lemma.

Lemma 6 *For every two even integers t and $m \geq 4$, there exists a 0/1-valued matrix M with t rows and m columns for which the following holds:*

1. *In each row and each column of M , exactly half of the elements are 1 and the other half are 0.*
2. *For every integer $2 \leq x < m/2$, and for every subset $S \subseteq [m]$ of size x , the number of rows i such that $M[i, j] = 1$ for every $j \in S$ is at most $t \cdot \left(\frac{1}{2^x} + \sqrt{\frac{2x \ln m}{t}} \right)$, and at least $t \cdot \left(\frac{1}{2^x} \left(1 - \frac{2x^2}{m} \right) - \sqrt{\frac{2x \ln m}{t}} \right)$.*

Proof: Consider selecting a matrix M randomly as follows: Denote the first $t/2$ rows of M by F . For each row in F , pick, independently from the other $t/2 - 1$ rows in F , a random half of its elements to be 1, and the other half of the elements to be 0. Rows $t/2 + 1, \dots, t$ are the negations of rows $1, \dots, t/2$, respectively. Thus, in each row and each column of M , exactly half of the elements are 1 and the other half are 0.

Consider a fixed choice of x . For each row i between 1 and t , each subset of columns $S \subseteq [m]$ of size x , and $b \in \{0, 1\}$, define the indicator random variable $I_{S,i,b}$ to be 1 if and only if $M[i, j] = b$ for every $j \in S$. Hence,

$$\Pr[I_{S,i,b} = 1] = \frac{1}{2} \cdot \left(\frac{1}{2} - \frac{1}{m} \right) \cdot \dots \cdot \left(\frac{1}{2} - \frac{x-1}{m} \right). \quad (2.40)$$

Clearly, $\Pr[I_{S,i,b} = 1] < \frac{1}{2^x}$. On the other hand,

$$\Pr[I_{S,i,b} = 1] \geq \left(\frac{1}{2} - \frac{x}{m} \right)^x = \frac{1}{2^x} \left(1 - \frac{2x}{m} \right)^x \geq \frac{1}{2^x} \left(1 - \frac{2x^2}{m} \right). \quad (2.41)$$

where the last inequality is due to Bernoulli's inequality which states that $(1+x)^n > 1+nx$, for every real number $x > -1 \neq 0$ and an integer $n > 1$ ([MV70]).

Let $E_{S,b}$ denote the expected value of $\sum_{i=1}^{t/2} I_{S,i,b}$. From the fact that rows $t/2+1, \dots, t$ are the negations of rows $1, \dots, t/2$ it follows that $\sum_{i=t/2+1}^t I_{S,i,1} = \sum_{i=1}^{t/2} I_{S,i,0}$. Therefore, the expected number of rows $1 \leq i \leq t$ such that $M[i, j] = 1$ for every $j \in S$ is simply $E_{S,1} + E_{S,0}$ (that is, at most $t \cdot \frac{1}{2^x}$ and at least $t \cdot \frac{1}{2^x} \left(1 - \frac{2x^2}{m} \right)$). By the additive Chernoff bound,

$$\Pr \left[\left| \sum_{i=1}^{t/2} I_{S,i,b} - E_{S,b} \right| > \sqrt{\frac{tx \ln m}{2}} \right] < 2 \exp(-2(t/2)(2x \ln m)/t) = 2m^{-2x}. \quad (2.42)$$

Thus, by taking a union bound (over $b \in \{0, 1\}$),

$$\Pr \left[\left| \sum_{i=1}^t I_{S,i,1} - (E_{S,1} + E_{S,0}) \right| > \sqrt{2tx \ln m} \right] < 4m^{-2x}. \quad (2.43)$$

By taking a union bound over all subsets S we get that M has the desired properties with probability greater than 0. ■

We first define \mathcal{D}^+ , in which all distributions are identical. Specifically, for each $j \in [m]$:

$$D_j^+(i) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2} \\ \frac{1}{n} & \text{if } \frac{n}{2} < i \leq n \\ 0 & \text{o.w.} \end{cases} \quad (2.44)$$

We now turn to defining \mathcal{D}^- . Let M be a matrix as in Lemma 6 for $t = n/2$. For every $j \in [m]$:

$$D_j^-(i) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2} \\ \frac{2}{n} & \text{if } \frac{n}{2} < i \leq n \text{ and } M[i - n/2, j] = 1 \\ 0 & \text{o.w.} \end{cases} \quad (2.45)$$

For both \mathcal{D}^+ and \mathcal{D}^- , we refer to the elements $1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2}$ as the *heavy* elements, and to the elements $\frac{n}{2} \leq i \leq n$, as the *light* elements. Observe that each heavy element has exactly the same probability weight, $\frac{1}{n^{2/3}m^{1/3}}$, in all distributions D_j^+ and D_j^- . On the other hand, for each light element i , while $D_j^+(i) = \frac{1}{n}$ (for every j), in \mathcal{D}^- we have that $D_j^+(i) = \frac{2}{n}$ for half of the distributions, the distributions selected by the M , and $D_j^+(i) = 0$ for half of the distributions, the distributions which are not selected by M . We later use the properties of M to bound the ℓ_1 distance between the fingerprints' distributions of \mathcal{D}^+ and \mathcal{D}^- .

In order to prove that \mathcal{D}^- is $\Omega(1)$ -far from \mathcal{P}^{eq} , we first introduce some more notation (which will be used throughout the remainder of the proof of Theorem 1). For any integer x , we define the following two subsets of \mathbb{N}^m :

$$S_x \stackrel{\text{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m : (\forall j \in [m], a_j < 2) \wedge \left(\sum_{j=1}^m a_j = x \right) \right\}, \quad (2.46)$$

and

$$A_x \stackrel{\text{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m : (\exists j \in [m], a_j \geq 2) \wedge \left(\sum_{j=1}^m a_j = x \right) \right\} \quad (2.47)$$

Thus, S_x consists of vectors that contain exactly x coordinates that are 1, and all the rest are 0 (which corresponds to an element that was sampled once or 0 times by each distribution), while each vector in A_x must contain at least one coordinate that is 2 (which corresponds to an element that was sampled at least twice by at least one distribution). For $\vec{a} \in \mathbb{N}^m$, let $\text{sup}(\vec{a}) \stackrel{\text{def}}{=} \{j : a_j \neq 0\}$ denote the *support* of \vec{a} , and let

$$I_M(\vec{a}) \stackrel{\text{def}}{=} \left\{ i : D_j^-(i) = \frac{2}{n} \quad \forall j \in \text{sup}(\vec{a}) \right\}. \quad (2.48)$$

Note that in terms of the matrix M (based on which \mathcal{D}^- is defined), $I_M(\vec{a})$ consists of the rows in M whose restriction to the support of \vec{a} contains only 1's. In terms of the \mathcal{D}^- , it corresponds to the set of light elements that might have a sample histogram of \vec{a} , when sampling according to \mathcal{D}^- .

Lemma 7 For every $m > 5$ and for $n \geq c \ln m$ for some sufficiently large c , we have that $\sum_{j=1}^m \|D_j^- - D^*\|_1 > m/20$ for every distribution D^* over $[n]$. That is, the list \mathcal{D}^- is $(1/20)$ -far from \mathcal{P}^{eq} .

Proof: Consider any $\vec{a} \in S_2$. By Lemma 6, setting $t = n/2$, the size of $I_M(\vec{a})$, i.e. the number of light elements ℓ such that $D_j^-[\ell] = \frac{2}{n}$ for every $j \in \text{sup}(\vec{a})$, is at most $\frac{n}{2} \left(\frac{1}{4} + \sqrt{\frac{8 \ln m}{n}} \right)$. The same lower bound holds for the number of light elements ℓ such that $D_j^-[\ell] = 0$ for every $j \in \text{sup}(\vec{a})$. This implies that for every $j \neq j'$ in $[m]$, for at least $\frac{n}{2} - n \left(\frac{1}{4} + \sqrt{\frac{8 \ln m}{n}} \right)$ of the light elements, ℓ , we have that $D_j^-[\ell] = \frac{2}{n}$ while $D_{j'}^-[\ell] = 0$, or that $D_{j'}^-[\ell] = \frac{2}{n}$ while $D_j^-[\ell] = 0$. Therefore, $\|D_j^- - D_{j'}^-\|_1 \geq \frac{1}{2} - 2\sqrt{\frac{8 \ln m}{n}}$, which for $n \geq c \ln m$ and a sufficiently large constant c , is at least $\frac{1}{8}$. Thus, by the triangle inequality we have that for every D^* , $\sum_{j=1}^m \|D_j^- - D^*\|_1 \geq \lfloor \frac{m}{2} \rfloor \cdot \frac{1}{8}$, which greater than $m/20$ for $m > 5$. ■

In what follows we work towards establishing that Equation (2.27) in Theorem 2 holds for \mathcal{D}^+ and \mathcal{D}^- . Set $\kappa = \delta \cdot \frac{n^{2/3}}{m^{2/3}}$, where δ is a constant to be determined later. We shall use the shorthand $\vec{\lambda}^+$ for $\vec{\lambda}^{\mathcal{D}^+, \kappa}$, and $\vec{\lambda}^-$ for $\vec{\lambda}^{\mathcal{D}^-, \kappa}$ (recall that the notation $\vec{\lambda}^{\mathcal{D}, \kappa}$ was introduced in Equation (2.26)). By the definition of $\vec{\lambda}^+$, for each $\vec{a} \in \mathbb{N}^m$,

$$\vec{\lambda}^+(\vec{a}) = \sum_{i=1}^n \prod_{j=1}^m \frac{e^{-\kappa \cdot D_j^+(i)} \cdot (\kappa \cdot D_j^+(i))^{a_j}}{a_j!} \quad (2.49)$$

$$= \sum_{i=1}^{n^{2/3}m^{1/3}/2} \prod_{j=1}^m \left(\frac{(\delta/m)^{a_j}}{a_j!} \cdot e^{-\frac{\delta}{m}} \right) + \sum_{i=n/2+1}^n \prod_{j=1}^m \left(\frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} \cdot e^{-\frac{\delta}{n^{1/3}m^{2/3}}} \right) \quad (2.50)$$

$$= \frac{n^{2/3}m^{1/3}}{2e^\delta} \prod_{j=1}^m \frac{(\delta/m)^{a_j}}{a_j!} + \frac{n}{2e^{\delta(m/n)^{1/3}}} \prod_{j=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!}. \quad (2.51)$$

By the construction of M , for every light i , $\sum_{j=1}^m D_j^-(i) = \frac{2}{n} \cdot \frac{m}{2} = \frac{m}{n}$. Therefore,

$$\vec{\lambda}^-(\vec{a}) = \frac{n^{2/3}m^{1/3}}{2e^\delta} \prod_{j=1}^m \frac{(\delta/m)^{a_j}}{a_j!} + \frac{1}{e^{\delta(m/n)^{1/3}}} \sum_{i \in I_M(\vec{a})} \prod_{j=1}^m \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!}. \quad (2.52)$$

Hence, $\vec{\lambda}^+(\vec{a})$ and $\vec{\lambda}^-(\vec{a})$ differ only on the term which corresponds to the contribution of the light elements. We start by considering the contribution to Equation (2.27) of histogram vectors $\vec{a} \in S_1$ (i.e., vectors of the form $(0, \dots, 0, 1, 0, \dots, 0)$ which correspond to the number of elements that are sampled only by one distribution, once).

Lemma 8

$$\sum_{\vec{a} \in S_1} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 = 0. \quad (2.53)$$

Proof: For every $\vec{a} \in S_1$, the size of $I_M(\vec{a})$ is $\frac{n}{4}$, thus,

$$\sum_{i \in I_M(\vec{a})} \prod_{j=1}^m \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} = \frac{n}{2} \prod_{j=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!}. \quad (2.54)$$

By Equations (2.51) and (2.52), it follows that $\left| \vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}) \right| = 0$ for every $\vec{a} \in S_1$. The lemma follows by applying Equation (2.1). ■

We now turn to bounding the contribution to Equation (2.27) of histogram vectors $\vec{a} \in A_2$ (i.e., vectors of the form $(0, \dots, 0, 2, 0, \dots, 0)$ which correspond to the number of elements that are sampled only by one distribution, twice).

Lemma 9

$$\left\| \text{poi}(\vec{\lambda}^+(A_2)) - \text{poi}(\vec{\lambda}^-(A_2)) \right\|_1 \leq 3\delta. \quad (2.55)$$

Proof: For every $\vec{a} \in A_2$, the size of $I_M(\vec{a})$ is $\frac{n}{4}$, thus,

$$\sum_{i \in I_M(\vec{a})} \prod_{j=1}^m \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} = n \prod_{j=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!}. \quad (2.56)$$

By Equations (2.51), (2.52) and (2.57) it follows that

$$\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}) = \frac{n}{2e^{\delta(m/n)^{1/3}}} \prod_{j=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} = \frac{n^{1/3}\delta^2}{4e^{\delta(m/n)^{1/3}}m^{4/3}}, \quad (2.57)$$

and that

$$\vec{\lambda}^-(\vec{a}) \geq \frac{n^{2/3}m^{1/3}}{2e^\delta} \prod_{j=1}^m \frac{(\delta/m)^{a_j}}{a_j!} = \frac{n^{2/3}\delta^2}{4e^\delta m^{5/3}}. \quad (2.58)$$

By Equations (2.57) and (2.58) we have that

$$\frac{(\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}))^2}{\vec{\lambda}^-(\vec{a})} \leq \frac{e^{\delta-2\delta(m/n)^{1/3}}\delta^2}{4m} \leq \frac{\delta^2}{m}. \quad (2.59)$$

By Equation (2.59) and the fact that $|A_2| = m$ we get

$$\sum_{\vec{a} \in A_2} \frac{(\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}))^2}{\vec{\lambda}^-(\vec{a})} \leq m \cdot \frac{\delta^2}{m} = \delta^2 \quad (2.60)$$

The lemma follows by applying Lemma 2. ■

Recall that for a subset I of \mathbb{N}^m , $\text{poi}(\vec{\lambda}(I))$ denotes the multivariate Poisson distributions restricted to the coordinates of $\vec{\lambda}$ that are indexed by the vectors in I . We separately deal with S_x where $2 \leq x < m/2$, and $x \geq m/2$, where our main efforts are with respect to the former, as the latter correspond to very low probability events.

Lemma 10 For $m \geq 16$, $n \geq cm \ln m$ (where c is a sufficiently large constant) and for $\delta \leq 1/16$

$$\left\| \text{poi}(\vec{\lambda}^+\left(\bigcup_{2 \leq x < m/2} S_x\right)) - \text{poi}(\vec{\lambda}^-\left(\bigcup_{2 \leq x < m/2} S_x\right)) \right\|_1 \leq 32\delta. \quad (2.61)$$

Proof: Let \vec{a} be a vector in S_x then by the definition of S_x , every coordinate of \vec{a} is 0 or 1. Therefore we make the following simplification of Equation (2.51): For each $\vec{a} \in \bigcup_{x=2}^{m/2-1} S_x$,

$$\vec{\lambda}^+(\vec{a}) = \frac{n^{2/3}m^{1/3}}{2e^\delta} \cdot \left(\frac{\delta}{m}\right)^x + \frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{\delta}{n^{1/3}m^{2/3}}\right)^x \quad (2.62)$$

By Lemma 6, for every $\vec{a} \in \bigcup_{x=2}^{m/2-1} S_x$ the size of $I_M(\vec{a})$ is at most $\frac{n}{2} \cdot \left(\frac{1}{2^x} + \sqrt{\frac{4x \ln m}{n}}\right)$. By Equation (2.52) this implies that

$$\vec{\lambda}^-(\vec{a}) \leq \frac{n^{2/3}m^{1/3}}{2e^\delta} \cdot \left(\frac{\delta}{m}\right)^x + \frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{1}{2^x} + \sqrt{\frac{4x \ln m}{n}}\right) \left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x. \quad (2.63)$$

On the other hand, the size of $I_M(\vec{a})$ is at least $\max\left\{0, \left(\frac{1}{2^x} \left(1 - \frac{2x^2}{m}\right) - \sqrt{\frac{4x \ln m}{n}}\right)\right\}$, so

$$\begin{aligned} \vec{\lambda}^-(\vec{a}) &\geq \frac{n^{2/3}m^{1/3}}{2e^\delta} \cdot \left(\frac{\delta}{m}\right)^x \\ &+ \max\left\{0, \frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{1}{2^x} \left(1 - \frac{2x^2}{m}\right) - \sqrt{\frac{4x \ln m}{n}}\right) \left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x\right\}. \end{aligned} \quad (2.64)$$

It follows that

$$(\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}))^2 \leq \left(\frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x \cdot \left(\frac{1}{2^x} \frac{2x^2}{m} + \sqrt{\frac{4x \ln m}{n}}\right)\right)^2, \quad (2.65)$$

and so

$$\frac{(\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}))^2}{\vec{\lambda}^-(\vec{a})} \leq \frac{e^\delta n^{4/3}}{2e^{2\delta(m/n)^{1/3}} m^{1/3}} \cdot \left(\frac{4\delta}{n^{2/3}m^{1/3}}\right)^x \cdot \left(\frac{2x^2}{2^x m} + \sqrt{\frac{4x \ln m}{n}}\right)^2 \quad (2.66)$$

$$\leq \frac{n^{4/3}}{m^{1/3}} \cdot \left(\frac{4\delta}{n^{2/3}m^{1/3}}\right)^x \cdot \frac{x}{m} \cdot \left(\frac{2x^2}{2^x \sqrt{m}} + \sqrt{\frac{4m \ln m}{n}}\right)^2 \quad (2.67)$$

$$\leq \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{4x^{1/x}\delta}{n^{2/3}m^{1/3}}\right)^x \quad (2.68)$$

$$\leq \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{8\delta}{n^{2/3}m^{1/3}}\right)^x, \quad (2.69)$$

where in Equation (2.68) we used the fact that $n \geq cm \ln m$ for some sufficiently large constant c , and $\frac{2x^2}{2^x \sqrt{m}} \leq \frac{1}{2}$ for every $2 \leq x < m/2$ and $m \geq 16$. Summing over all $\vec{a} \in \bigcup_{x=2}^{m/2-1} S_x$ we get:

$$\sum_{\vec{a} \in \bigcup_{x=2}^{m/2-1} S_x} \frac{(\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}))^2}{\vec{\lambda}^-(\vec{a})} \leq \sum_{x=2}^{\infty} \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{8\delta m^{2/3}}{n^{2/3}}\right)^x \quad (2.70)$$

$$= \sum_{x=0}^{\infty} 64\delta^2 \cdot \left(\frac{8\delta m^{2/3}}{n^{2/3}}\right)^x \quad (2.71)$$

$$\leq \frac{64\delta^2}{1-8\delta} \quad (2.72)$$

$$\leq 128\delta^2 \quad (2.73)$$

where in Equation (2.71) we used the fact that $n > m$, and Equation (2.73) holds for $\delta \leq 1/16$. The lemma follows by applying Lemma 2. ■

Lemma 11 For $n \geq m$, $m \geq 12$ and $\delta \leq 1/4$,

$$\sum_{x \geq m/2} \sum_{\vec{a} \in S_x} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 \leq 32\delta^3. \quad (2.74)$$

Proof: We first observe that $|S_x| \leq m^x/x$ for every $x \geq 6$. To see why this is true, observe that $|S_x|$ equals the number of possibilities of arranging x balls in m bins, i.e.,

$$|S_x| = \binom{m+x-1}{x} \quad (2.75)$$

$$\leq \frac{(m+x)^x}{x!} \quad (2.76)$$

$$\leq \frac{(2m)^x}{x!} \quad (2.77)$$

$$= \frac{2^x}{(x-1)!} \cdot \frac{m^x}{x} \quad (2.78)$$

$$\leq \frac{m^x}{x}, \quad (2.79)$$

where inequality (2.79) holds since $m \geq 12$ and thus $x \geq 6$. By Equations (2.51) and (2.52) (and the fact that $|x-y| \leq \max\{x,y\}$ for every positive real numbers x,y),

$$\sum_{x \geq m/2} \sum_{\vec{a} \in S_x} \left| \vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}) \right| \leq \sum_{x \geq m/2} \sum_{\vec{a} \in S_x} \frac{n}{2} \prod_{j=1}^m \left(\frac{2\delta}{n^{1/3}m^{2/3}} \right)^{a_j} \quad (2.80)$$

$$= \sum_{x \geq m/2} \sum_{\vec{a} \in S_x} \frac{n}{2} \left(\frac{2\delta}{n^{1/3}m^{2/3}} \right)^{\sum_{j=1}^m a_j} \quad (2.81)$$

$$\leq \sum_{x=m/2}^{\infty} \frac{m^x}{x} \cdot \frac{n}{2} \left(\frac{2\delta}{n^{1/3}m^{2/3}} \right)^x \quad (2.82)$$

$$\leq \sum_{x=m/2}^{\infty} \frac{2m^x}{m} \cdot \frac{n}{2} \left(\frac{2\delta}{n^{1/3}m^{2/3}} \right)^x \quad (2.83)$$

$$= \frac{n}{m} \sum_{x=m/2}^{\infty} \left(\frac{2\delta m^{1/3}}{n^{1/3}} \right)^x \quad (2.84)$$

$$= 8\delta^3 \sum_{x=m/2-3}^{\infty} \left(\frac{2\delta m^{1/3}}{n^{1/3}} \right)^x \quad (2.85)$$

$$\leq \frac{8\delta^3}{1-2\delta} \quad (2.86)$$

$$\leq 16\delta^3 \quad (2.87)$$

where in Equation (2.86) we used the fact that $n \geq m$ and Equation (2.87) holds for $\delta \leq 1/4$. The lemma follows by applying Equation (2.1). ■

We finally turn to the contribution of $\vec{a} \in A_x$ such that $x \geq 3$.

Lemma 12 For $n \geq m$ and $\delta \leq 1/4$,

$$\sum_{x \geq 3} \sum_{\vec{a} \in A_x} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 \leq 16\delta^3. \quad (2.88)$$

Proof: We first observe that $|A_x| \leq m^{x-1}$ for every x . To see why this is true, observe that $|A_x|$ equals the number of possibilities of arranging $x - 1$ balls, where one ball is a ‘‘special’’ (‘‘double’’) ball in m bins. By Equations (2.51) and (2.52) (and the fact that $|x - y| \leq \max\{x, y\}$ for every positive real numbers x, y),

$$\sum_{x \geq 3} \sum_{\vec{a} \in A_x} \left| \vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}) \right| \leq \sum_{x \geq 3} \sum_{\vec{a} \in A_x} \frac{n}{2} \prod_{j=1}^m \left(\frac{2\delta}{n^{1/3} m^{2/3}} \right)^{a_j} \quad (2.89)$$

$$= \sum_{x \geq 3} \sum_{\vec{a} \in A_x} \frac{n}{2} \left(\frac{2\delta}{n^{1/3} m^{2/3}} \right)^{\sum_{j=1}^m a_j} \quad (2.90)$$

$$\leq \sum_{x=3}^{\infty} m^{x-1} \cdot \frac{n}{2} \left(\frac{2\delta}{n^{1/3} m^{2/3}} \right)^x \quad (2.91)$$

$$= \frac{n}{2m} \sum_{x=3}^{\infty} \left(\frac{2\delta m^{1/3}}{n^{1/3}} \right)^x \quad (2.92)$$

$$= 4\delta^3 \sum_{x=0}^{\infty} \left(\frac{2\delta m^{1/3}}{n^{1/3}} \right)^x \quad (2.93)$$

$$\leq \frac{4\delta^3}{1 - 2\delta} \quad (2.94)$$

$$\leq 8\delta^3 \quad (2.95)$$

where in Equation (2.94) we used the fact that $n \geq m$ and Equation (2.95) holds for $\delta \leq 1/4$. The lemma follows by applying Equation (2.1). ■

We are now ready to finalize the proof of Theorem 1.

Proof of Theorem 1: Let \mathcal{D}^+ and \mathcal{D}^- be as defined in Equations (2.44) and (2.45), respectively, and recall that $\kappa = \delta \cdot \frac{n^{2/3}}{m^{2/3}}$ (where δ will be set subsequently). By the definition of the distributions in \mathcal{D}^+ and \mathcal{D}^- , the probability weight assigned to each element is at most $\frac{1}{n^{2/3} m^{1/3}} = \frac{\delta}{\kappa \cdot m}$, as required by Theorem 2. By Lemma 7, \mathcal{D}^- is $(1/20)$ -far from \mathcal{P}^{eq} . Therefore, it remains to establish that Equation (2.27) holds for \mathcal{D}^+ and \mathcal{D}^- . Consider the following partition of \mathbb{N}^m :

$$\left\{ \{\vec{a}\}_{\vec{a} \in S_1, A_2}, \bigcup_{2 \leq x < m/2} S_x, \{\vec{a}\}_{\vec{a} \in \bigcup_{x \geq m/2} S_x}, \{\vec{a}\}_{\vec{a} \in \bigcup_{x \geq 3} A_x} \right\}, \quad (2.96)$$

where $\{\vec{a}\}_{\vec{a} \in T}$ denotes the list of all singletons of elements in T . By Lemma 1 it follows that

$$\begin{aligned}
\left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 &\leq \sum_{\vec{a} \in S_1} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 \\
&\quad + \left\| \text{poi}(\vec{\lambda}^+(A_2)) - \text{poi}(\vec{\lambda}^-(A_2)) \right\|_1 \\
&\quad + \left\| \text{poi}(\vec{\lambda}^+(\bigcup_{2 \leq x < m/2} S_x)) - \text{poi}(\vec{\lambda}^-(\bigcup_{2 \leq x < m/2} S_x)) \right\|_1 \\
&\quad + \sum_{x \geq m/2} \sum_{\vec{a} \in S_x} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 \\
&\quad + \sum_{x \geq 3} \sum_{\vec{a} \in A_x} \left\| \text{poi}(\vec{\lambda}^+(\vec{a})) - \text{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 . \tag{2.97}
\end{aligned}$$

For $\delta < 1/16$ we get by Lemmas 8–12 that

$$\left\| \text{poi}(\vec{\lambda}^+) - \text{poi}(\vec{\lambda}^-) \right\|_1 \leq 35\delta + 48\delta^3 , \tag{2.98}$$

which is less than $\frac{16}{30} - \frac{352\delta}{5}$ for $\delta = 1/200$. ■

2.1.4 A lower bound for testing Independence

Corollary 4 *Given a joint distribution Q over $[m] \times [n]$ impossible to test if Q is independent or $1/48$ -far from independent using $o(n^{2/3}m^{1/3})$ samples.*

Proof: Follows directly from Lemma 13 and Theorem 1. ■

2.2 Algorithms for Testing Equivalence in the Sampling Model

In this section we establish the following Theorems.

Theorem 5 *Let \mathcal{D} be a list of m distributions over $[n]$. It is possible to test whether $\mathcal{D} \in \mathcal{P}^{\text{eq}}$ in the unknown-weights sampling model using a sample of size $\tilde{O}((n^{2/3}m^{1/3} + m) \cdot \text{poly}(1/\epsilon))$.*

Theorem 6 *Let \mathcal{D} be a list of m distributions over $[n]$. It is possible to test whether $\mathcal{D} \in \mathcal{P}_{m,n}^{\text{eq}}$ in the known-weights sampling model using a sample of size $\tilde{O}((n^{1/2}m^{1/2} + n) \cdot \text{poly}(1/\epsilon))$.*

We shall actually prove a stronger version of Theorem 6 (see Theorem 12) which allows the distributions in \mathcal{D} to be close (as a function of n or alternatively coordinate-wise) and not necessarily identical.

Thus, when the weight vector \vec{w} is known, and in particular when all weights are equal (the uniform sampling model) we get a combined upper bound of $\tilde{O}(\min\{n^{2/3}m^{1/3} + m, n^{1/2}m^{1/2} + n\} \cdot \text{poly}(1/\epsilon))$. Namely, as long as $n \geq m$ the complexity (in terms of the dependence on n and m) grows like $\tilde{O}(n^{2/3}m^{1/3})$, and when $m \geq n$ it grows like $\tilde{O}(n^{1/2}m^{1/2})$.

In order to prove Theorem 5 we shall consider a (related) property of joint distributions over $[n] \times [m]$. Specifically, we are interested in determining whether a distribution Q over $[n] \times [m]$ is a *product* distribution $Q_1 \times Q_2$ where Q_1 is a distribution over $[n]$ and Q_2 is a distribution over $[m]$ (i.e., $Q(i, j) = Q_1(i) \cdot Q_2(j)$ for every $(i, j) \in [n] \times [m]$). In other words, if we denote by $\pi_1 Q$ the marginal distribution according to Q of the first coordinate, i , and by $\pi_2 Q$ the marginal distribution of the second coordinate, j , then we ask whether $\pi_1 Q$ and $\pi_2 Q$ are independent. With a slight abuse of the terminology, we shall say in such a case that Q is *independent*.

As we observe in the the next lemma, the problem of testing independence of a joint distribution and the problem of testing equivalence of a list of distributions in the (not necessarily uniform) sampling model, are closely related.

Lemma 13 *If there exists an algorithm T for testing whether a joint distribution Q over $[m] \times [n]$ is independent using a sample of size $s(m, n, \epsilon)$, then there exists an algorithm T' for testing whether $\mathcal{D} \in \mathcal{P}_{m,n}^{\text{eq}}$ in the unknown-weights sampling model using a sample of size $s(m, n, \epsilon/3)$.*

If T is provided with (and uses) an explicit description of the marginal distribution $\pi_2 Q$, then the claim holds for T' in the known-weights sampling model.

Proof: Given a sample $\{(i_\ell, j_\ell)\}_{\ell=1}^s(m, n, \epsilon/3)$ generated according to \mathcal{D} in the sampling model with a weight vector $\vec{w} = (w_1, \dots, w_m)$, the algorithm T' simply runs T on the sample and returns the answer that T gives. If \vec{w} is known, then T' provides T with \vec{w} (as the marginal distribution of j). If D_1, \dots, D_m are identical and equal to some D^* , then for each $(i, j) \in [n] \times [m]$ we have that the probability of getting (i, j) in the sample is $w_j \cdot D^*(i)$. That is, the joint distribution of the first and second coordinates is independent and therefore T (and hence T') accepts with probability at least $2/3$.

On the other hand, suppose that \mathcal{D} is ϵ -far from $\mathcal{P}_{m,n}^{\text{eq}}$, that is, $\sum_{j=1}^m w_j \cdot \|D_j - D^*\|_1 > \epsilon$ for every distribution, D^* over $[n]$. In such a case, in particular we have that $\sum_{j=1}^m w_j \cdot \|D_j - \bar{D}\|_1 > \epsilon$, where \bar{D} is the distribution over $[n]$ such that $\bar{D}(i) = \sum_{j=1}^m w_j \cdot D_j(i)$. By Proposition 1 in [BFF⁺01], the joint distribution Q over i and j (determined by the list \mathcal{D} and the sampling process) is $\delta/3$ -far from independent, so T (and hence T') rejects with probability greater than $2/3$. ■

2.2.1 Proof of Theorem 5

By Lemma 13, in order to prove Theorem 5 it suffices to design an algorithm for testing independence of a joint distribution (with the complexity stated in the theorem). Indeed, testing independence was studied in [BFF⁺01]. However, there was a certain flaw in one of the claims on which their analysis built (Theorem 15 in [BFF⁺01], which is attributed to [BFR⁺10]), and hence we fix the flaw next (building on [BFR⁺10]).

Given a sampling access to a pair of distributions \mathbf{p} and \mathbf{q} and bounds on their ℓ_∞ -norm $b_{\mathbf{p}}$ and $b_{\mathbf{q}}$, respectively, the algorithm **Bounded- L_∞ -Closeness-Test** (Algorithm 1 in Figure 2.1) tests the closeness of \mathbf{p} and \mathbf{q} . The sample complexity of the algorithm depends on $b_{\mathbf{p}}$ and $b_{\mathbf{q}}$, as described in the next theorem.

We note that the check in step 8 in the algorithm was added to the original ℓ_2 -Distance-Test of [BFR⁺10] to achieve a tighter bound on the sample complexity.

For a multiset of sample points F over a domain R and an element $i \in R$, let $\text{occ}(i, F)$ denote the number of times that i appears in the sample F and define the *collision count* of F to be $\text{coll}(F) \stackrel{\text{def}}{=} \sum_{i \in R} \binom{\text{occ}(i, F)}{2}$.

Theorem 7 *Let \mathbf{p} and \mathbf{q} be two distributions over the same finite domain R . Suppose that $\|\mathbf{p}\|_\infty \leq b_{\mathbf{p}}$ and $\|\mathbf{q}\|_\infty \leq b_{\mathbf{q}}$ where $b_{\mathbf{q}} \geq b_{\mathbf{p}}$. Algorithm Bounded- L_∞ -Closeness-Test($\mathbf{p}, \mathbf{q}, b_{\mathbf{p}}, b_{\mathbf{q}}, \epsilon$) is such that:*

1. *If $\|\mathbf{p} - \mathbf{q}\|_1 \leq \epsilon/(2|R|^{1/2})$, then the test accepts with probability at least $2/3$.*
2. *If $\|\mathbf{p} - \mathbf{q}\|_1 > \epsilon$, then the test rejects with probability at least $2/3$.*

The algorithm takes $O\left(|R| \cdot b_{\mathbf{p}}^{1/2}/\epsilon^2 + |R|^2 \cdot b_{\mathbf{q}} \cdot b_{\mathbf{p}}/\epsilon^4\right)$ sample points from each distribution.

Proof: Following the analysis of [BFR⁺10, Lemma 5], we have that:

Algorithm 1: Bounded-L_∞-Closeness-Test	
Input: $\mathbf{p}, \mathbf{q}, b_{\mathbf{p}}, b_{\mathbf{q}}, \epsilon$	
1	Take samples $F_{\mathbf{p}}^1$ and $F_{\mathbf{p}}^2$ from \mathbf{p} , each of size t , where $t = O\left(R \cdot b_{\mathbf{p}}^{1/2}/\epsilon^2 + R ^2 \cdot b_{\mathbf{q}} \cdot b_{\mathbf{p}}/\epsilon^4\right)$;
2	Take samples $F_{\mathbf{q}}^2$ and $F_{\mathbf{q}}^1$ from \mathbf{q} , each of size t ;
	<i>/* $r_{\mathbf{p}}$ is the the number of self collisions in $F_{\mathbf{p}}^1$. */</i>
3	Let $r_{\mathbf{p}} = \text{coll}(F_{\mathbf{p}}^1)$;
	<i>/* $r_{\mathbf{q}}$ is the the number of self collisions in $F_{\mathbf{q}}^1$. */</i>
4	Let $r_{\mathbf{q}} = \text{coll}(F_{\mathbf{q}}^1)$;
	<i>/* $s_{\mathbf{p}, \mathbf{q}}$ is the number of collisions between $F_{\mathbf{p}}^2$ and $F_{\mathbf{q}}^2$. */</i>
5	Let $s_{\mathbf{p}, \mathbf{q}} = \sum_{i \in R} (\text{occ}(i, F_{\mathbf{p}}^2) \cdot \text{occ}(i, F_{\mathbf{q}}^2))$;
6	Define $r \stackrel{\text{def}}{=} \frac{2t}{t-1} (r_{\mathbf{p}} + r_{\mathbf{q}})$;
7	Define $s \stackrel{\text{def}}{=} 2s_{\mathbf{p}, \mathbf{q}}$;
8	if $r_{\mathbf{q}} > (7/4) \binom{t}{2} b_{\mathbf{p}}$ then output REJECT ;
9	Define $\delta \stackrel{\text{def}}{=} \epsilon/ R ^{1/2}$;
10	if $r - s > t^2 \delta^2 / 2$ then output REJECT ;
11	output ACCEPT ;

Figure 2.1: The algorithm for testing ℓ_1 distance when L_∞ is bounded

$$\text{Exp}[r - s] = t^2 \|\mathbf{p} - \mathbf{q}\|_2^2, \quad (2.99)$$

and we have the following bounds on the variances of $r_{\mathbf{p}}$, $r_{\mathbf{q}}$ and s (for some constant c):

$$\text{Var}[s] \leq ct^2 \sum_{\ell \in R} \mathbf{p}(\ell) \mathbf{q}(\ell) + ct^3 \sum_{\ell \in R} (\mathbf{p}(\ell) \mathbf{q}(\ell)^2 + \mathbf{p}(\ell)^2 \mathbf{q}(\ell)), \quad (2.100)$$

$$\text{Var}[r_{\mathbf{p}}] \leq ct^2 \sum_{\ell \in R} \mathbf{p}(\ell)^2 + ct^3 \sum_{\ell \in R} \mathbf{p}(\ell)^3, \quad (2.101)$$

and

$$\text{Var}[r_{\mathbf{q}}] \leq ct^2 \sum_{\ell \in R} \mathbf{q}(\ell)^2 + ct^3 \sum_{\ell \in R} \mathbf{q}(\ell)^3. \quad (2.102)$$

Using the bounds we have on the ℓ_∞ norms of \mathbf{p} and \mathbf{q} we get (possibly for a larger constant c):

$$\text{Var}[s] \leq ct^2 \|\mathbf{p}\|_\infty + ct^3 (\|\mathbf{p}\|_\infty \|\mathbf{q}\|_2^2 + \|\mathbf{p}\|_\infty^2) \leq ct^2 b_{\mathbf{p}} + ct^3 (b_{\mathbf{p}} \|\mathbf{q}\|_2^2 + b_{\mathbf{p}}^2), \quad (2.103)$$

$$\text{Var}[r_{\mathbf{p}}] \leq ct^2 \|\mathbf{p}\|_2^2 + ct^3 \|\mathbf{p}\|_\infty \|\mathbf{p}\|_2^2 \leq ct^2 \|\mathbf{p}\|_\infty + ct^3 \|\mathbf{p}\|_\infty^2 \leq ct^2 b_{\mathbf{p}} + ct^3 b_{\mathbf{p}}^2, \quad (2.104)$$

and

$$\text{Var}[r_{\mathbf{q}}] \leq ct^2 \|\mathbf{q}\|_2^2 + ct^3 \|\mathbf{q}\|_\infty \|\mathbf{q}\|_2^2 \leq ct^2 \|\mathbf{q}\|_2^2 + ct^3 b_{\mathbf{q}} \|\mathbf{q}\|_2^2. \quad (2.105)$$

First, we prove that the tester distinguishes with high constant probability between the case that $\|\mathbf{q}\|_2^2 > 2b_{\mathbf{p}}$ and the case that $\|\mathbf{q}\|_2^2 \leq (3/2)b_{\mathbf{p}}$ by rejecting (w.h.p.) when $r_{\mathbf{q}} > (7/4)\binom{t}{2}b_{\mathbf{p}}$. Notice that by the triangle inequality $\|\mathbf{p} - \mathbf{q}\|_2 \geq \|\mathbf{q}\|_2 - \|\mathbf{p}\|_2$. Thus, if $\|\mathbf{q}\|_2^2 > (3/2)b_{\mathbf{p}}$ and $\|\mathbf{p}\|_2^2 \leq b_{\mathbf{p}}$ then it follows that $\|\mathbf{p} - \mathbf{q}\|_2 \geq \sqrt{(3/2)b_{\mathbf{p}}} - b_{\mathbf{p}}^{1/2}$. Therefore, by the fact that $b_{\mathbf{p}} \geq 1/|R|$, we obtain that $\|\mathbf{p} - \mathbf{q}\|_1 \geq \|\mathbf{p} - \mathbf{q}\|_2 \geq (\sqrt{(3/2)} - 1)/|R|^{1/2}$ which is greater than $\epsilon/(2|R|^{1/2})$ for $\epsilon \leq 1$. Consider first the case that $\|\mathbf{q}\|_2^2 > 2b_{\mathbf{p}}$, so that $\text{Exp}[r_{\mathbf{q}}] > 2\binom{t}{2}b_{\mathbf{p}}$. Then we can bound the probability that the tester accepts, that is, that $r_{\mathbf{q}} \leq (7/4)\binom{t}{2}b_{\mathbf{p}}$, by the probability that $r_{\mathbf{q}} < (7/8)\text{Exp}[r_{\mathbf{q}}]$. In the case that $\|\mathbf{q}\|_2^2 \leq (3/2)b_{\mathbf{p}}$, so that $\text{Exp}[r_{\mathbf{q}}] \leq (3/2)\binom{t}{2}b_{\mathbf{p}}$, we can bound the probability that the tester rejects, that is, that $r_{\mathbf{q}} > (7/4)\binom{t}{2}b_{\mathbf{p}}$, by the probability that $r_{\mathbf{q}} > (7/6)\text{Exp}[r_{\mathbf{q}}]$. Then the probability to accept when $\|\mathbf{q}\|_2^2 > 2b_{\mathbf{p}}$ and reject when $\|\mathbf{q}\|_2^2 \leq b_{\mathbf{p}}$ is upper bounded by $\Pr[|r_{\mathbf{q}} - \text{Exp}[r_{\mathbf{q}}]| > \text{Exp}[r_{\mathbf{q}}]/8]$. Now, using the upper bound on the variance of $r_{\mathbf{q}}$ that we have (the first bound in Equation (2.105)), the fact that for every distribution \mathbf{q} over R , $\|\mathbf{q}\|_2^2 \leq 1/|R|$ and $\text{Exp}[r_{\mathbf{q}}] = \binom{t}{2} \|\mathbf{q}\|_2^2$, we have that

$$\Pr[|r_{\mathbf{q}} - \text{Exp}[r_{\mathbf{q}}]| > \text{Exp}[r_{\mathbf{q}}]/8] \leq \frac{64\text{Var}[r_{\mathbf{q}}]}{\text{Exp}^2[r_{\mathbf{q}}]} \quad (2.106)$$

$$\leq \frac{c \cdot (t^2 \|\mathbf{q}\|_2^2 + t^3 \|\mathbf{q}\|_\infty \|\mathbf{q}\|_2^2)}{t^4 \|\mathbf{q}\|_2^4} \quad (2.107)$$

$$= \frac{c}{t^2 \|\mathbf{q}\|_2^2} + \frac{c \|\mathbf{q}\|_\infty}{t \|\mathbf{q}\|_2^2} \quad (2.108)$$

$$\leq \frac{c|R|}{t^2} + \frac{c|R| \|\mathbf{q}\|_\infty}{t}, \quad (2.109)$$

If we want this to be a small constant, then we need to take t so that:

$$t = \Omega\left(|R|^{1/2} + |R|b_{\mathbf{q}}\right). \quad (2.110)$$

Next, we prove that the tester distinguishes between the case that $\|\mathbf{p} - \mathbf{q}\|_2 > \delta$ and $\|\mathbf{p} - \mathbf{q}\|_2 \leq \delta/2$ by rejecting when $r - s > t^2\delta^2/2$. We have that $\text{Exp}[r - s] = t^2\|\mathbf{p} - \mathbf{q}\|_2^2$. Chebyshev gives us that

$\Pr[|A - \text{Exp}[A]| > \rho] \leq \text{Var}[A]/\rho^2$, and so, for the case $\|\mathbf{p} - \mathbf{q}\|_2 > \delta$ (i.e. $\text{Exp}[r - s] > t^2\delta^2$) we have that

$$\Pr[r - s < t^2\delta^2/2] \leq \Pr[|(r - s) - \text{Exp}[r - s]| < t^2\delta^2/2] \quad (2.111)$$

$$\leq \frac{4\text{Var}[r - s]}{t^4\delta^4}, \quad (2.112)$$

and similarly, for the case $\|\mathbf{p} - \mathbf{q}\|_2 \leq \delta/2$ (i.e. $\text{Exp}[r - s] \leq t^2\delta^2/4$) we have that

$$\Pr[r - s \geq t^2\delta^2/2] \leq \Pr[|(r - s) - \text{Exp}[r - s]| < t^2\delta^2/4] \quad (2.113)$$

$$\leq \frac{16\text{Var}[r - s]}{t^4\delta^4}. \quad (2.114)$$

That is, we want $\frac{\text{Var}[r-s]}{t^4\delta^4}$ which is of the order of $\frac{\text{Var}[r-s] \cdot |R|^2}{t^4\epsilon^4}$ to be a small constant. If we use $\text{Var}[r - s] = \frac{4t^2}{(t-1)^2} (\text{Var}[r_{\mathbf{p}}] + \text{Var}[r_{\mathbf{q}}]) + \text{Var}[s]$, then we need to ensure that each of $\frac{\text{Var}[r_{\mathbf{p}}] \cdot |R|^2}{t^4\epsilon^4}$, $\frac{\text{Var}[r_{\mathbf{q}}] \cdot |R|^2}{t^4\epsilon^4}$ and $\frac{\text{Var}[s] \cdot |R|^2}{t^4\epsilon^4}$ is a small constant, which by Equations (2.103), (2.104), (2.105), and the premise that $\|\mathbf{q}\|_2^2 \leq 2b_{\mathbf{p}}$, holds when

$$t = \Omega\left(|R| \cdot b_{\mathbf{p}}^{1/2}/\epsilon^2 + |R|^2 \cdot b_{\mathbf{q}} \cdot b_{\mathbf{p}}/\epsilon^4\right), \quad (2.115)$$

since both $b_{\mathbf{p}}, b_{\mathbf{q}} \geq 1/|R|$, this dominates the sample complexity. ■

As a corollary of Theorem 7 we obtain:

Theorem 8 *Let Q be a distribution over $[n] \times [m]$ such that Q satisfies: $\|\pi_1 Q\|_{\infty} \leq b_1$, $\|\pi_2 Q\|_{\infty} \leq b_2$ and $b_1 \leq b_2$. There is a test that takes $O(nmb_1^{1/2}b_2^{1/2}/\epsilon^2 + n^2m^2b_1^2b_2/\epsilon^4)$ samples from Q , such that if Q is independent, then the test accepts with probability at least $2/3$ and if Q is ϵ -far from independent, then the test rejects with probability at least $2/3$.*

Proof: By the premise of the theorem we have that $\|Q\|_{\infty} \leq b_1$ and that $\|\pi_1 Q \times \pi_2 Q\|_{\infty} \leq b_1 \cdot b_2$. Applying Theorem 7 we can test if Q is identical to $\pi_1 Q \times \pi_2 Q$ using sample of size $O(nmb_1^{1/2}b_2^{1/2}/\epsilon^2 + n^2m^2b_1^2b_2/\epsilon^4)$ from¹ Q . If Q is independent, then Q equals $\pi_1 Q \times \pi_2 Q$ and the tester accepts with probability at least $2/3$. If Q is ϵ -far from independent, then in particular Q is ϵ -far from $\pi_1 Q \times \pi_2 Q$ and the tester rejects with probability at least $2/3$. ■

Applying Theorem 8 with $b_1 = 1/n^{2/3}m^{1/3}$, $b_2 = 1/m$, and combining that in the sample analysis of TestLightIndependence [BFF⁺01], the following theorem is obtained:

Theorem 9 [BFF⁺01] *There is an algorithm that given a distribution Q over $[n] \times [m]$ and an $\epsilon > 0$,*

- *If Q is independent then the test accepts with high probability.*
- *If Q is ϵ -far from independent then the test rejects with high probability.*

The algorithm uses $\tilde{O}((n^{2/3}m^{1/3} + m)\text{poly}(\epsilon^{-1}))$ samples.

Finally, Theorem 5 follows by combining Theorem 9 with Lemma 13.

¹We obtain a sample from $\pi_1 Q \times \pi_2 Q$ by simply taking two independent samples from Q , (i_1, j_1) and (i_2, j_2) and considering (i_1, j_2) as a sample from $\pi_1 Q \times \pi_2 Q$.

2.2.2 Proof of Theorem 6

In the proof of Theorem 6 we exploit the knowledge of the weights over the distributions, i.e. \vec{w} , which allows us to improve on the result which is stated in Theorem 5, when $m \geq n$. In fact in Theorem 12 we prove a stronger statement than the one in Theorem 12.

2.3 Algorithms for Testing Tolerant Equivalence in the Sampling Model

2.3.1 Algorithm for Testing Tolerant Identity in the Sampling Model

In this section we prove Theorem 10 which is a restatement of theorems in [Whi] and [BFF⁺01]. In order to state it we introduce the following new definitions.

Definition 2 For two parameters $\alpha, \beta \in (0, 1)$, we say that a distribution \mathbf{p} is an (α, β) -multiplicative approximation of a distribution \mathbf{q} (over the same domain R) if the following holds.

- For every $i \in R$ such that $\mathbf{q}(i) \geq \alpha$ we have that $\mathbf{q}(i) \cdot (1 - \beta) \leq \mathbf{p}(i) \leq \mathbf{q}(i) \cdot (1 + \beta)$.
- For every $i \in R$ such that $\mathbf{q}(i) < \alpha$ we have that $\mathbf{p}(i) < \alpha \cdot (1 + \beta)$.

Definition 3 For $\alpha \in (0, 1)$, we say that a distribution \mathbf{p} is an α -additive approximation of a distribution \mathbf{q} (over the same domain R) if for every $i \in R$, $|\mathbf{p}(i) - \mathbf{q}(i)| \leq \alpha$.

Theorem 10 (Adapted from [Whi], [BFF⁺01]) Given sample access to \mathbf{p} , a black-box distribution over a finite domain R , and \mathbf{q} , an explicitly specified distribution over R , for every $0 < \epsilon \leq 1/3$, algorithm *Test-Tolerant-Identity*($\mathbf{p}, \mathbf{q}, n, \epsilon$) is such that:

1. If $\|\mathbf{p} - \mathbf{q}\|_1 > 13\epsilon$, it rejects with high constant probability.
2. If \mathbf{q} is a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of some \mathbf{q}' such that $\|\mathbf{p} - \mathbf{q}'\|_1 \leq \frac{72\epsilon^2}{\ell\sqrt{n}}$, where $\ell = \log(n/\epsilon) / \log(1 + \epsilon)$, it accepts with high constant probability (in particular, if \mathbf{q} is a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of \mathbf{p} or if $\|\mathbf{p} - \mathbf{q}\|_1 \leq \frac{72\epsilon^2}{\ell\sqrt{n}}$, the test accepts with high constant probability).

The algorithm takes $\tilde{O}(\sqrt{n}\text{poly}(\epsilon^{-1}))$ samples from \mathbf{p} .

In the proof of Theorem 10 we shall use the following lemmas.

Definition 4 ([BFF⁺01]) Given an explicit distribution \mathbf{p} over R , *Bucket*($\mathbf{p}, R, \alpha, \beta$) is the partition $\{R_0, \dots, R_\ell\}$ of R with $\ell = \log(1/\alpha) / \log(1 + \beta)$, $R_0 = \{i : \mathbf{p}(i) \leq \alpha\}$, such that for all j in $[\ell]$,

$$R_j = \{i : \alpha(1 + \beta)^{j-1} < \mathbf{p}(i) \leq \alpha(1 + \beta)^j\} \quad (2.116)$$

Definition 5 ([BFF⁺01]) Given a random variable \mathbf{p} over R , and a partition $\mathcal{R} = \{R_1, \dots, R_\ell\}$ of R , the coarsening $\mathbf{p}_{\langle \mathcal{R} \rangle}$ is the random variable over $[\ell]$ with distribution $\mathbf{p}_{\langle \mathcal{R} \rangle}(i) = \mathbf{p}(R_i)$.

Theorem 11 ([BFF⁺01]) Let \mathbf{p} be a black-box distribution over a finite domain R and let S be a sample set from \mathbf{p} . $\text{coll}(S)/\binom{|S|}{2}$ approximates $\|\mathbf{p}\|_2^2$ to within a factor of $(1 \pm \epsilon)$, with probability at least $1 - \delta$, provided that $|S| = \Omega(\sqrt{|R|}\epsilon^{-2} \log(1/\delta))$.

Lemma 14 ([BFF⁺01]) Let \mathbf{p}, \mathbf{q} be distributions over R and let $R' \subseteq R$, then $\|\mathbf{p}_{|R'} - \mathbf{q}_{|R'}\|_1 \leq 2\|\mathbf{p} - \mathbf{q}\|_1 / \mathbf{p}(R')$.

Lemma 15 ([BFF⁺01]) For any distribution \mathbf{p} over R , $\|\mathbf{p}\|_2^2 - \|U_R\|_2^2 = \|\mathbf{p} - U_R\|_2^2$.

Let \mathbf{p} be a distribution over some finite domain R , and let R' be a subset of R such that $\mathbf{p}(R') > 0$ where $\mathbf{p}(R') = \sum_{i \in R'} \mathbf{p}(i)$. Denote by $\mathbf{p}_{|R'}$ the restriction of \mathbf{p} to R' , i.e., $\mathbf{p}_{|R'}$ is a distribution over R' such that for every $i \in R'$, $\mathbf{p}_{|R'}(i) = \frac{\mathbf{p}(i)}{\mathbf{p}(R')}$.

Lemma 16 (Based on [BFF⁺01]) Let \mathbf{p}, \mathbf{q} be distributions over R and let $R' \subseteq R$, then $\sum_{i \in R'} |\mathbf{p}(i) - \mathbf{q}(i)| \leq |\mathbf{p}(R') - \mathbf{q}(R')| + \mathbf{q}(R') \|\mathbf{p}_{|R'} - \mathbf{q}_{|R'}\|_1$.

Proof:

$$\sum_{i \in R'} |\mathbf{p}(i) - \mathbf{q}(i)| \leq \sum_{i \in R'} \left| \frac{\mathbf{p}(i)(\mathbf{p}(R') - \mathbf{q}(R'))}{\mathbf{p}(R')} \right| + \sum_{i \in R'} \left| \frac{\mathbf{p}(i)\mathbf{q}(R')}{\mathbf{p}(R')} - \mathbf{q}(i) \right| \quad (2.117)$$

$$= |\mathbf{p}(R') - \mathbf{q}(R')| + \sum_{i \in R'} \left| \frac{\mathbf{p}(i)\mathbf{q}(R')}{\mathbf{p}(R')} - \mathbf{q}(i) \right| \quad (2.118)$$

$$= |\mathbf{p}(R') - \mathbf{q}(R')| + \sum_{i \in R'} \mathbf{q}(R') \cdot \left| \frac{\mathbf{p}(i)}{\mathbf{p}(R')} - \frac{\mathbf{q}(i)}{\mathbf{q}(R')} \right| \quad (2.119)$$

$$= |\mathbf{p}(R') - \mathbf{q}(R')| + \mathbf{q}(R') \cdot \|\mathbf{p}_{|R'} - \mathbf{q}_{|R'}\|_1 \quad (2.120)$$

■

Lemma 17 Let \mathbf{p}, \mathbf{q} be distributions over a finite domain R and let $R' \subseteq R$ be a subset of R such that for every $i \in R'$ it holds that

$$\mathbf{p}(i)(1 - \epsilon) \leq \mathbf{q}(i) \leq \mathbf{p}(i)(1 + \epsilon), \quad (2.121)$$

then for every $i \in R'$,

$$\mathbf{p}_{|R'}(i) \frac{(1 - \epsilon)}{(1 + \epsilon)} \leq \mathbf{q}_{|R'}(i) \leq \mathbf{p}_{|R'}(i) \frac{(1 + \epsilon)}{(1 - \epsilon)} \quad (2.122)$$

Proof: Equation (2.121) implies that $\mathbf{p}(R')(1 - \epsilon) \leq \mathbf{q}(R') \leq \mathbf{p}(R')(1 + \epsilon)$ and therefore $\frac{1}{1 + \epsilon} \leq \frac{\mathbf{p}(R')}{\mathbf{q}(R')} \leq \frac{1}{1 - \epsilon}$. Thus, we obtain that $\frac{\mathbf{p}(i)}{\mathbf{p}(R')} \cdot \frac{(1 - \epsilon)}{(1 + \epsilon)} \leq \frac{\mathbf{q}(i)}{\mathbf{q}(R')} \leq \frac{\mathbf{p}(i)}{\mathbf{p}(R')} \cdot \frac{(1 + \epsilon)}{(1 - \epsilon)}$, and the lemma follows. ■

Proof of Theorem 10: The algorithm **Test-Tolerant-Identity** is given in Figure 2.2. Let E_1 be the event that for every i in $[\ell]$ we have that m_i approximates $\|\mathbf{p}_{|R_i}\|_2^2$ to within a factor of $(1 \pm \epsilon^2)$. By Theorem 11,

Algorithm 2: Test-Tolerant-Identity**Input:** Sampling access to $\mathbf{p}, \mathbf{q}, n, \epsilon$

- 1 $\mathcal{R} \stackrel{\text{def}}{=} \{R_0, \dots, R_\ell\} = \text{Bucket}(\mathbf{q}, n, \epsilon/n, \epsilon/24)$;
- 2 Let S be a set of $\tilde{\Theta}(\sqrt{n}\epsilon^{-5} \log n)$ samples from \mathbf{p} ;
- 3 Let H be the set of all x such that $\mathbf{q}(x) > \epsilon(1 + \epsilon)/n$;
- 4 **foreach** $R_i \subseteq H$ **do**
- 5 Let $S_i = S \cap R_i$;
- 6 **if** $\mathbf{q}(R_i) \geq \epsilon/\ell$ **then**
- 7 **if** $|S_i| < \Theta(\sqrt{n}\epsilon^{-4} \log \ell)$ **then output REJECT**;
- 8 Let $m_i = \text{coll}(S_i) / \binom{|S_i|}{2}$;
- 9 **if** $m_i > \frac{(1+\epsilon^2)^2}{|R_i|}$ **then output REJECT**;
- 10 Take $\Theta(\epsilon^{-2}\ell \log \ell)$ samples and obtain a $\epsilon/(4\ell)$ -additive approximations $\tilde{\mathbf{p}}_{\langle \mathcal{R} \rangle}$ and $\tilde{\mathbf{q}}_{\langle \mathcal{R} \rangle}$ of $\mathbf{p}_{\langle \mathcal{R} \rangle}$ and $\mathbf{q}_{\langle \mathcal{R} \rangle}$, respectively;
- 11 **if** $\|\tilde{\mathbf{p}}_{\langle \mathcal{R} \rangle} - \tilde{\mathbf{q}}_{\langle \mathcal{R} \rangle}\|_1 > 3\epsilon/2$ **then output REJECT**;
- 12 **output ACCEPT**;

Figure 2.2: The algorithm for tolerant identity testing

if S_i is such that $|S_i| = \Theta(\sqrt{n}\epsilon^{-4} \log \ell)$ then E_1 occurs with probability at least $8/9$. Let E_2 be the event that for every i in $[\ell]$ we have that $|(S_i|/|S|) - \mathbf{p}(R_i)| \leq \epsilon/(2\ell)$. By Hoeffding's inequality E_2 occurs with probability at least $8/9$ for $|S| = \tilde{\Omega}(\ell^2\epsilon^{-2})$. Let E_3 be the event that $\tilde{\mathbf{p}}_{\langle \mathcal{R} \rangle}$ and $\tilde{\mathbf{q}}_{\langle \mathcal{R} \rangle}$ are $\epsilon/(2\ell)$ -additive approximations of $\mathbf{p}_{\langle \mathcal{R} \rangle}$ and $\mathbf{q}_{\langle \mathcal{R} \rangle}$, respectively. By taking $\Theta(\epsilon^{-2}\ell^2 \log \ell)$ samples, E_3 occurs with probability at least $8/9$.

Let \mathbf{p} and \mathbf{q} be as described in Case 1, i.e. $\|\mathbf{p} - \mathbf{q}\|_1 > 13\epsilon$. Suppose the algorithm accepts \mathbf{p} and \mathbf{q} . Conditioned on $E_1 \cap E_3$, this implies that for each partition R_i for which steps (7) - (9) were performed, which are those for which $\mathbf{q}(R_i) \geq \epsilon/\ell$, we have $\|\mathbf{p}_{|R_i}\|_2^2 \leq \frac{(1+\epsilon^2)^2}{|R_i|} \cdot \frac{1}{1-\epsilon^2}$, which is at most $\frac{1+4\epsilon^2}{|R_i|}$ for $0 < \epsilon \leq 1/3$. Thus, by Lemma 15 it follows that,

$$\|\mathbf{p}_{|R_i} - U_{|R_i}\|_2^2 = \|\mathbf{p}_{|R_i}\|_2^2 - \|U_{|R_i}\|_2^2 \leq \frac{4\epsilon^2}{|R_i|}. \quad (2.123)$$

From the bucketing definition we have that for every $i \in [\ell]$,

$$\|\mathbf{q}_{|R_i} - U_{|R_i}\|_2^2 \leq \frac{\epsilon^2}{|R_i|} \leq \frac{\epsilon^2}{|R_i|}. \quad (2.124)$$

By the triangle inequality we obtain from Equations (2.123) and (2.124) that $\|\mathbf{p}_{|R_i} - \mathbf{q}_{|R_i}\|_2^2 \leq \frac{9\epsilon^2}{|R_i|}$ and thus $\|\mathbf{p}_{|R_i} - \mathbf{q}_{|R_i}\|_1 \leq 3\epsilon$. We also have that the sum of $\mathbf{q}(R_i)$ over all R_i for which steps (7) - (9) were not performed is at most $\ell \cdot (\epsilon/\ell) + n \cdot (\epsilon(1 + \epsilon)^2/n) < 4\epsilon$. For those R_i we use the trivial bound $\|\mathbf{p}_{|R_i} - \mathbf{q}_{|R_i}\|_1 \leq 2$. Also, $\|\mathbf{p}_{\langle \mathcal{R} \rangle} - \mathbf{q}_{\langle \mathcal{R} \rangle}\|_1 \leq 2\epsilon$ by step (11). So by Lemma 16 we get that $\|\mathbf{p} - \mathbf{q}\|_1 \leq 13\epsilon$

in contradiction to our assumption. Therefore, the test accepts \mathbf{p} and \mathbf{q} with probability at most $1/3$ (the bound on the probability of $\bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_3$).

We next turn to proving the second item in the theorem. Suppose \mathbf{q} is a $(\epsilon/n, (\epsilon/24))$ -multiplicative approximation of some \mathbf{q}' such that \mathbf{p} is $\frac{72\epsilon^2}{\ell\sqrt{n}}$ -close to \mathbf{q}' . From the bucketing definition we have that for every $i \in [\ell]$ and for every $x \in R_i$,

$$\frac{1}{(1 + (\epsilon/24))} \cdot \frac{\mathbf{q}(R_i)}{|R_i|} \leq \mathbf{q}(x) \leq (1 + (\epsilon/24)) \cdot \frac{\mathbf{q}(R_i)}{|R_i|}. \quad (2.125)$$

Since \mathbf{q} is a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of \mathbf{q}' , we get by Lemma 17 that for every $R_i \subseteq H$ and every $x \in H$,

$$\frac{\mathbf{q}'(x)}{\mathbf{q}'(R_i)} \cdot \frac{(1 - (\epsilon/24))}{(1 + (\epsilon/24))} \leq \frac{\mathbf{q}(x)}{\mathbf{q}(R_i)} \leq \frac{\mathbf{q}'(x)}{\mathbf{q}'(R_i)} \cdot \frac{(1 + (\epsilon/24))}{(1 - (\epsilon/24))} \quad (2.126)$$

Combining Equations (2.125) and (2.126) we get that

$$\frac{(1 - (\epsilon/24))}{(1 + (\epsilon/24))^2} \cdot \frac{\mathbf{q}'(R_i)}{|R_i|} \leq \mathbf{q}'(x) \leq \frac{(1 + (\epsilon/24))^2}{(1 - (\epsilon/24))} \cdot \frac{\mathbf{q}'(R_i)}{|R_i|}, \quad (2.127)$$

and thus for $0 < \epsilon \leq 1/2$,

$$\frac{(1 - (\epsilon/2))}{|R_i|} \leq \frac{\mathbf{q}'(x)}{\mathbf{q}'(R_i)} \leq \frac{(1 + (\epsilon/2))}{|R_i|}. \quad (2.128)$$

By Equation (2.128) we obtain that

$$\|\mathbf{q}'_{|R_i} - U_{|R_i}\|_2 \leq \epsilon/(2\sqrt{|R_i|}). \quad (2.129)$$

For all subsets $R_i \subseteq H$ with $\mathbf{q}(R_i) \geq \epsilon/\ell$ we have that $\mathbf{q}'(R_i) \geq \epsilon/(1 + \epsilon)\ell$, combined with the fact that $\|\mathbf{p} - \mathbf{q}'\|_1 \leq \frac{72\epsilon^2}{\ell\sqrt{n}}$ we get by Lemma 14 (for sufficiently large n) that,

$$\|\mathbf{p}_{|R_i} - \mathbf{q}'_{|R_i}\|_1 \leq \epsilon/(2\sqrt{n}). \quad (2.130)$$

This implies that

$$\|\mathbf{p}_{|R_i} - \mathbf{q}'_{|R_i}\|_2 \leq \|\mathbf{p}_{|R_i} - \mathbf{q}'_{|R_i}\|_1 \leq \epsilon/(2\sqrt{n}) < \epsilon/(2\sqrt{|R_i|}). \quad (2.131)$$

Then by the triangle inequality we get that,

$$\|\mathbf{p}_{|R_i} - U_{|R_i}\|_2 \leq \|\mathbf{p}_{|R_i} - \mathbf{q}'_{|R_i}\|_2 + \|\mathbf{q}'_{|R_i} - U_{|R_i}\|_2 \leq \epsilon/\sqrt{|R_i|}. \quad (2.132)$$

Therefore, by Lemma 15 it follows that,

$$\|\mathbf{p}_{|R_i}\|_2^2 = \|\mathbf{p}_{|R_i} - U_{|R_i}\|_2^2 + \|U_{|R_i}\|_2^2 \leq (1 + \epsilon^2)/|R_i|. \quad (2.133)$$

Then conditioned on $E_1 \cap E_2$ the algorithm will pass on all such subsets; Since \mathbf{q} is $\epsilon/2$ -close to \mathbf{q}' , by the triangle inequality \mathbf{p} is ϵ -close to \mathbf{q} and thus conditioned on E_3 the algorithm will pass step (11) as well. Thus the algorithm will pass with probability at least $2/3$.

Finally, the sample complexity is $\tilde{O}(\sqrt{n}\epsilon^{-5})$ from step (2), which dominates the sample complexity of step (10). ■

2.3.2 Algorithm for Testing Tolerant Equivalence in the Known-Weights Sampling Model

In this section we prove Theorem 12. We note that in the proof of the theorem we essentially describe a tolerant tester for the property of independence of two random variables.

Theorem 12 *Let \mathcal{D} be a list of $[m]$ distributions over $[n]$ and let \vec{w} be a weight vector over $[m]$. Denote by $Q^{\mathcal{D}, \vec{w}}$ the joint distribution over $[n] \times [m]$ such that $Q^{\mathcal{D}, \vec{w}}(i, j) = w_j \cdot D_j(i)$. There is a test that works in the Known-Weights sampling model, which takes $\tilde{O}((n^{1/2}m^{1/2} + n)\text{poly}(1/\epsilon))$ samples from \mathcal{D} , and whose output satisfies the following:*

- If \mathcal{D} is $\frac{\epsilon^2}{24\ell\sqrt{n}}$ -close from being in \mathcal{P}^{eq} , where $\ell = \log(n/\epsilon)/\log(1 + \epsilon)$, or if $Q^{\mathcal{D}, \vec{w}}$ is a $(\epsilon/n, \epsilon/120)$ -multiplicative approximation of $\pi_1 Q^{\mathcal{D}, \vec{w}} \times \pi_2 Q^{\mathcal{D}, \vec{w}}$, then the test accepts with probability at least $2/3$
- If \mathcal{D} is 19ϵ -far from being in \mathcal{P}^{eq} , then the test rejects with probability at least $2/3$.

In the proof of Theorem 12 we shall use the following lemma:

Lemma 18 *Let Q be a joint distribution over $[n] \times [m]$. Let \tilde{Q}^1 be a (α_1, β_1) -multiplicative approximation of $\pi_1 Q$. Let \tilde{Q}^2 be a (α_2, β_2) -multiplicative approximation of $\pi_2 Q$. Denote by A_1 the set of all $i \in [n]$ such that $\tilde{Q}^1(i) \geq \alpha_1(1 + \beta_1)$. Denote by A_2 the set of all $j \in [m]$ such that $\tilde{Q}^2(j) \geq \alpha_2(1 + \beta_2)$. For every $B_1 \subseteq A_1$ and every $B_2 \subseteq A_2$, $(\tilde{Q}^1 \times \tilde{Q}^2)_{|B_1 \times B_2}$ is a $(0, \frac{2(\beta_1 + \beta_2)}{(1 - \beta_1) \cdot (1 - \beta_2)})$ -multiplicative approximation of $(\pi_1 Q \times \pi_2 Q)_{|B_1 \times B_2}$.*

Proof: For every $(i, j) \in B_1 \times B_2$ we have that

$$\pi_1 Q(i) \cdot \pi_2 Q(j) \cdot (1 - \beta_1) \cdot (1 - \beta_2) \leq \tilde{Q}^1(i) \cdot \tilde{Q}^2(j) \leq \pi_1 Q(i) \cdot \pi_2 Q(j) \cdot (1 + \beta_1) \cdot (1 + \beta_2). \quad (2.134)$$

From the facts that $\frac{(1 + \beta_1) \cdot (1 + \beta_2)}{(1 - \beta_1) \cdot (1 - \beta_2)} = 1 + \frac{2(\beta_1 + \beta_2)}{(1 - \beta_1) \cdot (1 - \beta_2)}$ and $\frac{(1 - \beta_1) \cdot (1 - \beta_2)}{(1 + \beta_1) \cdot (1 + \beta_2)} > 1 - \frac{2(\beta_1 + \beta_2)}{(1 - \beta_1) \cdot (1 - \beta_2)}$, and from Lemma 17 the lemma follows. ■

Proof of Theorem 12: The test referred to in the statement of the theorem is **Test-Tolerant-Equivalence-Known-Weights** (Algorithm 3 in Figure 2.3). Let E_1 be the event that \tilde{Q}^1 is a $(\epsilon/n, \epsilon/120)$ -multiplicative approximation of $\pi_1 Q$, as defined in Definition 2. By applying Chernoff's inequality and the union bound, E_1 occurs with probability at least $8/9$ (for a sufficiently large constant in the $\Theta(\cdot)$ notation for the sample size). By Lemma 18, conditioned on E_1 , we have that $(\tilde{Q}^1 \times \vec{w})_{|H \times [m]}$ is a $(0, \epsilon/24)$ -multiplicative approximation of $(\pi_1 Q \times \pi_2 Q)_{|H \times [m]}$. Thus, $\left\| (\tilde{Q}^1 \times \vec{w})_{|H \times [m]} - (\pi_1 Q \times \vec{w})_{|H \times [m]} \right\|_1 \leq \epsilon$. Let E_2 be the event that the application of Test-Tolerant-Identity returned a correct answer, as defined by Theorem 10. We run the amplified version of Test-Tolerant-Identity, therefore the additional parameter, which is the confidence parameter is set to $1/9$, i.e. E_2 occurs with probability at least $8/9$.

Algorithm 3: Test-Tolerant-Equivalence-Known-Weights

Input: Parameter $0 < \epsilon \leq 1/3$, sampling access to a list of distributions, \mathcal{D} , over $[n]$, in the Known-Weights sampling model

- 1 Let Q denote $Q^{\mathcal{D}, \vec{w}}$;
- 2 Take $\Theta(\epsilon^{-3} n \log n)$ samples and obtain a $(\epsilon/n, \epsilon/120)$ -multiplicative approximation, \tilde{Q}^1 , of $\pi_1 Q$;
- 3 Let H be the set of all $i \in [n]$ such that $\tilde{Q}^1(i) > \epsilon(1 + \epsilon)/n$ and let L be $[n] \setminus H$;
- 4 **if** $\text{Test-Tolerant-Identity}(Q_{H \times [m]}, (\tilde{Q}^1 \times \vec{w})_{|H \times [m]}, |H| \cdot m, \epsilon, 1/9) = \text{REJECT}$
then output REJECT;
- 5 $\mathcal{I} \stackrel{\text{def}}{=} \{H \times [m], L \times [m]\}$;
- 6 Take $\Theta(\epsilon^{-2})$ samples and obtain a $(\epsilon/2)$ -additive approximations $\tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2}$ and $\tilde{Q}_{\langle \mathcal{I} \rangle}$ of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle}$ and $Q_{\langle \mathcal{I} \rangle}$, respectively;
- 7 **if** $\|\tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2} - \tilde{Q}_{\langle \mathcal{I} \rangle}\|_1 > 2\epsilon$ **then output REJECT**;
- 8 **output ACCEPT**;

Figure 2.3: The algorithm for testing tolerant equivalence in the known-weights sampling model

Let \mathcal{D} be 19ϵ -far from being in \mathcal{P}^{eq} and assume the test accepts. Conditioned on E_2 this implies that $\|Q_{|H \times [m]} - (\tilde{Q}^1 \times \vec{w})_{|H \times [m]}\|_1 \leq 13\epsilon$. By the triangle inequality, we obtain that conditioned on $E_1 \cap E_2$,

$$\|Q_{|H \times [m]} - (\pi_1 Q \times \vec{w})_{|H \times [m]}\|_1 \leq \epsilon + 13\epsilon < 14\epsilon. \quad (2.135)$$

From the fact that $Q(L \times [m]) \leq \epsilon$ we have that

$$Q(L \times [m]) \cdot \|Q_{L \times [m]} - (\pi_1 Q \times \vec{w})_{L \times [m]}\|_1 \leq 2\epsilon. \quad (2.136)$$

Let E_3 be the event that $\tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2}$ and $\tilde{Q}_{\langle \mathcal{I} \rangle}$ are $\epsilon/2$ -additive approximations of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle}$ and $Q_{\langle \mathcal{I} \rangle}$, respectively. By taking $\Theta(\epsilon^{-2})$ samples, E_3 occurs with probability at least $8/9$. Conditioned on E_3 , we have that,

$$\|(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle} - Q_{\langle \mathcal{I} \rangle}\|_1 \leq 3\epsilon. \quad (2.137)$$

Combining Equations (2.135) - (2.137), by Lemma 16, we have that

$$\|(\pi_1 Q \times \pi_2 Q) - Q\|_1 \leq 3\epsilon + 14\epsilon + 2\epsilon = 19\epsilon. \quad (2.138)$$

Hence \mathcal{D} is 19ϵ -close to being in \mathcal{P}^{eq} , in contradiction to our assumption, thus the test accepts with probability at most $1/3$.

On the other hand, let \mathcal{D} be $\frac{\epsilon^2}{24\ell\sqrt{n}}$ -close from being in \mathcal{P}^{eq} or $\pi_1 Q^{\mathcal{D}, \vec{w}} \times \pi_2 Q^{\mathcal{D}, \vec{w}}$ is a $(\epsilon/n, \epsilon/120)$ -multiplicative approximation of $Q^{\mathcal{D}, \vec{w}}$ and assume the test rejects. In case the test rejects on Step (4) then

conditioned on E_2 , we get by Theorem 10 that $(\tilde{Q}^1 \times \vec{w})_{|H \times [m]}$ is not a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of some \mathbf{q}' such that $\|Q_{|H \times [m]} - \mathbf{q}'\|_1 \leq \frac{72\epsilon^2}{\ell\sqrt{n}}$. Conditioned on E_1 , we have that $(\tilde{Q}^1 \times \vec{w})_{|H \times [m]}$ is a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of $(\pi_1 Q \times \vec{w})_{|H \times [m]}$. Thus, conditioned on $E_1 \cap E_2$, we obtain that $\|Q - \pi_1 Q \times \vec{w}\|_1 > \frac{72\epsilon^2}{\ell\sqrt{n}}$. By Proposition 1 in [BFF⁺01] this implies that \mathcal{D} is $\frac{24\epsilon^2}{\ell\sqrt{n}}$ -far from being in \mathcal{P}^{eq} . By setting $\mathbf{q}' = Q_{|H \times [m]}$ we also have that $(\tilde{Q}^1 \times \vec{w})_{|H \times [m]}$ is not a $(\epsilon/n, \epsilon/24)$ -multiplicative approximation of $Q_{|H \times [m]}$. For the sake of simplicity, denote $(\tilde{Q}^1 \times \vec{w})_{|H \times [m]}$ by A and $(\pi_1 Q \times \vec{w})_{|H \times [m]}$ by B . So, there exists $(i, j) \in H \times [m]$ that satisfies either

$$A_{|H \times [m]}(i, j) > (1 + (\epsilon/24))B_{|H \times [m]}(i, j) \quad (2.139)$$

or

$$A_{|H \times [m]}(i, j) < (1 - (\epsilon/24))B_{|H \times [m]}(i, j). \quad (2.140)$$

By Lemma 18, we get that $A_{|H \times [m]}$ is a $(0, \epsilon/30)$ -multiplicative approximation of $B_{|H \times [m]}$. Therefore, by Equations (2.139) and (2.140), either it holds that

$$Q_{|H \times [m]}(i, j) < \frac{1 + (\epsilon/30)}{1 + (\epsilon/24)}B_{|H \times [m]}(i, j) \quad (2.141)$$

or that,

$$Q_{|H \times [m]}(i, j) > \frac{1 - (\epsilon/30)}{1 - (\epsilon/24)}B_{|H \times [m]}(i, j). \quad (2.142)$$

Since $Q(H \times [m]) = B(H \times [m])$, we obtain from Equations (2.141) and (2.142) that either $Q(i, j) < \frac{1 + (\epsilon/30)}{1 + (\epsilon/24)}B(i, j)$ or $Q(i, j) > \frac{1 - (\epsilon/30)}{1 - (\epsilon/24)}B(i, j)$, which by a simple calculation implies that Q is not a $(\epsilon/n, \epsilon/120)$ -multiplicative approximation of $\pi_1 Q \times \vec{w}$.

Alternatively, in case the test reject on Step (7) then by the triangle inequality we get that conditioned on E_3 , Q is ϵ -far from $\pi_1 Q \times \pi_2 Q$. In both cases we get a contradiction to our assumption and therefore the algorithm accepts \mathcal{D} with probability at most $1/3$ (which is the upper bound on the probability of $\bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_3$).

The sample complexity of Step (4) is bounded by $\tilde{O}(n^{1/2}m^{1/2}\text{poly}(\epsilon^{-1}))$ so the overall sample complexity is $\tilde{O}((n^{1/2}m^{1/2} + n)\text{poly}(\epsilon^{-1}))$. ■

2.3.3 Algorithm for Testing Tolerant Equivalence in the Unknown-Weights Sampling Model

In this section we prove the following theorem:

Theorem 13 *Let \mathcal{D} be a list of m distributions over $[n]$. It is possible to distinguish between the case that \mathcal{D} is $\frac{36\epsilon^3}{\ell\sqrt{n}}$ -close to being in \mathcal{P}^{eq} , where $\ell = \log(n/\epsilon)/\log(1 + \epsilon)$ and the case that \mathcal{D} is 25ϵ -far from being in \mathcal{P}^{eq} in the unknown-weights sampling model using a sample of size $\tilde{O}((n^{2/3}m^{1/3} + m) \cdot \text{poly}(1/\epsilon))$.*

Proof of Theorem 13: The algorithm referred to in the statement of the theorem is **Test-Tolerant-Equivalence-Unknown-Weights** (given in Figure 2.4). Let E_1 to be the event that \tilde{Q}^1 is a $(\epsilon/n^{2/3}m^{1/3}, \epsilon/250)$ -multiplicative approximation of $\pi_1 Q$. For a sample of size $\Theta(\epsilon^{-3}n^{2/3}m^{1/3} \log n)$, we get, by Chernoff's inequality, that E_1 occurs with probability at least $20/21$. Let E_2 be the event that \tilde{Q}^2 is a $(\epsilon/m, \epsilon/250)$ -multiplicative approximation of $\pi_2 Q$. By taking sample of size $\Theta(\epsilon^{-3}m \log m)$, E_2 occurs with probability at least $20/21$. By Lemma (18), for every $0 < \epsilon \leq 1/3$, we get, condition on $E_1 \cap E_2$, that $(\tilde{Q}^1 \times \tilde{Q}^2)_{|H_n \times H_m}$ is a $(0, \epsilon/24)$ -multiplicative approximation of $(\pi_1 Q \times \pi_2 Q)_{|H_n \times H_m}$. Thus, conditioned on $E_1 \cap E_2$, we have that

$$\left\| (\tilde{Q}^1 \times \tilde{Q}^2)_{|H_n \times H_m} - (\pi_1 Q \times \pi_2 Q)_{|H_n \times H_m} \right\|_1 \leq \epsilon. \quad (2.143)$$

Let E_3 be the event that the application of Test-Tolerant-Identity returned a correct answer, as defined by Theorem 10. E_3 occurs with probability at least $20/21$.

Let \mathcal{D} be 25-far from being in \mathcal{P}^{eq} and assume the algorithm accepts. Then either Test-Tolerant-Identity returns accept or $\tilde{Q}_{|H_n \times H_m} < 3\epsilon/2$. Consider the case that Test-Tolerant-Identity returns accept. Conditioned on E_3 , by Theorem 10, we have that $\left\| (\tilde{Q}^1 \times \tilde{Q}^2)_{|H_n \times H_m} - Q_{|H_n \times H_m} \right\|_1 \leq 13\epsilon$, then by the triangle inequality and Equation (2.143) we obtain that

$$\left\| (\pi_1 Q \times \pi_2 Q)_{|H_n \times H_m} - Q_{|H_n \times H_m} \right\|_1 \leq 13\epsilon + \epsilon = 14\epsilon. \quad (2.144)$$

Consider the case $\tilde{Q}_{|H_n \times H_m} < 3\epsilon/2$. Let E_7 be the event that $\left| \tilde{Q}_{H_n \times H_m} - Q(H_n \times H_m) \right| \leq \epsilon/2$. By taking $\Theta(\epsilon^{-2})$ samples, E_7 occurs with probability at least $20/21$. Then we have that

$$Q(H_n \times H_m) \leq \epsilon. \quad (2.145)$$

Let E_4 be the event that all applications of Bounded- ℓ_∞ -Closeness-Test returned a correct answer, as defined by Theorem 7. By the union bound, E_4 occurs with probability at least $20/21$. Conditioned on E_4 , we obtain that every R_i that passes step (14) satisfies the following

$$\left\| (\pi_1 Q \times \pi_2 Q)_{|L_n \times R_i} - Q_{|L_n \times R_i} \right\|_1 \leq \epsilon. \quad (2.146)$$

Let E_5 to be the event that for every i in $[\ell]$ we have that $||S_i|/|S| - Q(R_i \times L_n)| \leq \epsilon/(2\ell)$. By Hoeffding's inequality E_5 occurs with probability at least $20/21$ for $|S| = \tilde{\Omega}(\ell^2 \epsilon^{-2})$. From the fact that for every R_i that doesn't enter step (14) we have that $|S_i|/|S| < \epsilon/\ell$, we obtain, conditioned on E_5 , that

$$Q(L \times R_i) \leq 2\epsilon/\ell. \quad (2.147)$$

Let E_6 be the event that $\tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2}$ and $\tilde{Q}_{\langle \mathcal{I} \rangle}$ are $\epsilon/(2\ell)$ -additive approximations of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle}$ and $Q_{\langle \mathcal{I} \rangle}$, respectively. By taking $\Theta(\epsilon^{-2} \ell^2 \log \ell)$ samples, E_6 occurs with probability at least $20/21$. Conditioned on E_6 , we have that,

$$\left\| (\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle} - Q_{\langle \mathcal{I} \rangle} \right\|_1 \leq 3\epsilon. \quad (2.148)$$

Conditioned on $E_1 \cap E_2$, for $0 < \epsilon \leq 1/5$ we have that

$$Q(H_m \times L_m) \leq 3\epsilon/2. \quad (2.149)$$

For every $I \in \mathcal{I}$ we have the following trivial bound

$$\|(\pi_1 Q \times \pi_2 Q)|_I - Q|_I\|_1 \leq 2. \quad (2.150)$$

Combining Equations (2.144) - (2.150), by Lemma 16, we have that

$$\|(\pi_1 Q \times \pi_2 Q) - Q\|_1 \leq 3\epsilon + 14\epsilon + \epsilon + \ell \cdot (2\epsilon/\ell) \cdot 2 + 3/2 \cdot 2 = 25\epsilon. \quad (2.151)$$

Therefore, \mathcal{D} is 25ϵ -close to being in \mathcal{P}^{eq} in contradiction to our assumption. It follows that the algorithm accepts \mathcal{D} with probability at most $1/3$.

On the other hand, let \mathcal{D} be $\frac{36\epsilon^3}{\ell\sqrt{n}}$ -close to being in \mathcal{P}^{eq} and assume the algorithm rejects. Conditioned on $E_1 \cap E_2$, we have that $(\tilde{Q}^1 \times \tilde{Q}^2)|_{H_n \times H_m}$ is a $(0, \epsilon/24)$ -multiplicative approximation of $(\pi_1 Q \times \pi_2 Q)|_{H_n \times H_m}$. Therefore, conditioned on $E_1 \cap E_2 \cap E_3 \cap E_7$, if we reject on step (9), then we obtain by Theorem 10 that

$$\|Q|_{H_n \times H_m} - (\pi_1 Q \times \pi_2 Q)|_{H_n \times H_m}\|_1 > 72 \cdot \frac{\epsilon^2}{\ell\sqrt{n}}. \quad (2.152)$$

It follows, by Lemma 14, that $\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \frac{\pi_1 Q(H_n) \cdot \pi_2 Q(H_m)}{2} \cdot 72 \cdot \frac{\epsilon^2}{\ell\sqrt{n}} \geq \frac{36\epsilon^3}{\ell\sqrt{n}}$. If we reject on step (14), then conditioned on $E_4 \cap E_5$, there is R_i such that $Q(L_n \times R_i) \geq \epsilon/\ell$ in which the following holds,

$$\|(\pi_1 Q \times \pi_2 Q)|_{L_n \times R_i} - Q|_{L_n \times R_i}\|_1 > \epsilon/(2\sqrt{n}). \quad (2.153)$$

Thus, by Lemma 14, $\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \frac{Q(L_n \times R_i)}{2} \cdot \epsilon/(2\sqrt{n}) \geq \epsilon^2/(4\ell\sqrt{n})$. If we reject on step (17), then conditioned on E_6 it follows that $\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \epsilon$. Then we get a contradiction to our assumption which implies that the algorithm accepts \mathcal{D} with probability at least $2/3$. We note that we run the amplified version of Test-Tolerant-Identity and Bounded- ℓ_∞ -Closeness-Test and that the additional parameter in the application of Test-Tolerant-Identity and Bounded- ℓ_∞ -Closeness-Test is the confidence parameter. To achieve $(1 - \delta)$ confidence, the amplified algorithm takes the majority result of $\Theta(\log 1/\delta)$ applications of the original algorithm. In addition, both algorithms are applied on restricted domains ($H_n \times H_m$ in Test-Tolerant-Identity and $L_n \times R_i$ in Bounded- ℓ_∞ -Closeness-Test). This affects the sample complexity only by a factor of $\text{poly}(1/\epsilon, \log n)$. Therefore, the sample complexity is $\tilde{O}((n^{2/3}m^{1/3} + m) \cdot \text{poly}(1/\epsilon))$ as required. ■

2.4 A Lower Bound of $\Omega(n^{1/2}m^{1/2})$ for Testing Equivalence in the Uniform Sampling Model

In this section we prove the following theorem:

Theorem 14 *Testing the property $\mathcal{P}_{m,n}^{\text{eq}}$ in the uniform sampling model for every $\epsilon \leq 1/2$ and $m \geq 64$ requires $\Omega(n^{1/2}m^{1/2})$ samples.*

We assume without loss of generality that n is even (or else, we set the probability weight of the element n to 0 in all distributions considered, and work with $n - 1$ that is even). Define \mathcal{H}_n to be the set of all distributions over $[n]$ that have probability $\frac{2}{n}$ on exactly half of the elements and 0 on the other half. Define \mathcal{H}_n^m to be the set of all possible lists of m distributions from \mathcal{H}_n . Define \mathcal{U}_n^m to consist of a single list of m distributions that are identical to U_n , where U_n denotes the uniform distribution over $[n]$. Thus the single list in \mathcal{U}_n^m belongs to $\mathcal{P}_{m,n}^{\text{eq}}$. On the other hand we show that \mathcal{H}_n^m contains mostly lists of distributions that are $\Omega(1)$ -far from $\mathcal{P}_{m,n}^{\text{eq}}$. However, we also show that any tester in the uniform sampling model that takes less than $n^{1/2}m^{1/2}/6$ samples can't distinguish between \mathcal{D} that was uniformly drawn from \mathcal{H}_n^m and $\mathcal{D} = (U_n, \dots, U_n) \in \mathcal{U}_n^m$. Details follow.

Lemma 19 *For every $m \geq 3$, with probability at least $\left(1 - \frac{2}{\sqrt{m}}\right)$ over the choice of $\mathcal{D} \in \mathcal{H}_n^m$ we have that \mathcal{D} is $(1/2)$ -far from $\mathcal{P}_{m,n}^{\text{eq}}$.*

Proof: We need to prove that with probability at least $\left(1 - \frac{2}{\sqrt{m}}\right)$ over the choice of $\mathcal{D} \in \mathcal{H}_n^m$, for every $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$ which corresponds to a distribution (i.e., $v_i \geq 0$ for every $i \in [n]$ and $\sum_{i=1}^n v_i = 1$),

$$\frac{1}{m} \sum_{j=1}^m \|D_j - \mathbf{v}\|_1 > \frac{1}{2}. \quad (2.154)$$

We shall actually prove a slightly more general statement. Namely, that Equation (2.154) holds for every vector $\mathbf{v} \in \mathbb{R}^n$. We define the function, $\text{med}^{\mathcal{D}} : [n] \rightarrow [0, 1]$, such that $\text{med}^{\mathcal{D}}(i) = \mu_{\frac{1}{2}}(D_1(i), \dots, D_m(i))$, where $\mu_{\frac{1}{2}}(x_1, \dots, x_m)$ denotes the median of x_1, \dots, x_m (where if m is even, it is the value in position $\frac{m}{2}$ in sorted non-decreasing order). The sum $\sum_{i=1}^m |x_i - c|$ is minimized when $c = \mu_{\frac{1}{2}}(x_1, \dots, x_m)$. Therefore, for every \mathcal{D} and every vector $\mathbf{v} \in \mathbb{R}^n$,

$$\sum_{j=1}^m \|D_j - \text{med}^{\mathcal{D}}\|_1 \leq \sum_{j=1}^m \|D_j - \mathbf{v}\|_1. \quad (2.155)$$

Recall that for every $\mathcal{D} = (D_1, \dots, D_m)$ in \mathcal{H}_n^m , and for each $(i, j) \in [n] \times [m]$, we have that either $D_j(i) = \frac{2}{n}$, or $D_j(i) = 0$. Thus, $\text{med}^{\mathcal{D}}(i) = 0$ when $D_j(i) = 0$ for at least half of the j 's in $[m]$ and $\text{med}^{\mathcal{D}}(i) = \frac{2}{n}$ otherwise. We next show that for every $(i, j) \in [n] \times [m]$, the probability over $\mathcal{D} \in \mathcal{H}_n^m$ that $D_j(i)$ will have the same value as $\text{med}^{\mathcal{D}}(i)$ is just a little bit bigger than half. More precisely, we show that for every $(i, j) \in [n] \times [m]$:

$$\Pr_{\mathcal{D} \in \mathcal{H}_n^m} [D_j(i) \neq \text{med}^{\mathcal{D}}(i)] \geq \frac{1}{2} \left(1 - \frac{1}{\sqrt{m}}\right). \quad (2.156)$$

Fix $(i, j) \in [n] \times [m]$, and consider selecting \mathcal{D} uniformly at random from \mathcal{H}_n^m . Suppose we first determine the values $D_{j'}(i)$ for $j' \neq j$, and set $D_j(i)$ in the end. For each (i, j') the probability that $D_{j'}(i) = 0$ is $1/2$, and the probability that $D_{j'}(i) = \frac{2}{n}$ is $1/2$. If more than $m/2$ of the outcomes are 0, or more than $m/2$ are

$\frac{2}{n}$, then the value of $med^{\mathcal{D}}(i)$ is already determined. Conditioned on this we have that the probability that $D_j(i) \neq med^{\mathcal{D}}(i)$ is exactly $1/2$. On the other hand, if at most $m/2$ are 0 and at most $m/2$ are $\frac{2}{n}$ (that is, for odd m there are $(m-1)/2$ that are 0 and $(m-1)/2$ that are $\frac{2}{n}$, and for even m there are $m/2$ of one kind and $(m/2) - 1$ of the other) then necessarily $med^{\mathcal{D}}(i) = D_j(i)$. We thus bound the probability of this event. First consider the case that m is odd (so that $m-1$ is even).

$$\Pr \left[Bin \left(x, \frac{1}{2} \right) = \frac{x}{2} \right] = \binom{x}{\frac{x}{2}} \cdot \frac{1}{2^x} = \frac{x!}{\frac{x!}{2! \cdot \frac{x!}{2!}} \cdot 2^x} \cdot \frac{1}{2^x} \quad (2.157)$$

By Stirling's approximation, $x! = \sqrt{2\pi x} \left(\frac{x}{e}\right)^x e^{\lambda_x}$, where $\frac{1}{12x+1} < \lambda_x < \frac{1}{12x}$, thus,

$$\frac{x!}{\frac{x!}{2! \cdot \frac{x!}{2!}} \cdot 2^x} < \frac{\sqrt{2\pi x} \left(\frac{x}{e}\right)^x e^{\frac{1}{12x}}}{\left(\sqrt{2\pi x/2} \left(\frac{x/2}{e}\right)^{x/2} e^{\frac{1}{12x/2+1}}\right)^2} \cdot \frac{1}{2^x} \quad (2.158)$$

$$= \frac{e^{\frac{1}{12x} - \frac{2}{6x+1}}}{\sqrt{\pi x/2}} \quad (2.159)$$

$$< \frac{1}{\sqrt{\pi x/2}} \quad (2.160)$$

$$\leq \frac{1}{\sqrt{m}}, \quad (2.161)$$

where Inequalities (2.160) and (2.161) hold for $m \geq 3$. In case m is even, the probability (over the choice of $D_{j'}(i)$ for $j' \neq j$) that $med^{\mathcal{D}}(i)$ is determined by $D_j(i)$ is $\Pr \left[Bin \left(x, \frac{1}{2} \right) = \frac{x+1}{2} \right] \leq \Pr \left[Bin \left(x, \frac{1}{2} \right) = \frac{x}{2} \right]$. Hence, Equation (2.156) holds for all m and we obtain that

$$\mathbb{E}_{\mathcal{D} \in \mathcal{H}_n^m} \left[\sum_{j=1}^m \|D_j - med^{\mathcal{D}}\|_1 \right] = \sum_{i=1}^m \sum_{j=1}^n \mathbb{E}_{\mathcal{D} \in \mathcal{H}_n^m} [|D_j(i) - med^{\mathcal{D}}(i)|] \quad (2.162)$$

$$= m \cdot n \cdot \Pr_{\mathcal{D} \in \mathcal{H}_n^m} [D_j(i) \neq med^{\mathcal{D}}(i)] \cdot \frac{2}{n} \quad (2.163)$$

$$\geq m \cdot n \cdot \frac{1}{2} \left(1 - \frac{1}{\sqrt{m}} \right) \cdot \frac{2}{n} \quad (2.164)$$

$$= m - \sqrt{m}, \quad (2.165)$$

while,

$$\sum_{j=1}^m \|D_j - med^{\mathcal{D}}\|_1 = \sum_{i=1}^m \sum_{j=1}^n |D_j(i) - med^{\mathcal{D}}(i)| \quad (2.166)$$

$$\leq \sum_{j=1}^n \frac{m}{2} \frac{2}{n} \quad (2.167)$$

$$= m. \quad (2.168)$$

Assume for the sake of contradiction that

$$\Pr_{\mathcal{D} \in \mathcal{H}_n^m} \left[\sum_{j=1}^m \|D_j - med^{\mathcal{D}}\|_1 \leq m/2 \right] > \frac{2}{\sqrt{m}}, \quad (2.169)$$

then by Equation (2.168) we have,

$$\mathbb{E}_{\mathcal{D} \in \mathcal{H}_n^m} \left[\sum_{j=1}^m \|D_j - \text{med}^{\mathcal{D}}\|_1 \right] < \frac{2}{\sqrt{m}} \cdot \frac{m}{2} + \left(1 - \frac{2}{\sqrt{m}}\right) \cdot m \quad (2.170)$$

$$= m - \sqrt{m}, \quad (2.171)$$

which contradicts Equation (2.165). ■

Recall that for an element $i \in [n]$ and a distribution $D_j, j \in [m]$, we let $a_{i,j}$ denote the number of times the pair (i, j) appears in the sample (when the sample is selected in the uniform sampling model). Thus $(a_{i,1}, \dots, a_{i,m})$ is the *sample histogram* of the element i . Since the sample points are selected independently, a sample is simply the union of the histograms of the different elements, or equivalently, a matrix M in $\mathbb{N}^{n \times m}$.

Lemma 20 *Let \mathcal{U} be the distribution of the histogram of q samples taken from the uniform distribution over $[n] \times [m]$, and let \mathcal{H} be the distribution of the histogram of q samples taken from a random list of distributions in \mathcal{H}_n^m , then,*

$$\|\mathcal{U} - \mathcal{H}\|_1 \leq \frac{4q^2}{mn} \quad (2.172)$$

Proof: For every matrix $M \in \mathbb{N}^{n \times m}$, let A_M be the event of getting the histogram M ; For every $\vec{x} = (x_1, \dots, x_m) \in \mathbb{N}^m$, let $B_{\vec{x}}$ be the event of getting a histogram M such that for every $j \in [m]$, $\sum_{i \in [n]} M[i, j] = x_j$; Let C be the event of getting a histogram M such that there exists $(i, j) \in [n] \times [m]$ such that $M[i, j] \geq 2$; Let $V = \{B_{\vec{x}} : \Pr_{\mathcal{H}}(B_{\vec{x}} \cap \bar{C}) > 0\}$ (where \bar{C} denotes the event complementary to C). In order to bound the statistical distance between \mathcal{H} and \mathcal{U} , we use the fact that, for every $B_{\vec{x}} \in V$, given the occurrence of $B_{\vec{x}} \cap \bar{C}$, i.e., given the histogram projected on the first coordinate and given that there were no collisions, \mathcal{H} and \mathcal{U} are equivalent. More formally,

$$\|\mathcal{U} - \mathcal{H}\|_1 = \sum_{A_M \subseteq C} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| + \sum_{A_M \subseteq \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| \quad (2.173)$$

$$\leq \Pr_{\mathcal{U}}(C) + \Pr_{\mathcal{H}}(C) + \sum_{A_M \subseteq \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)|. \quad (2.174)$$

We start by bounding the third term in Equation (2.174).

$$\sum_{A_M \subseteq \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| = \sum_{B_{\vec{x}}} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| \quad (2.175)$$

$$= \sum_{B_{\vec{x}} \in V} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| \quad (2.176)$$

$$+ \sum_{B_{\vec{x}} \in \bar{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)|. \quad (2.177)$$

We next bound the expression in Equation (2.176).

$$\begin{aligned} & \sum_{B_{\vec{x}} \in V} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| \\ &= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} \Pr_{\mathcal{U}}(A_M | B_{\vec{x}} \cap \bar{C}) \cdot |\Pr_{\mathcal{U}}(\bar{C} | B_{\vec{x}}) - \Pr_{\mathcal{H}}(\bar{C} | B_{\vec{x}})| \end{aligned} \quad (2.178)$$

$$= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) |\Pr_{\mathcal{U}}(\bar{C} | B_{\vec{x}}) - \Pr_{\mathcal{H}}(\bar{C} | B_{\vec{x}})| \quad (2.179)$$

$$= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) |(1 - \Pr_{\mathcal{U}}(C | B_{\vec{x}})) - (1 - \Pr_{\mathcal{H}}(C | B_{\vec{x}}))| \quad (2.180)$$

$$= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) |\Pr_{\mathcal{U}}(C | B_{\vec{x}}) - \Pr_{\mathcal{H}}(C | B_{\vec{x}})| \quad (2.181)$$

$$\leq \Pr_{\mathcal{U}}(C) + \Pr_{\mathcal{H}}(C) , \quad (2.182)$$

where in Equation (2.178) we used the fact that $\Pr_{\mathcal{U}}(B_{\vec{x}}) = \Pr_{\mathcal{H}}(B_{\vec{x}})$ for every $B_{\vec{x}}$, and that $\Pr_{\mathcal{U}}(A_M | B_{\vec{x}} \cap \bar{C}) = \Pr_{\mathcal{H}}(A_M | B_{\vec{x}} \cap \bar{C})$ for every $B_{\vec{x}} \in V$ and $M \in \mathbb{N}^{n \times m}$. Turning to the expression in Equation (2.177),

$$\sum_{B_{\vec{x}} \in \bar{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} |\Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M)| = \sum_{B_{\vec{x}} \in \bar{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \bar{C}} \Pr_{\mathcal{U}}(A_M) \quad (2.183)$$

$$\leq \sum_{B_{\vec{x}} \in \bar{V}} \Pr_{\mathcal{U}}(B_{\vec{x}}) \quad (2.184)$$

$$= \sum_{B_{\vec{x}} \in \bar{V}} \Pr_{\mathcal{H}}(B_{\vec{x}}) \quad (2.185)$$

$$= \sum_{B_{\vec{x}} \in \bar{V}} \Pr_{\mathcal{H}}(B_{\vec{x}} \cap C) \quad (2.186)$$

$$\leq \Pr_{\mathcal{H}}(C) . \quad (2.187)$$

We thus obtain that $\|\mathcal{U} - \mathcal{H}\|_1 \leq 2\Pr_{\mathcal{U}}(C) + 3\Pr_{\mathcal{H}}(C)$. If we take q uniform independent samples from $[\ell]$, then by a union bound over the q samples, the probability to get a collision is at most $\frac{1}{\ell} + \frac{2}{\ell} + \dots + \frac{q-1}{\ell}$ which is $\frac{q^2}{2\ell}$. Thus, $2\Pr_{\mathcal{U}}(C) + 3\Pr_{\mathcal{H}}(C) \leq 2 \cdot \frac{q^2}{2mn} + 3 \cdot \frac{q^2}{mn} = \frac{4q^2}{mn}$, and the lemma follows. ■

Proof of Theorem 14: Assume there is a tester, T , for the property $\mathcal{P}_{m,n}^{\text{eq}}$ in the uniform sampling model, which takes $q \leq m^{1/2}n^{1/2}/6$ samples. By Lemma 19,

$$\Pr_{\mathcal{D} \in \mathcal{H}_n^m} [A \text{ accepts } \mathcal{D}] \leq \frac{2}{\sqrt{m}} \cdot 1 + \left(1 - \frac{2}{\sqrt{m}}\right) \cdot \frac{1}{3} \quad (2.188)$$

$$= \frac{1}{3} \left(1 + \frac{4}{\sqrt{m}}\right) \quad (2.189)$$

$$\leq \frac{1}{2} \quad (2.190)$$

where the last inequality holds for $m \geq 64$. By Lemma 20, for $q \leq m^{1/2}n^{1/2}/6$, $\frac{1}{2}\|\mathcal{U} - \mathcal{H}\|_1 \leq \frac{1}{18}$, while by Equation (2.190), $(\Pr_{\mathcal{D} \in \mathcal{U}_n^m} [A \text{ accepts } \mathcal{D}] - \Pr_{\mathcal{D} \in \mathcal{H}_n^m} [A \text{ accepts } \mathcal{D}]) \geq \frac{2}{3} - \frac{1}{2} > \frac{1}{18}$. ■

Algorithm 4: Test-Tolerant-Equivalence-Unknown-Weights

Input: Parameter $0 < \epsilon \leq 1/8$, sampling access to a list of distributions, \mathcal{D} , over $[n]$, in the Unknown-Weights sampling model

- 1 Let Q denote $Q^{\mathcal{D}, \vec{w}}$;
- 2 Take $\Theta(\epsilon^{-3} n^{2/3} m^{1/3} \log n)$ samples and obtain a $(\epsilon/(n^{2/3} m^{1/3}), \epsilon/250)$ -multiplicative approximation \tilde{Q}^1 of $\pi_1 Q$;
- 3 Let H_n be the set of all $i \in [n]$ such that $\tilde{Q}^1(i) > \epsilon(1 + \epsilon)/(n^{2/3} m^{1/3})$ and let $L_n = [n] \setminus H_n$;
- 4 Take $\Theta(\epsilon^{-3} m \log m)$ samples and obtain a $(\epsilon/m, \epsilon/250)$ -multiplicative approximation \tilde{Q}^2 of $\pi_2 Q$;
- 5 $\mathcal{R} \stackrel{\text{def}}{=} \{R_0, \dots, R_\ell\} = \text{Bucket}(\tilde{Q}^2, m, (1 + \epsilon)\epsilon/m, \epsilon)$;
- 6 Let $L_m = R_0$ and let $H_m = [m] \setminus L_m$;
- 7 Take $\Theta(\epsilon^{-2})$ samples and let $\tilde{Q}_{H_n \times H_m}$ be the fraction of samples in $H_n \times H_m$;
- 8 **if** $\tilde{Q}_{H_n \times H_m} \geq 3\epsilon/2$ **then**
 - 9 **if** Test-Tolerant-Identity
 $(Q|_{H_n \times H_m}, (\tilde{Q}^1 \times \tilde{Q}^2)|_{H_n \times H_m}, |H| \cdot m, \epsilon, 1/21) = \text{REJECT}$ **then output REJECT**;
- 10 Let S be a set of $\tilde{\Theta}(\ell^2 \epsilon^{-2})$ samples;
- 11 **foreach** R_i **do**
 - 12 Let $S_i = S \cap (L_n \times R_i)$;
 - 13 **if** $|S_i|/|S| \geq \epsilon/\ell$ **then**
 - 14 **if** Bounded- ℓ_∞ -Closeness-Test
 $((\pi_1 Q \times \pi_2 Q)|_{L_n \times R_i}, Q|_{L_n \times R_i}, 1/(n^{2/3} m^{4/3}), 1/(n^{2/3} m^{1/3}), \epsilon, 1/(21\ell)) = \text{REJECT}$ **then output REJECT**;
- 15 $\mathcal{I} \stackrel{\text{def}}{=} \{H_n \times H_m, H_n \times L_m, L_n \times R_0, \dots, L_n \times R_\ell\}$;
- 16 Take $\Theta(\epsilon^{-2} \ell^2 \log \ell)$ samples and obtain a $\epsilon/(2\ell)$ -additive approximations $\tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2}$ and $\tilde{Q}_{\langle \mathcal{I} \rangle}$ of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle}$ and $Q_{\langle \mathcal{I} \rangle}$, respectively;
- 17 **if** $\left\| \tilde{Q}_{\langle \mathcal{I} \rangle}^{1 \times 2} - \tilde{Q}_{\langle \mathcal{I} \rangle} \right\|_1 > 2\epsilon$ **then output REJECT**;
- 18 **output ACCEPT**;

Figure 2.4: The algorithm for testing tolerant equivalence in the unknown-weights sampling model

Chapter 3

Clusterability Testing

3.1 Testing (k, β) -Clusterability in the Query Model

In this section we consider an extension of the property $\mathcal{P}_{m,n}^{\text{eq}}$ studied in the previous sections. Namely, rather than asking whether all distributions in a list \mathcal{D} are the same, we ask whether there exists a partition of \mathcal{D} into at most k lists, such that within each list all distributions are the the same (or close). That is, we are interested in the following a *clustering* problem:

Definition 6 Let \mathcal{D} be a list of m distributions over $[n]$. We say that \mathcal{D} is (k, β) -clusterable if there exists a partition of \mathcal{D} to k lists, $\{\mathcal{D}_i\}_{i=1}^k$ such that for every $i \in [k]$ and every $D, D' \in \mathcal{D}_i$, $\|D - D'\|_1 \leq \beta$.

In particular, for $k = 1$ and $\beta = 0$, we get the property $\mathcal{P}_{m,n}^{\text{eq}}$. We study testing (k, β) -clusterability (for $k \geq 1$) in the query model. The question for $k > 1$ in the (uniform) sampling model remains open.

We start by noting that if we allow a linear (or slightly higher) dependence on n , then it is possible (by adapting the algorithm we give below), to obtain a tester that works for any ϵ and β . The complexity of this tester is $\tilde{O}(n \cdot k \cdot \text{poly}(1/\epsilon))$. However, if we want a dependence on n that grows slower than $n^{1-o(1)}$, then it is not possible to get such a result even for $m = 2$ (and $k = 1$). This is true since distinguishing between the case that a pair of distributions are β -close and the case that they are β' -far for constant β and β' requires $n^{1-o(1)}$ samples [Val08b]. We also note that for $\beta = 0$ the dependence on n must be at least $\Omega(n^{2/3})$ (for $m = 2$ and $k = 1$) [Val08b]. Our algorithm works for $\beta = 0$ and slightly more generally, for $\beta = O(\epsilon/\sqrt{n})$, has no dependence on m , has almost linear dependence on k , and its dependence on n grows like $\tilde{O}(n^{2/3})$.

Theorem 15 *Algorithm Test-Clusterability (see Figure 3.1) is a testing algorithm for (k, β) -clusterability of a list of distributions in the query model, which works for every $\epsilon > 8\beta n^{1/2}$, and performs $\tilde{O}(n^{2/3} \cdot k \cdot \text{poly}(1/\epsilon))$ sampling queries.*

We build on the following theorem.

Theorem 16 ([BFR⁺]) Given parameter δ , and sampling access to distributions \mathbf{p}, \mathbf{q} over $[n]$, there is a test, ℓ_1 -Distance-Test(p, q, ϵ, δ), which takes $O(\epsilon^{-4}n^{2/3} \log n \log \delta^{-1})$ samples from each distribution and for which the following holds.

- If $\|\mathbf{p} - \mathbf{q}\|_1 \leq \epsilon/(4n^{1/2})$, then the test accepts with probability at least $1 - \delta$.
- If $\|\mathbf{p} - \mathbf{q}\|_1 > \epsilon$, then the test rejects with probability at least $1 - \delta$.

Our algorithm is an adaptation of the diameter-clustering tester of [ADPR03], which applies to clustering vectors in \mathbb{R}^d , and is given in Figure 3.1. While often clustering algorithms rely on a method of evaluating distances between the objects that they cluster, the algorithm from [BFR⁺10] only distinguishes pairs of distributions that are very close from those that are ϵ -far (in ℓ_1 distance). Still, this is enough information in conjunction with the algorithm of [ADPR03] to construct a good distribution (k, b) -clusterability tester. In addition, by applying a small change, the algorithm can find an approximately good clustering, as described in the proof of Theorem 15.

<p>Algorithm 5: Test-Clusterability</p> <hr/> <p>Input: Parameters k, β and ϵ, and access in the query model to a list \mathcal{D} of m distributions over $[n]$</p> <ol style="list-style-type: none"> 1 Pick rep_1 uniformly from \mathcal{D}; 2 $i := 1$; 3 $\text{find_new_rep} := \text{true}$; 4 while ($i < k + 1$) and ($\text{find_new_rep} = \text{true}$) do <li style="padding-left: 20px;">5 Uniformly and independently select $2 \ln(6(k + 1))/\epsilon$ distributions from \mathcal{D}; <li style="padding-left: 20px;">6 foreach <i>selected distribution</i> D do <li style="padding-left: 40px;">7 $\text{find_new_rep} := \text{true}$; <li style="padding-left: 40px;">8 for $\ell := 1$ to i do <li style="padding-left: 60px;">9 if ℓ_1-Distance-Test ($D, \text{rep}_\ell, \epsilon/2, \epsilon/12(k + 1) \ln(6(k + 1))$) then <li style="padding-left: 80px;">10 $\text{find_new_rep} := \text{false}$; <li style="padding-left: 40px;">11 if $\text{find_new_rep} = \text{true}$ then <li style="padding-left: 60px;">12 $i := i + 1$; <li style="padding-left: 60px;">13 $\text{rep}_i = D$; <li style="padding-left: 60px;">14 break; 15 if $i \leq k$ then output ACCEPT; 16 else output REJECT;

Figure 3.1: The algorithm for testing clusterability

Proof of Theorem 15: Assume all applications of ℓ_1 -Distance-Test returned a correct answer, as defined by Theorem 16. By the union bound, this happens with probability at least $5/6$. Let us refer to this event as E_1 . Conditioned on E_1 , the clustering algorithm rejects only if it finds $k + 1$ distributions in \mathcal{D} such that the ℓ_1 distance between every two of them is greater than $\frac{\epsilon/2}{4n^{1/2}} \geq \beta$. Thus, if \mathcal{D} is (k, β) -clusterable, then it will be accepted with probability at least $5/6$.

We thus turn to the case that \mathcal{D} is ϵ -far from being (k, β) -clusterable. In this case we claim that as long as there are $t \leq k$ representatives, $\text{rep}_1, \dots, \text{rep}_t$, the number of distributions $D_j \in \mathcal{D}$ such that $\|D_j - \text{rep}_t\|_1 > \epsilon/2$ is at least $\epsilon m/2$. To verify this, assuming in contradiction that there are less than $\epsilon m/2$ such distributions. But then, by modifying each of these distributions so that it equals rep_1 , and modifying each of the other distributions so that it equals the representative it is most close to, we get a list that is $(k, 0)$ -clusterable (at a total cost of less than ϵm).

Since in each iteration of the while loop, there are less than $k + 1$ representative distributions, at least $\frac{\epsilon m}{2}$ of the distributions in \mathcal{D} are $\frac{\epsilon}{2}$ -far from any of the former representative distributions. Therefore, conditioned on E_1 , for every iteration of the while loop, the probability that a new representative is not found is less than $(1 - \epsilon/2)^{\frac{2 \ln(6(k+1))}{\epsilon}} < e^{\ln(6(k+1))} = \frac{1}{6(k+1)}$. By applying the union bound, the algorithm rejects \mathcal{D} with probability greater than $2/3$. Since there are $O(\log k/\epsilon)$ iterations, and in each there is a single application of the ℓ_1 -distance test, by Theorem 16 the total number of samples used is as stated. We note that if we change the algorithm to continue finding new representatives even after finding $k + 1$ representatives then the algorithm would find a set of representatives, S , such that at most ϵm of the distributions in \mathcal{D} are ϵ -far from any representatives in S . ■

Bibliography

- [AAK⁺07] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k -wise and almost k -wise independence. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 496–505, 2007.
- [ACS10] M. Adamaszek, A. Czumaj, and C. Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 56–65, 2010.
- [ADPR03] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. *SIAM Journal on Discrete Math*, 16(3):393–417, 2003.
- [AIOR09] A. Andoni, P. Indyk, K. Onak, and R. Rubinfeld. External sampling. In *Automata, Languages and Programming: Thirty-Sixth International Colloquium (ICALP)*, pages 83–94, 2009.
- [AMS99] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58, 1999.
- [AS92] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley, New York, 1992.
- [Bat01] T. Batu. *Testing properties of distributions*. PhD thesis, Computer Science department, Cornell University, 2001.
- [BDKR05] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.
- [BFF⁺01] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, 2001.
- [BFR⁺] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that distributions are close. This is a long version of [BFR⁺10], and is currently available from the authors’ web-pages.

- [BFR⁺10] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *CoRR*, abs/1009.5397, 2010.
- [BFRV10] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. Technical Report TR10-027, Electronic Colloquium on Computational Complexity (ECCC), 2010.
- [BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 381–390, 2004.
- [BNNR09] K. Do Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld. Sublinear time algorithms for earth mover’s distance. In *CoRR abs/0904.0292*, 2009.
- [BO08] V. Braverman and R. Ostrovsky. Measuring k -wise independence of streaming data. In *CoRR abs/0806.4790*, 2008.
- [BO10a] V. Braverman and R. Ostrovsky. Measuring independence of datasets. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing (STOC)*, pages 271–280, 2010.
- [BO10b] V. Braverman and R. Ostrovsky. Zero-one frequency laws. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing (STOC)*, pages 281–290, 2010.
- [BYJK⁺02] Z. Bar-Yossef, T.S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Proceedings of RANDOM*, 2002.
- [CK04] D. Coppersmith and R. Kumar. An improved data stream algorithm for frequency moments. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 151–156, 2004.
- [CMIM03] G. Cormode, M.Datar, P. Indyk, and S. Muthukrishnan. Comparing data stream using hamming norms (how to zero in). *IEEE Trans. Knowl. Data Eng.*, 15(3):529–540, 2003.
- [CS07] A. Czumaj and C. Sohler. Testing expansion in bounded-degree graphs. In *Proceedings of the Forty-Eighth Annual Symposium on Foundations of Computer Science (FOCS)*, pages 570–578, 2007.
- [Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.
- [Fel67] W. Feller. *An introduction to probability theory and its applications / William Feller*. Wiley, New York ; Sydney :, 3rd ed. edition, 1967.

- [FKSV99] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1 -difference algorithm for massive data streams (extended abstract). In *Proceedings of the Fortieth Annual Symposium on Foundations of Computer Science (FOCS)*, 1999.
- [FM08] E. Fischer and A. Matsliah. Testing graph isomorphism. *SIAM Journal on Computing*, 38(1):207–225, 2008.
- [FS00] J. Fong and M. Strauss. An approximate L^p -difference algorithm for massive data streams. In *Annual Symposium on Theoretical Aspects of Computer Science*, 2000.
- [GMV09] S. Guha, A. McGregor, and S. Venkatasubramanian. Sub-linear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5, 2009.
- [GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity*, 7(20), 2000.
- [Har75] Bernard Harris. The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis János Bolyai*, 16:323–355, 1975. Topics in Information Theory.
- [IKOS09] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai. Extracting correlations. In *Proceedings of the Fiftieth Annual Symposium on Foundations of Computer Science (FOCS)*, pages 261–270, 2009.
- [IM08] P. Indyk and A. McGregor. Declaring independence via the sketching of sketches. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 737–745, 2008.
- [Knu69] D. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison Wesley, Phillipines, 1969.
- [KS08] S. Kale and C. Seshadhri. Testing expansion in bounded degree graphs. In *Automata, Languages and Programming: Thirty-Fifth International Colloquium (ICALP)*, pages 527–538, 2008. A preliminary version appeared in ECCO, TR07-076.
- [Ma81] S.-K. Ma. Calculation of entropy from data of motion. *J. of Statistical Physics*, 26(2):221–240, 1981.
- [MV70] Dragoslav S. Mitrinovic and P. M. Vasic. *Analytic inequalities / D.S. Mitrinovi ; in cooperation with P.M. Vasic*. Springer-Verlag, Berlin ; New York :, 1970.
- [NBS04] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E*, 69(056111), 2004.

- [NS07] A. Nachmias and A. Shapira. Testing the expansion of a graph. Technical Report TR07-118, Electronic Colloquium on Computational Complexity (ECCC), 2007.
- [Pan03] L. Paninski. Estimation of information-theoretic quantities and discrete distributions. *Neural Computation*, 15:1191–1254, 2003.
- [Pan04] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.
- [Pan08] L. Paninski. Testing for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.
- [Roo99] B. Roos. On the rate of multivariate poisson convergence. *J. Multivar. Anal.*, 69(1):120–134, 1999.
- [RRRS07] S. Raskhodnikova, D. Ron, R. Rubinfeld, and A. Smith. Sublinear algorithms for approximating string compressibility. In *Proceedings of the Eleventh International Workshop on Randomization and Computation (RANDOM)*, pages 609–623, 2007.
- [RRSS09] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.
- [RS04] R. Rubinfeld and R. Servedio. Testing monotone high-dimensional distributions. Manuscript, 2004.
- [RX10] R. Rubinfeld and N. Xie. Testing non-uniform k -wise independent distributions over product spaces. In *Automata, Languages and Programming: Thirty-Seventh International Colloquium (ICALP)*, pages 565–581, 2010.
- [SKSB98] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197–200, 1998.
- [Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, 2001.
- [Val08a] P. Valiant. Testing symmetric properties of distributions. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 383–392, 2008.
- [Val08b] P. Valiant. *Testing symmetric properties of distributions*. PhD thesis, CSAIL, MIT, 2008.
- [Whi] Patrick White. Testing random variables for independence and identity. Unpublished manuscript.

- [WW95] D. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. Part I. Bayes estimators and the Shannon entropy. *Physical Review E*, 52(6):6841–6854, 1995.
- [Yam95] Kenji Yamanishi. Probably almost discriminative learning. *Machine Learning*, 18(1):23–50, 1995.