THEORY OF COMPUTING, Volume 9 (8), 2013, pp. 295–347
www.theoryofcomputing.org

Testing Properties of Collections of Distributions*

Reut Levi† Dana Ron‡ Ronitt Rubinfeld§

Received June 14, 2011; Revised August 28, 2012; Published March 10, 2013

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.

ACM Classification: G.3, F.2

AMS Classification: 68Q25

Key words and phrases: property testing, distributions, independence

*A preliminary version of this paper appeared in the the proceedings of the Second Symposium on Innovations in Computer Science (ICS) 2011 [32].

†Supported by the Israel Science Foundation grant numbers 1147/09 and 246/08

‡Supported by the Israel Science Foundation grant number 246/08.

§Supported by NSF grants 0732334 and 0728645, Marie Curie Reintegration grant PIRG03-GA-2008-231077 and the Israel Science Foundation grant nos. 1147/09 and 1675/09.

 DOI: 10.4086/toc.2013.v009a008

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 [25, 39] or similar to another known distribution [12], and testing whether a joint distribution is independent [10]. Related tasks concern sublinear estimation of various measures of a distribution, such as its entropy [9, 26] or its support size [41]. Recently, general techniques have been designed to obtain nearly tight lower bounds on such testing and estimation problems [48, 47].

These types of questions have arisen in several disparate areas, including physics [33, 45, 36], cryptography and pseudorandom number generation [31], statistics [19, 27, 50, 38, 39, 37], learning theory [51], property testing of graphs and sequences (e. g., [25, 20, 30, 35, 40, 23]) and streaming algorithms (e. g., [4, 21, 24, 26, 18, 17, 7, 28, 15, 16, 14, 29]). 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 [11, 48]. 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 Our contributions

1.1.1 The models

We begin by proposing two models that describe possible access patterns to multiple distributions $D_1,\ldots,D_m$ 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 independent 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 is selected uniformly in $[m]$ and $j \in [n]$ is distributed according to D_i. In the second model, referred to as the *query model*, at each time step, the algorithm is allowed to specify $i \in [m]$ and receives j that is distributed according to D_i. 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. The ℓ_1-distance between two probability distributions D_1, D_2 over domain $[n]$,

$$\|D_1, D_2\|_1 = \sum_{i\in[n]} |D_1(i) - D_2(i)|\,,$$

is perhaps the most standard measure of distance between distributions, as it measures the maximum difference between the probability of *any event* (i. e., set $S \subseteq [n]$) occurring according to one distribution as compared to the other distribution. In other words, if the distance is small, then the distributions are essentially indistinguishable in terms of their behavior. 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 also consider an extension to the sampling model in which the distribution over distributions (that is, the distribution of the index i) is non-uniform (i. e., is determined by a weight w_i) and the distance measure is adapted accordingly (see in the preliminaries section for details).

1.1.2 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 [12], an algorithm is given that uses $\tilde{O}(n^{2/3})$ samples[1] and distinguishes between the case that the two distributions are ε-far and the case that they are $O(\varepsilon/\sqrt{n})$-close. This algorithm has been shown to be nearly tight (in terms of the dependence on n) by Valiant [47]. 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 $D_1, \ldots, D_m$, in which the goal of the tester is to distinguish the case in which all distributions are the same (or, slightly more generally, that there is a distribution D^* for which $(1/m)\sum_{i=1}^m \|D_i - D^*\|_1 \leq \mathrm{poly}(\varepsilon)/\sqrt{n}$), from the case in which there is no distribution D^* for which $(1/m)\sum_{i=1}^m \|D_i - D^*\|_1 \leq \varepsilon$. 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 [12] 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_i (which generated j) 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

[1] For a function f over $t \geq 1$ variables $n_1, \ldots, n_t$ we use $\tilde{O}(f(n_1, \ldots, n_t))$ as a shorthand for $O(f(n_1, \ldots, n_t)\log^k f(n_1, \ldots, n_t))$ for some constant k.

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 $(1 \pm \varepsilon/c)$ 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 $[m] \times [n]$ which was studied in [10]. The $\tilde{O}(n^{2/3}m^{1/3} + m)$ algorithm follows from [10] after we fill in a certain gap in the analysis of their algorithm due to an imprecision of a claim given in [11]. The $\tilde{O}(n^{1/2}m^{1/2} + n)$ algorithm exploits the fact that i is selected uniformly (or, more generally, according to a known weight w_i) 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. Indeed, we are able to 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 [12], 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 [47] 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 [10] for testing independence). As in [47], 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.

1.1.3 Testing clusterability in the query model

The second property that we consider is a natural generalization of the equivalence property. The question we ask is whether the distributions can be partitioned into at most k subsets (clusters), such that within every 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 algorithms works by combining the diameter clustering algorithm of [3] (for points in a general metric space where the algorithm has access to the corresponding distance matrix) with the tester for equivalence of a pair of distributions of [12]. We observe that the above-mentioned lower bound of $\Omega(n^{2/3})$ in [47] for testing equivalence of a pair of distributions implies that sample complexity of $\tilde{O}(n^{2/3})$ for the clustering problem is tight to within polylogarithmic factors in n, for constant k. This is true because one can reduce testing equivalence of a pair of distributions to testing clusterability for a constant number of clusters.[2]

1.1.4 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 [47]). 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 [10]: Let Q be a distribution over pairs of elements drawn from $[m]\times[n]$ (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 [10], contains a known gap in the proof. Similar gaps in the lower bounds of [12] for testing the equivalence of a pair of distributions and of [9] for estimating the entropy of a distribution were settled by the work of [47], which applies to symmetric properties. Since independence is not a symmetric property, the work of [47] 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 [10], we get nearly tight bounds on the complexity of testing independence.

1.2 Other related work

Other works on testing and estimating properties of (single or pairs of) distributions include [8, 26, 13, 43, 2, 44, 6, 1, 5].

1.3 Open problems and further research

There are several possible directions for further research on testing properties of collections of distributions, and we next 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 $\Theta(1/\varepsilon)$ distributions uniformly and estimating the mean of each. On the other hand, in the sampling model an $\Omega(\sqrt{m})$ lower

[2]Consider the following reduction. Let D_1^* and D_2^* be a pair of distributions over $[n]$ that we wish to test. We wish to accept if $D_1^* = D_2^*$ and to reject if $\|D_1^* - D_2^*\|_1 > \varepsilon$. Let $\mathcal{D}$ be a collection of m distributions over domain $[2n]$ such that an ε-fraction of the distributions are identical to D_1^* and an ε-fraction are identical to D_2^*. We select the remaining $t = (1-2\varepsilon)m$ distribution, denoted by $D_1,\ldots,D_t$, so that they have the following properties. First, the support of each is (a subset of) $[n+1,\ldots,2n]$ (so that it is disjoint from the support of D_1^* and D_2^*). Second, the collection $\{D_1,\ldots D_t\}$ is $(k-1,0)$-clusterable (see Definition 7.1) and $\varepsilon/(1-2\varepsilon)$-far from being $(k-2,0)$-clusterable. The reduction follows from the fact that $\mathcal{D}$ is $(k,0)$-clusterable if and only if $D_1^* = D_2^*$ and ε-far from being $(k,0)$-clusterable if and only if $\|D_1^* = D_2^*\|_1 > \varepsilon$.

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.

1.4 Organization

We start by providing notation and definitions in Section 2. In Section 3 we give the lower bound of $\Omega(n^{2/3}m^{1/3})$ for testing equivalence in the uniform sampling model, which is the main technical contribution of this paper. In Section 4 we give our second lower bound (of $\Omega(n^{1/2}m^{1/2})$) for testing equivalence and our algorithms for the problem follow in Sections 5 and 6. We conclude with our algorithm for testing clusterability in the query model in Section 7.

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 $D_i : [n] \to [0,1]$ and $\sum_{j=1}^{n} D_i(j) = 1$ for every $1 \le i \le m$. 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 $0 \le \varepsilon \le 1$, we say that $\mathcal{D}$ is *ε-far from* (*having*) $\mathcal{P}$ if $(1/m)\sum_{i=1}^{m} \|D_i - D_i^*\|_1 > \varepsilon$ for every list $\mathcal{D}^* = (D_1^*,\dots,D_m^*)$ that has the property $\mathcal{P}$. Note that the statistical distance (also known as the total variation distance) between two discrete distributions, D and D^* over $[n]$, that is, $\max_{S\subseteq[n]} |\sum_{i\in S} D(i) - \sum_{i\in S} D^*(i)|$, equals $\|D - D^*\|_1/2$.

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.

The Query Model In this model the testing algorithm may indicate an index $1 \le i \le m$ of its choice and it gets an independent sample j distributed according to D_i.

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 i is selected uniformly (we refer to this as the *Uniform* sampling model) and j is an independent sample distributed according to D_i.

We also consider a generalization in which there is an underlying weight vector $\mathbf{w} = (w_1,\dots,w_m)$ (where $\sum_{i=1}^{m} w_i = 1$), and the distribution D_i is selected according to $\mathbf{w}$. In this case the notion of "ε-far" needs to be modified accordingly. We shall say that $\mathcal{D}$ is ε-far from $\mathcal{P}$ *with respect to* $\mathbf{w}$ if $\sum_{i=1}^{m} w_i \cdot \|D_i - D_i^*\|_1 > \varepsilon$ 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 $\mathcal{P}^{\text{eq}}_{m,n}$ (or has the property $\mathcal{P}^{\text{eq}}_{m,n}$) if $D_i = D_{i'}$ for all $1 \le i,i' \le m$.

3 A lower bound for testing equivalence in the uniform sampling model

In this section we prove the following theorem:

Theorem 3.1. *Any testing algorithm for the property $\mathcal{P}^{\mathrm{eq}}_{m,n}$ in the uniform sampling model for every $\varepsilon \le 1/20$ 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 3.1 consists of two parts. The first is a general theorem (Theorem 3.5) concerning testing symmetric properties of lists of distributions. This theorem extends a lemma of Valiant [47, 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 3.5 is applied. Our analysis uses a technique called *Poissonization* [46] (which was used in the past in the context of lower bounds for testing and estimating properties of distributions in [41, 48, 47]), and hence we first introduce some preliminaries concerning Poisson distributions. We later provide some intuition regarding the benefits of Poissonization.

3.1 Preliminaries concerning Poisson distributions

For a positive real number λ, the Poisson distribution $\mathrm{poi}(\lambda)$ takes the value $x \in \mathbb{N}$ (where $\mathbb{N} = \{0,1,2,\dots\}$) with probability $\mathrm{poi}(x;\lambda) = \mathrm{e}^{-\lambda}\lambda^x/x!$. The expectation and variance of $\mathrm{poi}(\lambda)$ 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 [41, Claim A.2]):

$$\|\mathrm{poi}(\lambda_1) - \mathrm{poi}(\lambda_2)\|_1 \le 2|\lambda_1 - \lambda_2|\,. \tag{3.1}$$

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

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, $\mathrm{poi}(\vec{\lambda}(\vec{a}))$ 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 $\mathrm{poi}(\vec{\lambda}(I))$ 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 [47, Lemma 4.5.3] that

$$\left\|\mathrm{poi}(\vec{\lambda}^+) - \mathrm{poi}(\vec{\lambda}^-)\right\|_1 \le \sum_{j=1}^{d} \left\|\mathrm{poi}(\lambda_j^+) - \mathrm{poi}(\lambda_j^-)\right\|_1\,. \tag{3.2}$$

For our purposes we generalize equation (3.2) in the following lemma.

Lemma 3.2. *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, and for any partition $\{I_i\}_{i=1}^{\ell}$ of $[d]$,*

$$\left\|\mathrm{poi}(\vec{\lambda}^+) - \mathrm{poi}(\vec{\lambda}^-)\right\|_1 \le \sum_{i=1}^{\ell} \left\|\mathrm{poi}(\vec{\lambda}^+(I_i)) - \mathrm{poi}(\vec{\lambda}^-(I_i))\right\|_1\,. \tag{3.3}$$

Proof. Let $\{I_i\}_{i=1}^{\ell}$ be a partition of $[d]$, let $\vec{i}$ denote $(i_1,\dots i_d)$, by the triangle inequality we have that for every $k\in[\ell]$,

$$\left|\mathrm{poi}(\vec{i};\vec{\lambda}^+)-\mathrm{poi}(\vec{i};\vec{\lambda}^-)\right| = \left|\prod_{j\in[d]}\mathrm{poi}(i_j;\lambda_j^+)-\prod_{j\in[d]}\mathrm{poi}(i_j;\lambda_j^-)\right| \tag{3.4}$$

$$\leq \left|\prod_{j\in[d]}\mathrm{poi}(i_j;\lambda_j^+)-\prod_{j\in[d]\setminus I_k}\mathrm{poi}(i_j;\lambda_j^+)\prod_{j\in I_k}\mathrm{poi}(i_j;\lambda_j^-)\right| \tag{3.5}$$

$$+\left|\prod_{j\in[d]\setminus I_k}\mathrm{poi}(i_j;\lambda_j^+)\prod_{j\in I_k}\mathrm{poi}(i_j;\lambda_j^-)-\prod_{j\in[d]}\mathrm{poi}(i_j;\lambda_j^-)\right|. \tag{3.6}$$

Hence, we obtain that

$$\left\|\mathrm{poi}(\vec{\lambda}^+)-\mathrm{poi}(\vec{\lambda}^-)\right\|_1 = \sum_{\vec{i}\in\mathbb{N}^d}\left|\mathrm{poi}(\vec{i};\vec{\lambda}^+)-\mathrm{poi}(\vec{i};\vec{\lambda}^-)\right| \tag{3.7}$$

$$\leq \left\|\mathrm{poi}(\vec{\lambda}^+(I_k))-\mathrm{poi}(\vec{\lambda}^-(I_k))\right\|_1 \tag{3.8}$$

$$+\left\|\mathrm{poi}(\vec{\lambda}^+([d]\setminus I_k))-\mathrm{poi}(\vec{\lambda}^-([d]\setminus I_k))\right\|_1 . \tag{3.9}$$

Thus, the lemma follows by induction on ℓ. □

We shall also make use of the following lemma.

Lemma 3.3. *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,*

$$\left\|\mathrm{poi}(\vec{\lambda}^+)-\mathrm{poi}(\vec{\lambda}^-)\right\|_1 \leq 2\sqrt{2\sum_{j=1}^{d}\frac{(\lambda_j^- -\lambda_j^+)^2}{\lambda_j^-}}\,. \tag{3.10}$$

Proof. In order to prove the lemma we shall use the *KL-divergence* between distributions. For two distributions p_1 and p_2 over a domain X, we set

$$D_{\mathrm{KL}}(p_1\|\,p_2)\stackrel{\mathrm{def}}{=}\sum_{x\in X}p_1(x)\cdot\ln\frac{p_1(x)}{p_2(x)}\,.$$

Let $\vec{\lambda}^+=(\lambda_1^+\dots,\lambda_d^+)$, $\vec{\lambda}^-=(\lambda_1^-\dots,\lambda_d^-)$ and let $\vec{i}$ denote $(i_1,\dots i_d)$. We have that

$$\ln\frac{\mathrm{poi}(\vec{i};\vec{\lambda}^+)}{\mathrm{poi}(\vec{i};\vec{\lambda}^-)} = \sum_{j=1}^{d}\ln\left(e^{\lambda_j^- -\lambda_j^+}\left(\lambda_j^+/\lambda_j^-\right)^{i_j}\right) \tag{3.11}$$

$$= \sum_{j=1}^{d}\left((\lambda_j^- -\lambda_j^+)+i_j\cdot\ln(\lambda_j^+/\lambda_j^-)\right) \tag{3.12}$$

$$\leq \sum_{j=1}^{d}\left((\lambda_j^- -\lambda_j^+)+i_j\cdot(\lambda_j^+/\lambda_j^- -1)\right), \tag{3.13}$$

where in the last inequality we used the fact that $\ln x \le x-1$ for every $x>0$. Therefore, we obtain that

$$D_{\mathrm{KL}}\left(\mathrm{poi}(\vec{\lambda}^{+}) \| \mathrm{poi}(\vec{\lambda}^{-})\right) = \sum_{\vec{i}\in\mathbb{N}^d} \mathrm{poi}(\vec{i};\vec{\lambda}^{+}) \cdot \ln \frac{\mathrm{poi}(\vec{i};\vec{\lambda}^{+})}{\mathrm{poi}(\vec{i};\vec{\lambda}^{-})} \tag{3.14}$$

$$\le \sum_{j=1}^{d} \left((\lambda_j^{-} - \lambda_j^{+}) + \lambda_j^{+} \cdot (\lambda_j^{+}/\lambda_j^{-} - 1) \right) \tag{3.15}$$

$$= \sum_{j=1}^{d} \frac{(\lambda_j^{-} - \lambda_j^{+})^2}{\lambda_j^{-}}\,, \tag{3.16}$$

where in equation (3.15) we used the facts that $\sum_{i\in\mathbb{N}} \mathrm{poi}(i;\lambda) = 1$ and $\sum_{i\in\mathbb{N}} \mathrm{poi}(i;\lambda)\cdot i = \lambda$. The ℓ_1 distance is related to the KL-divergence by

$$\|D - D'\|_1 \le 2\sqrt{2 D_{\mathrm{KL}}(D \| D')}$$

and thus we obtain the lemma. □

The next lemma bounds the probability that a Poisson random variable is significantly smaller than its expected value.

Lemma 3.4. *Let* $X \sim \mathrm{poi}(\lambda)$*, then,*

$$\Pr[X < \lambda/2] < (3/4)^{\lambda/4}\,. \tag{3.17}$$

Proof. Consider the matching between j and $j+\lambda/2$ for every $j = 0,\ldots,\lambda/2-1$. We consider the ratio between $\mathrm{poi}(j;\lambda)$ and $\mathrm{poi}(j+\lambda/2;\lambda)$:

$$\frac{\mathrm{poi}(j+\lambda/2;\lambda)}{\mathrm{poi}(j;\lambda)} = \frac{\mathrm{e}^{-\lambda}\cdot\lambda^{j+\lambda/2}/(j+\lambda/2)!}{\mathrm{e}^{-\lambda}\cdot\lambda^{j}/j!} \tag{3.18}$$

$$= \frac{\lambda^{\lambda/2}}{(j+\lambda/2)(j+\lambda/2-1)\cdots(j+1)} \tag{3.19}$$

$$= \frac{\lambda}{j+\lambda/2}\cdot\frac{\lambda}{j+\lambda/2-1}\cdots\frac{\lambda}{j+1} \tag{3.20}$$

$$\ge \frac{\lambda}{\lambda-1}\cdot\frac{\lambda}{\lambda-2}\cdots\frac{\lambda}{\lambda/2} \tag{3.21}$$

$$> \left(\frac{\lambda}{(3/4)\lambda}\right)^{\lambda/4} \tag{3.22}$$

$$= (4/3)^{\lambda/4}\,. \tag{3.23}$$

This implies that

$$\Pr[X < \lambda/2] = \frac{\Pr[X<\lambda/2]}{\Pr[\lambda/2 \le X < \lambda]}\cdot\Pr[\lambda/2 \le X < \lambda] \tag{3.24}$$

$$< \frac{\Pr[X<\lambda/2]}{\Pr[\lambda/2 \le X < \lambda]} \tag{3.25}$$

$$< (3/4)^{\lambda/4}\,, \tag{3.26}$$

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

$$p^{\mathcal{D},\kappa}(j;\vec{a}) \stackrel{\text{def}}{=} \prod_{i=1}^{m} \mathrm{poi}(a_i;\kappa \cdot D_i(j)) . \tag{3.27}$$

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

$$\vec{\lambda}^{\mathcal{D},\kappa}(\vec{a}) = \sum_{j=1}^{n} p^{\mathcal{D},\kappa}(j;\vec{a}) . \tag{3.28}$$

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 $j \in [n]$. In the next subsection we show that the quantity $\vec{\lambda}^{\mathcal{D},\kappa}(\vec{a})$ characterizes the view of the tester, provided that it tests a symmetric property, and thus a lower bound can be proved via a construction of two instances of collections that agree on this quantity while one instance has the property and the other is far from having the property.

3.2 Testability of symmetric properties of lists of distributions

In this subsection we prove the following theorem (which is used to prove Theorem 3.1).

Theorem 3.5. *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 $0 < \delta \le 1/2$. If*

$$\left\| \mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^+,\kappa}\right) - \mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^-,\kappa}\right) \right\|_1 < \frac{31}{60} - \frac{352\delta}{5} , \tag{3.29}$$

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.

A high-level discussion of the proof of Theorem 3.5 For an element $j \in [n]$ and a distribution D_i, $i \in [m]$, let $\alpha_{i,j}$ be the number of times the pair (i, j) appears in the sample (when the sample is selected according to some sampling model). Thus $(\alpha_{1,j}, \dots, \alpha_{m,j})$ is the *sample histogram* of the element j. 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 j such that $(\alpha_{1,j}, \dots, \alpha_{m,j}) = \vec{a}$. As shown in [12], 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.

For example, consider the distributions D_1 and D_2 over $\{1,2,3\}$ such that $D_1[j] = 1/3$ for every $j \in \{1,2,3\}$, $D_2[1] = D_2[2] = 1/2$ and $D_2[3] = 0$. Assume that we sample (D_1, D_2) four times, according to the uniform sampling model and we get the samples $(1,1),(2,1),(2,2),(1,3)$, where the first coordinate denotes the distribution, and the second coordinate denotes the element. Then the sample histogram of element 1 is $(1,1)$ because 1 was selected once by D_1 and once by D_2. For the elements $2,3$ we have the sample histograms $(0,1)$ and $(1,0)$, respectively. The fingerprint of the sample is $(0,1,1,0,1,0,0,\dots)$ for the following order of histograms: $((0,0),(0,1),(1,0),(2,0)(1,1),(0,2),(3,0),\dots)$.

In order to prove Theorem 3.5, 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. For each choice of element $j \in [n]$ and a distribution D_i, the number of times the sample (i,j) appears, i. e., $\alpha_{i,j}$, depends on the number of times the other samples appear simply because the total number of samples is fixed. Furthermore, for each histogram $\vec{a}$, the number of elements with sample histogram identical to $\vec{a}$ is dependent on the number of times the other histograms appear, because the number of samples is fixed. For instance, in the example above, if we know that we have the histogram $(0,1)$ once and the histogram $(1,1)$ once, then we know that third histogram cannot be $(2,0)$. In addition, it is dependent because the number of elements is fixed.

We thus see that the distribution of the fingerprints is rather difficult to analyze (and therefore it is difficult to bound the statistical distance between two different such distributions). Therefore, we would like to break as much of the above dependencies. To this end we define a slightly different process for generating the samples that involves *Poissonization* [46]. In the Poissonized process the number of samples we take from each distribution D_i, denoted by κ'_i, is distributed according to the Poisson distribution. We prove that, while the overall number of samples the Poissonized process takes is bigger just by a constant factor from the uniform process, we get with very high probability that $\kappa'_i > \kappa_i$, for every i, where κ_i is the number of samples taken from D_i. This implies that if we prove a lower bound for algorithms that receive samples generated by the Poissonized process, then we obtain a related lower bound for algorithms that work in the uniform sampling model.

As opposed to the process that takes a fixed number of samples according to the uniform sampling model, the benefit of the Poissonized process is that the $\alpha_{i,j}$'s determined by this process are independent. Therefore, the type of sample histogram that element j has is completely independent of the types of sample histograms the other elements have. We get that the fingerprint distribution is a generalized multinomial distribution, which has been studied by Roos [42] (the connection is due to Valiant [48]).

Definition 3.6. 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 m independent samples $\kappa_1,\dots,\kappa_m$ from $\mathrm{poi}(\kappa)$,
- Return κ_i independent samples distributed according to D_i for each $i \in [m]$.

Lemma 3.7. *Assume 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 $\kappa \geq c$ for some sufficiently large constant c, and works for every $\varepsilon \geq \varepsilon_0$ 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 2κ, that works for every $\varepsilon \geq \varepsilon_0$ and whose success probability is at least $19/30$.*

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 T' be a tester in the Poissonized uniform sampling model with parameter 2κ. The tester T' receives a set of samples, S', where each sample in the set is distributed uniformly and independently over the m distributions, and their number, κ', is distributed as $\kappa' \sim \mathrm{poi}(2\kappa m)$ (see [22, p. 216]). By Lemma 3.4 we have that,

$$\Pr\left[\kappa' < \kappa m\right] \leq (3/4)^{\kappa m/2} . \tag{3.30}$$

If $\kappa' \geq \kappa m$ then T' picks a random subset of samples, $S \subset S'$, of size κm and simulates T on S. Otherwise it "accepts" or "rejects" with equal probability. Since the samples in S are distributed as κm samples drawn according to the uniform sampling model, if equation (3.30) holds then T' simulates a tester in the uniform sampling model successfully.

The probability that $\kappa' \geq \kappa m$ is at least $1-(3/4)^{\kappa m/2}$, which is greater than $4/5$ for $\kappa > c$ 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 3.7 it suffices to consider samples that are generated in the Poissonized uniform sampling model. The process for generating a sample $\{\alpha_{1,j},\dots,\alpha_{m,j}\}_{j\in[n]}$ (recall that $\alpha_{i,j}$ is the number of times that element j was selected by distribution D_i) in the κ-Poissonized model is equivalent to the following process: For each $i \in [m]$ and $j \in [n]$, independently select $\alpha_{i,j}$ according to $\mathrm{poi}(\kappa \cdot D_i(j))$ (see [22, p. 216]). Thus the probability of getting a particular histogram $\vec{a} = (a_1,\dots,a_m)$ for element j is $p^{\mathcal{D},\kappa}(j;\vec{a})$ (as defined in equation (3.27)). We can represent the event that the histogram of element j is $\vec{a}$ by a Bernoulli random vector $\vec{b}_j$ that is indexed by all $\vec{v} \in \mathbb{N}^m$, is 1 in the coordinate corresponding to $\vec{a}$, and is 0 elsewhere. Given this representation, the fingerprint of the sample corresponds to $\sum_{j=1}^n \vec{b}_j$.

As defined above, the dimension of the $\vec{b}_j$'s is unbounded. This is simply because $\mathrm{poi}(\lambda)$ is unbounded for every $\lambda > 0$, and so every possible histogram has non-zero probability. However, we may consider vectors of finite dimension (we show in the proof of Theorem 3.5 that it is sufficient) and map the histograms that we do not consider to the vector $(0,\dots,0)$. Roos's theorem, stated next, shows that the distribution of the fingerprints can be approximated by a multivariate Poisson distribution (the Poisson here is related to the fact that the fingerprints' distributions are generalized multinomial distributions and not related to the Poisson from the Poissonization process). For simplicity, the theorem is stated for vectors $\vec{b}_j$ that are indexed directly, that is $\vec{b}_j = (b_{j,1},\dots,b_{j,h})$.

Theorem 3.8 ([42]). *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 $\Pr\left[\vec{b}_j = \vec{e}_\ell\right] = p_{j,\ell}$ and $\Pr\left[\vec{b}_j = (0,\dots,0)\right] = 1-\sum_{\ell=1}^h p_{j,\ell}$ (here $\vec{e}_\ell$ satisfies $e_{j,\ell}=1$ and $e_{j,\ell'}=0$ for every $\ell' \neq \ell$). Let us define an h-dimensional vector $\vec{\lambda} = (\lambda_1,\dots,\lambda_h)$ by $\lambda_\ell := \sum_{j=1}^n p_{j,\ell}$. Then*

$$\left\| D^{S_n} - \mathrm{poi}(\vec{\lambda}) \right\|_1 \leq \frac{88}{5} \sum_{\ell=1}^{h} \frac{\sum_{j=1}^n p_{j,\ell}^2}{\sum_{j=1}^n p_{j,\ell}} . \tag{3.31}$$

We next show how to obtain a bound on sums of the form given in equation (3.31) under appropriate conditions.

Lemma 3.9. *Given a list* $\mathcal{D} = (D_1, \ldots, D_m)$ *of* m *distributions over* $[n]$ *and a real number* $0 < \delta \leq 1/2$ *such that for all* $i \in [m]$ *and for all* $j \in [n]$, $D_i(j) \leq \delta/(m \cdot \kappa)$ *for some integer* κ*, we have that*

$$\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{j=1}^n p^{\mathcal{D},\kappa}(j;\vec{a})^2}{\sum_{j=1}^n p^{\mathcal{D},\kappa}(j;\vec{a})} \leq 2\delta\,. \tag{3.32}$$

Proof.

$$\begin{aligned}
\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{j=1}^n p^{\mathcal{D},\kappa}(j;\vec{a})^2}{\sum_{j=1}^n p^{\mathcal{D},\kappa}(j;\vec{a})} &\leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_j \left(p^{\mathcal{D},\kappa}(j;\vec{a}) \right) &(3.33)\\
&= \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_j \left(\prod_{i=1}^m \mathrm{poi}(a_i; \kappa \cdot D_i(j)) \right) &(3.34)\\
&\leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \left(\frac{\delta}{m} \right)^{a_1+\ldots+a_m} &(3.35)\\
&\leq \sum_{a=1}^{\infty} m^a \left(\frac{\delta}{m} \right)^a &(3.36)\\
&\leq 2\delta\,, &(3.37)
\end{aligned}$$

where the inequality in equation (3.37) holds for $\delta \leq 1/2$ and the inequality in equation (3.35) follows from:

$$\begin{aligned}
\mathrm{poi}(a; \kappa \cdot D_i(j)) &= \frac{\mathrm{e}^{-\kappa \cdot D_i(j)} (\kappa \cdot D_i(j))^a}{a!} &(3.38)\\
&\leq (\kappa \cdot D_i(j))^a &(3.39)\\
&\leq \left(\frac{\delta}{m} \right)^a, &(3.40)
\end{aligned}$$

and the proof is completed. □

Proof of Theorem 3.5. Let $D^{\mathcal{D}^-}$ and $D^{\mathcal{D}^+}$ denote the distribution of the fingerprints when taking samples in the κ-Poissonized model (see Definition 3.6) from $\mathcal{D}^-$ and $\mathcal{D}^+$, respectively. We first bound the ℓ_1-distance between $D^{\mathcal{D}^-}$ and $D^{\mathcal{D}^+}$ and then prove the desired result on the uniform sampling model by applying Lemma 3.7. Let $A \subset \mathbb{N}^m$ be some finite subset of histograms such that for both $\mathcal{D}^-$ and $\mathcal{D}^+$, the probability (in the κ-Poissonized model) that at least one of the elements has a sample histogram $\vec{a} \notin A$ is smaller than $1/120$ (it is easy to verify, for example by Markov's inequality, that such a subset exists). Denote by $\mathcal{F}$ the subset of fingerprints describing outcomes where for each element in $[n]$, the sample histogram is in A. Hence,

$$\sum_{\mathcal{F} \ni \vec{f}} \left| D^{\mathcal{D}^-}\left(\vec{f}\right) - D^{\mathcal{D}^+}\left(\vec{f}\right) \right| \leq \frac{1}{60}\,.$$

We next turn to bound
$$\sum_{\mathcal{F}\ni\vec{f}}\left|D^{\mathcal{D}^-}\left(\vec{f}\right)-D^{\mathcal{D}^+}\left(\vec{f}\right)\right| .$$

By the first premise of the theorem, $D_i^+(j), D_i^+(j) \le \delta/(\kappa m)-$ for every $i\in[m]$ and $j\in[n]$. By Lemma 3.9 this implies that equation (3.32) holds both for $\mathcal{D}=\mathcal{D}^+$ and for $\mathcal{D}=\mathcal{D}^-$. Combining this with Theorem 3.8 we get that the ℓ_1-distance between the fingerprint distribution over $\mathcal{F}$ when the sample is generated according to $\mathcal{D}^+$ (in the κ-Poissonized model) and the distribution $\mathrm{poi}\big(\vec{\lambda}^{\mathcal{D}^+,\kappa}\big)$ 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 (3.29) (concerning the ℓ_1 distance between $\mathrm{poi}\big(\vec{\lambda}^{\mathcal{D}^+,\kappa}\big)$ and $\mathrm{poi}\big(\vec{\lambda}^{\mathcal{D}^-,\kappa}\big)$) and the triangle inequality, we get that
$$\sum_{\bar{\mathcal{F}}\cup\mathcal{F}\ni\vec{f}}\left|D^{\mathcal{D}^-}\left(\vec{f}\right)-D^{\mathcal{D}^+}\left(\vec{f}\right)\right|\le 2\cdot\frac{176}{5}\delta+\frac{31}{60}-\frac{352\delta}{5}+\frac{1}{60}=\frac{16}{30},$$
which implies that the statistical distance is smaller than $8/30$. Thus a tester that accepts $\mathcal{D}^+$ and rejects $\mathcal{D}^-$ with probability greater than
$$\frac{1}{2}+\frac{1}{2}\cdot\frac{8}{30}=\frac{19}{30},$$
viewing only the above-mentioned fingerprints, does not exist. By Lemma 3.7 the theorem follows. □

3.3 Proof of Theorem 3.1

In this subsection we show how to apply Theorem 3.5 to two lists of distributions, $\mathcal{D}^+$ and $\mathcal{D}^-$, which we will define shortly, where $\mathcal{D}^+\in\mathcal{P}^{\mathrm{eq}}=\mathcal{P}^{\mathrm{eq}}_{m,n}$ while $\mathcal{D}^-$ is $(1/20)$-far from $\mathcal{P}^{\mathrm{eq}}$. Recall that by the premise of Theorem 3.1, $n\ge cm\log m$ for some sufficiently large constant $c>1$. In the proof it will be convenient to assume that m is even and that n (which corresponds in Lemma 3.10 to $2t$) 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 3.10. *For every two even integers m and t, there exists a $0/1$-valued matrix M with m rows and t 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 $2\le x<m/2$, and for every subset $S\subseteq[m]$ of size x, the number of columns j such that $M[i,j]=1$ for every $i\in S$ is at least*
$$t\cdot\left(\frac{1}{2^x}\left(1-\frac{2x^2}{m}\right)-\sqrt{\frac{2x\ln m}{t}}\right),$$
and at most
$$t\cdot\left(\frac{1}{2^x}+\sqrt{\frac{2x\ln m}{t}}\right).$$

Proof. Consider selecting a matrix M randomly as follows. Denote the first $t/2$ columns of M by F. For each column in F, pick, independently from the other $t/2-1$ columns in F, a random half of its elements to be 1, and the other half of the elements to be 0. Columns $t/2+1,\dots,t$ are the negations of columns $1,\dots,t/2$, 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 column j between 1 and t, each subset of rows $S \subseteq [m]$ of size x, and $b \in \{0,1\}$, define the indicator random variable $I_{S,j,b}$ to be 1 if and only if $M[i,j]=b$ for every $i \in S$. Hence,

$$\Pr[I_{S,j,b}=1] = \frac{1}{2}\cdot\left(\frac{1}{2}-\frac{1}{m}\right)\cdot\dots\cdot\left(\frac{1}{2}-\frac{x-1}{m}\right). \tag{3.41}$$

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

$$\begin{aligned}\Pr[I_{S,j,b}=1] &\geq \left(\frac{1}{2}-\frac{x}{m}\right)^x &(3.42)\\ &= \frac{1}{2^x}\left(1-\frac{2x}{m}\right)^x &(3.43)\\ &\geq \frac{1}{2^x}\left(1-\frac{2x^2}{m}\right), &(3.44)\end{aligned}$$

where the last inequality is due to Bernoulli's inequality which states that $(1+x)^n > 1+nx$, for every real, nonzero number $x>-1$ and an integer $n>1$ ([34]).

Let $E_{S,b}$ denote the expected value of $\sum_{j=1}^{t/2} I_{S,j,b}$. From the fact that columns $t/2+1,\dots,t$ are the negations of columns $1,\dots,t/2$ it follows that

$$\sum_{j=t/2+1}^{t} I_{S,j,1} = \sum_{j=1}^{t/2} I_{S,j,0}.$$

Therefore, the expected number of columns $1 \leq j \leq t$ such that $M[i,j]=1$ for every $i \in S$ is simply $E_{S,1}+E_{S,0}$ (that is, at most $t\cdot 1/2^x$ and at least $(t\cdot 1/2^x)\left(1-2x^2/m\right)$). By the additive Chernoff bound,

$$\begin{aligned}\Pr\left[\left|\sum_{j=1}^{t/2} I_{S,j,b}-E_{S,b}\right| > \sqrt{\frac{tx\ln m}{2}}\right] &< 2\exp(-2(t/2)(2x\ln m)/t) &(3.45)\\ &= 2m^{-2x}. &(3.46)\end{aligned}$$

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

$$\Pr\left[\left|\sum_{j=1}^{t} I_{S,j,1}-(E_{S,1}+E_{S,0})\right| > \sqrt{2tx\ln m}\right] < 4m^{-2x}. \tag{3.47}$$

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 $i \in [m]$:

$$D_i^+(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \le j \le \frac{n^{2/3}m^{1/3}}{2}\,, \\ \frac{1}{n} & \text{if } \frac{n}{2} < j \le n\,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.48}$$

We now turn to defining $\mathcal{D}^-$. Let M be a matrix as in Lemma 3.10 for $t = n/2$. For every $i \in [m]$:

$$D_i^-(j) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \le j \le \frac{n^{2/3}m^{1/3}}{2}\,, \\ \frac{2}{n} & \text{if } \frac{n}{2} < j \le n \text{ and } M[i, j-n/2] = 1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.49}$$

For both $\mathcal{D}^+$ and $\mathcal{D}^-$, we refer to the elements $1 \le j \le n^{2/3}m^{1/3}/2$ as the *heavy* elements, and to the elements $n/2 \le j \le n$, as the *light* elements. Observe that each heavy element has exactly the same probability weight, $1/(n^{2/3}m^{1/3})$, in all distributions D_i^+ and D_i^-. On the other hand, for each light element j, while $D_i^+(j) = 1/n$ (for every i), in $\mathcal{D}^-$ we have that $D_i^+(j) = 2/n$ for half of the distributions, the distributions selected by the M, and $D_i^+(j) = 0$ 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 $\mathcal{D}^-$.

A high-level discussion To gain some intuition before delving into the detailed proof, consider first the special case that $m = 2$ (which was studied by Valiant [48], and indeed the construction is the same as the one he analyzes (and was initially proposed in [11])). In this case each heavy element has probability weight $\Theta(1/n^{2/3})$ and we would like to establish a lower bound of $\Omega(n^{2/3})$ on the number of samples required to distinguish between $\mathcal{D}^+$ and $\mathcal{D}^-$. That is, we would like to show that the corresponding fingerprints' distributions when the sample is of size $o(n^{2/3})$ are very similar.

The first main observation is that since the probability weight of light elements is $\Theta(1/n)$ in both $\mathcal{D}^+$ and $\mathcal{D}^-$, the probability that a light element will appear more than twice in a sample of size $o(n^{2/3})$ is very small. That is (using the fingerprints of histograms notation we introduced previously), for each $\vec{a} = (a_1, a_2)$ such that $a_1 + a_2 > 2$, the sample will not include (with high probability) any light element j such that $\alpha_{1,j} = a_1$ and $\alpha_{2,j} = a_2$ (for both $\mathcal{D}^+$ and $\mathcal{D}^-$). Moreover, for every $x \in \{1,2\}$, the expected number of elements j such that $(\alpha_{1,j}, \alpha_{2,j}) = (x, 0)$ is the same in $\mathcal{D}^+$ and $\mathcal{D}^-$, as well as the variance (from symmetry, the same applies to $(0, x)$). Thus, most of the difference between the fingerprints' distributions is due to the numbers of elements j such that $(\alpha_{1,j}, \alpha_{2,j}) = (1,1)$. For this setting we do expect to see a non-negligble difference for light elements between $\mathcal{D}^+$ and $\mathcal{D}^-$ (in particular, we cannot get the $(1,1)$ histogram for light elements in $\mathcal{D}^-$, as opposed to $\mathcal{D}^+$).

Here is where the heavy elements come into play. Recall that in both $\mathcal{D}^+$ and $\mathcal{D}^-$ the heavy elements have the same probability weight, so that the expected number of heavy elements i such that $(a_{1,j}, a_{2,j}) = (1,1)$ is the same for $\mathcal{D}^+$ and $\mathcal{D}^-$. However, intuitively, the variance of these numbers for the heavy elements "swamps" the differences between the light elements so that it is not possible to distinguish between $\mathcal{D}^+$ and $\mathcal{D}^-$. The actual proof, which formalizes (and quantifies) this intuition, considers the difference between the values of the vectors $\vec{\lambda}^{\mathcal{D}^+,k}$ and $\vec{\lambda}^{\mathcal{D}^-,k}$ (as defined in equation (3.28)) in the coordinates corresponding to $\vec{a}$ such that $a_1 + a_2 = 2$. We can then apply Lemmas 3.2 and 3.3 to obtain equation (3.29) in Theorem 3.5.

Turning to $m > 2$, it is no longer true that in a sample of size $o(n^{2/3}m^{1/3})$ we will not get histogram vectors $\vec{a}$ such that $\sum_{i=1}^m a_i > 2$ for light elements. Thus we have to deal with many more vectors $\vec{a}$ (of dimension m) and to bound the total contribution of all of them to the difference between fingerprints of $\mathcal{D}^+$ and of $\mathcal{D}^-$. To this end we partition the set of all possible histograms' vectors into several subsets according to their Hamming weight $\sum_{i=1}^m a_i$ and depending on whether all $a_i's$ are in $\{0,1\}$, or there exists a least one a_i such that $a_i \geq 2$. In particular, to deal with the former (whose number, for each choice of Hamming weight x is relatively large, i. e., roughly m^x), we use the properties of the matrix M based on which $\mathcal{D}^-$ is defined. We note that from the analysis we see that, similarly to when $m = 2$, we need the variance of the heavy elements to play a role just for the cases where $\sum_{i=1}^m a_i = 2$ while in the other cases the total contribution of the light elements is rather small.
In the remainder of this section we provide the details of the analysis.

Before establishing that indeed $\mathcal{D}^-$ is $\Omega(1)$-far from $\mathcal{P}^{\mathrm{eq}}$, we introduce some more notation (which will be used throughout the remainder of the proof of Theorem 3.1). Let S_x be the set of vectors of dimension m 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). Let A_x be the set of vector of dimension m that their coordinates sum up to x but must contain at least one coordinate that is 2 (which corresponds to an element that was samples at least twice by at least one distribution). More formally, for any integer x, we define the following two subsets of $\mathbb{N}^m$:

$$S_x \stackrel{\mathrm{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m \ : \ \begin{array}{c} \sum_{i=1}^m a_i = x \text{ and} \\ \forall i \in [m], a_i < 2 \end{array} \right\}, \tag{3.50}$$

and

$$A_x \stackrel{\mathrm{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m \ : \ \begin{array}{c} \sum_{i=1}^m a_i = x \text{ and} \\ \exists i \in [m], a_i \geq 2 \end{array} \right\}. \tag{3.51}$$

For $\vec{a} \in \mathbb{N}^m$, let $\sup(\vec{a}) \stackrel{\mathrm{def}}{=} \{i : a_i \neq 0\}$ denote the *support* of $\vec{a}$, and let

$$I_M(\vec{a}) \stackrel{\mathrm{def}}{=} \left\{ j \ : \ D_i^-(j) = \frac{2}{n} \quad \forall i \in \sup(\vec{a}) \right\}. \tag{3.52}$$

Note that in terms of the matrix M (based on which $\mathcal{D}^-$ is defined), $I_M(\vec{a})$ consists of the columns 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 3.11. *For every $m > 5$ and for $n \geq c \ln m$ for some sufficiently large c, we have that*

$$\sum_{i=1}^m \|D_i^- - D^*\|_1 > \frac{m}{20}$$

for every distribution D^* over $[n]$. That is, the list $\mathcal{D}^-$ is $(1/20)$-far from $\mathcal{P}^{\mathrm{eq}}$.

Proof. Consider any $\vec{a} \in S_2$. By Lemma 3.10, setting $t = n/2$, the size of $I_M(\vec{a})$, i. e., the number of light elements ℓ such that $D_i^-[\ell] = 2/n$ for every $i \in \sup(\vec{a})$, is at most

$$\frac{n}{2}\left(\frac{1}{4} + \sqrt{\frac{8\ln m}{n}}\right).$$

The same upper bound holds for the number of light elements ℓ such that $D_i^-[\ell] = 0$ for every $i \in \sup(\vec{a})$. This implies that for every $i \neq i'$ 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 $D_i^-[\ell] = 2/n$ while $D_{i'}^-[\ell] = 0$, or that $D_{i'}^-[\ell] = 2/n$ while $D_i^-[\ell] = 0$. Therefore,

$$\|D_i^- - D_{i'}^-\|_1 \geq \frac{1}{2} - 2\sqrt{\frac{8\ln m}{n}},$$

which for $n \geq c\ln m$ and a sufficiently large constant c, is at least $1/8$. Thus, by the triangle inequality we have that for every D^*,

$$\sum_{i=1}^{m} \|D_i^- - D^*\|_1 \geq \left\lfloor \frac{m}{2} \right\rfloor \cdot \frac{1}{8},$$

which greater than $m/20$ for $m > 5$. □

In what follows we work towards establishing that equation (3.29) in Theorem 3.5 holds for $\mathcal{D}^+$ and $\mathcal{D}^-$. Set $\kappa = \delta \cdot n^{2/3}/m^{2/3}$, 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 (3.28)). By the definition of $\vec{\lambda}^+$, for each $\vec{a} \in \mathbb{N}^m$,

$$\begin{aligned}
\vec{\lambda}^+(\vec{a}) &= \sum_{j=1}^{n}\prod_{i=1}^{m} \frac{(\kappa \cdot D_i^+(j))^{a_i}}{\mathrm{e}^{\kappa \cdot D_i^+(j)} \cdot a_i!} && (3.53)\\
&= \sum_{j=1}^{n^{2/3}m^{1/3}/2}\prod_{i=1}^{m} \frac{(\delta/m)^{a_i}}{\mathrm{e}^{\delta/m} \cdot a_i!} + \sum_{j=n/2+1}^{n}\prod_{i=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_i}}{\mathrm{e}^{\delta/(n^{1/3}m^{2/3})} \cdot a_i!} && (3.54)\\
&= \frac{n^{2/3}m^{1/3}}{2\mathrm{e}^{\delta}}\prod_{i=1}^{m} \frac{(\delta/m)^{a_i}}{a_i!} + \frac{n}{2\mathrm{e}^{\delta(m/n)^{1/3}}}\prod_{i=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!}. && (3.55)
\end{aligned}$$

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

$$\vec{\lambda}^-(\vec{a}) = \frac{n^{2/3}m^{1/3}}{2\mathrm{e}^{\delta}}\prod_{i=1}^{m} \frac{(\delta/m)^{a_i}}{a_i!} + \frac{1}{\mathrm{e}^{\delta(m/n)^{1/3}}}\sum_{j\in I_M(\vec{a})}\prod_{i=1}^{m} \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!}. \tag{3.56}$$

Hence, $\vec{\lambda}^+(\vec{a})$ and $\vec{\lambda}^-(\vec{a})$ differ only on the term which corresponds to the contribution of the light elements. Equations (3.55) and (3.56) demonstrate why we choose M with the specific properties defined in Lemma 3.10. First of all, in order for every D_i^- to be a probability distribution, we want each row of M to sum up to exactly $n/4$. We also want each column of M to sum up to exactly $m/2$, in order to get $\prod_{i=1}^m \mathrm{e}^{-\kappa \cdot D_i^+(j)} = \prod_{i=1}^m \mathrm{e}^{-\kappa \cdot D_i^-(j)}$. Finally, we would have liked $|I_M(\vec{a})| \cdot \prod_{i=1}^m 2^{a_i}$ to equal $n/2$ for every $\vec{a}$. This would imply that $\vec{\lambda}^+(\vec{a})$ and $\vec{\lambda}^-(\vec{a})$ are equal. As we show below, this is in fact true for every $\vec{a} \in S_1$. For vectors $\vec{a} \in S_x$ where $x > 1$, the second condition in Lemma 3.10 ensures that $|I_M(\vec{a})|$ is sufficiently close to $(n/2) \cdot (1/2^x)$. This property of M is not necessary in order to bound the contribution of the vectors in A_x. The bound that we give for those vectors is less tight, but since there are fewer such vectors, it suffices.

We start by considering the contribution to equation (3.29) 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. We prove that in the Poissonized uniform sampling model, for every $\vec{a} \in S_1$ the number of elements with such sample histogram is distributed exactly the same in $\mathcal{D}^+$ and $\mathcal{D}^-$.

Lemma 3.12.

$$\sum_{\vec{a} \in S_1} \left\| \mathrm{poi}(\vec{\lambda}^+(\vec{a}) - \mathrm{poi}(\vec{\lambda}^-(\vec{a})) \right\|_1 = 0\,. \tag{3.57}$$

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

$$\sum_{j \in I_M(\vec{a})} \prod_{i=1}^m \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!} = \frac{n}{2} \prod_{i=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!}\,. \tag{3.58}$$

By equations (3.55) and (3.56), 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 (3.1). $\square$

We now turn to bounding the contribution to equation (3.29) 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 3.13.

$$\left\| \mathrm{poi}(\vec{\lambda}^+(A_2)) - \mathrm{poi}(\vec{\lambda}^-(A_2)) \right\|_1 \le 3\delta\,. \tag{3.59}$$

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

$$\sum_{j \in I_M(\vec{a})} \prod_{i=1}^m \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!} = n \prod_{i=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!}\,. \tag{3.60}$$

By equations (3.55), (3.56) and (3.60) it follows that

$$\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}) = \frac{n}{2\mathrm{e}^{\delta(m/n)^{1/3}}} \prod_{i=1}^m \frac{(\delta/(n^{1/3}m^{2/3}))^{a_i}}{a_i!} = \frac{n^{1/3}\delta^2}{4\mathrm{e}^{\delta(m/n)^{1/3}} m^{4/3}}\,, \tag{3.61}$$

and that

$$\vec{\lambda}^-(\vec{a}) \geq \frac{n^{2/3}m^{1/3}}{2\mathrm{e}^\delta}\prod_{i=1}^{m}\frac{(\delta/m)^{a_i}}{a_i!} = \frac{n^{2/3}\delta^2}{4\mathrm{e}^\delta m^{5/3}}\,. \tag{3.62}$$

By equations (3.61) and (3.62) we have that

$$\frac{\left(\vec{\lambda}^-(\vec{a})-\vec{\lambda}^+(\vec{a})\right)^2}{\vec{\lambda}^-(\vec{a})} \leq \frac{\mathrm{e}^{\delta-2\delta(m/n)^{1/3}}\delta^2}{4m} \leq \frac{\delta^2}{m}\,. \tag{3.63}$$

By equation (3.63) and the fact that $|A_2| = m$ we get

$$\sum_{\vec{a}\in A_2}\frac{\left(\vec{\lambda}^-(\vec{a})-\vec{\lambda}^+(\vec{a})\right)^2}{\vec{\lambda}^-(\vec{a})} \leq m\cdot\frac{\delta^2}{m} = \delta^2 \tag{3.64}$$

The lemma follows by applying Lemma 3.3. □

Recall that for a subset I of $\mathbb{N}^m$, $\mathrm{poi}(\vec{\lambda}(I))$ 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 3.14. *For $m \geq 16$, $n \geq cm\ln m$ (where c is a sufficiently large constant) and for $\delta \leq 1/16$*

$$\left\|\mathrm{poi}\left(\vec{\lambda}^+\left(\bigcup_{x=2}^{m/2-1}S_x\right)\right)-\mathrm{poi}\left(\vec{\lambda}^-\left(\bigcup_{x=2}^{m/2-1}S_x\right)\right)\right\|_1 \leq 32\delta\,. \tag{3.65}$$

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 (3.55): For each $\vec{a}\in\bigcup_{x=2}^{m/2-1}S_x$,

$$\vec{\lambda}^+(\vec{a}) = \frac{n^{2/3}m^{1/3}}{2\mathrm{e}^\delta}\cdot\left(\frac{\delta}{m}\right)^x + \frac{n}{2\mathrm{e}^{\delta(m/n)^{1/3}}}\cdot\left(\frac{\delta}{n^{1/3}m^{2/3}}\right)^x\,. \tag{3.66}$$

By Lemma 3.10, for every $\vec{a}\in\bigcup_{x=2}^{m/2-1}S_x$ the size of $I_M(\vec{a})$ is at most

$$\frac{n}{2}\cdot\left(\frac{1}{2^x}+\sqrt{\frac{4x\ln m}{n}}\right)$$

and at least

$$\frac{n}{2}\cdot\left(\frac{1}{2^x}-\frac{2x^2}{2^x m}-\sqrt{\frac{4x\ln m}{n}}\right)\,.$$

By equation (3.56) this implies that

$$\vec{\lambda}^-(\vec{a}) = \frac{n^{2/3}m^{1/3}}{2e^{\delta}} \cdot \left(\frac{\delta}{m}\right)^x + \frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{1}{2^x} + \eta\right)\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x , \tag{3.67}$$

where

$$-\left(\frac{2x^2}{2^x m} + \sqrt{\frac{4x\ln m}{n}}\right) \leq \eta \leq \sqrt{\frac{4x\ln m}{n}}$$

and thus

$$|\eta| \leq \sqrt{\frac{x}{m}} \cdot \left(\frac{2x^2}{2^x\sqrt{m}} + \sqrt{\frac{4m\ln m}{n}}\right) .$$

By the facts that $n \geq cm\ln m$ for some sufficiently large constant c, and that

$$\frac{2x^2}{2^x\sqrt{m}} \leq \frac{1}{2}$$

for every $2 \leq x < m/2$ and $m \geq 16$, we obtain that $|\eta| \leq \sqrt{\frac{x}{m}}$. So we have that

$$(\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}))^2 \leq \left(\frac{n}{2e^{\delta(m/n)^{1/3}}} \cdot \left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x \cdot \sqrt{\frac{x}{m}}\right)^2 \leq \frac{n^2}{4} \cdot \left(\frac{4\delta^2}{n^{2/3}m^{4/3}}\right)^x \cdot \frac{x}{m} . \tag{3.68}$$

and that

$$\vec{\lambda}^-(\vec{a}) \geq \frac{n^{2/3}m^{1/3}}{2e^{\delta}} \cdot \left(\frac{\delta}{m}\right)^x . \tag{3.69}$$

Then we get, for $\delta \leq 1/2$, that

$$\frac{\left(\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a})\right)^2}{\vec{\lambda}^-(\vec{a})} \leq \frac{e^{\delta}n^{4/3}}{2m^{1/3}} \cdot \left(\frac{4\delta}{n^{2/3}m^{1/3}}\right)^x \cdot \frac{x}{m} \tag{3.70}$$

$$\leq \frac{n^{4/3}}{m^{1/3}} \cdot \left(\frac{4\delta}{n^{2/3}m^{1/3}}\right)^x \cdot \frac{x}{m} \tag{3.71}$$

$$\leq \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{4x^{1/x}\delta}{n^{2/3}m^{1/3}}\right)^x \tag{3.72}$$

$$\leq \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{8\delta}{n^{2/3}m^{1/3}}\right)^x . \tag{3.73}$$

Summing over all

$$\vec{a} \in \bigcup_{x=2}^{m/2-1} S_x$$

we get:

$$\sum_{\vec{a}\in\bigcup_{x=2}^{m/2-1} S_x} \frac{(\vec{\lambda}^-(\vec{a})-\vec{\lambda}^+(\vec{a}))^2}{\vec{\lambda}^-(\vec{a})} \leq \sum_{x=2}^{\infty} \frac{n^{4/3}}{m^{4/3}} \cdot \left(\frac{8\delta m^{2/3}}{n^{2/3}}\right)^x \tag{3.74}$$

$$= \sum_{x=0}^{\infty} 64\delta^2 \cdot \left(\frac{8\delta m^{2/3}}{n^{2/3}}\right)^x \tag{3.75}$$

$$\leq \frac{64\delta^2}{1-8\delta} \tag{3.76}$$

$$\leq 128\delta^2 . \tag{3.77}$$

where in equation (3.75) we used the fact that $n > m$, and equation (3.77) holds for $\delta \leq 1/16$. The lemma follows by applying Lemma 3.3. □

Lemma 3.15. *For $n \geq m$, $m \geq 12$ and $\delta \leq 1/4$,* $\sum_{x\geq m/2}\sum_{\vec{a}\in S_x} \left\|\mathrm{poi}(\vec{\lambda}^+(\vec{a})) - \mathrm{poi}(\vec{\lambda}^-(\vec{a}))\right\|_1 \leq 32\delta^3$.

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.,

$$|S_x| = \binom{m}{x} \leq \left(\frac{\mathrm{e}m}{x}\right)^x \leq \frac{m^x}{x} . \tag{3.78}$$

where we have used the premise that $m \geq 12$ and thus $x \geq 6$. By equations (3.55) and (3.56) (and the fact that $|x-y| \leq \max\{x,y\}$ for every positive real numbers x,y),

$$\sum_{x\geq m/2}\sum_{\vec{a}\in S_x} \left|\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a})\right| \leq \sum_{x\geq m/2}\sum_{\vec{a}\in S_x} \frac{n}{2}\prod_{i=1}^{m}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^{a_i} \tag{3.79}$$

$$= \sum_{x\geq m/2}\sum_{\vec{a}\in S_x} \frac{n}{2}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^{\sum_{i=1}^m a_i} \tag{3.80}$$

$$\leq \sum_{x=m/2}^{\infty} \frac{m^x}{x}\cdot\frac{n}{2}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x \tag{3.81}$$

$$\leq \sum_{x=m/2}^{\infty} \frac{2m^x}{m}\cdot\frac{n}{2}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^x \tag{3.82}$$

$$= \frac{n}{m}\sum_{x=m/2}^{\infty}\left(\frac{2\delta m^{1/3}}{n^{1/3}}\right)^x \tag{3.83}$$

$$= 8\delta^3 \sum_{x=m/2-3}^{\infty}\left(\frac{2\delta m^{1/3}}{n^{1/3}}\right)^x \tag{3.84}$$

$$\leq \frac{8\delta^3}{1-2\delta} \tag{3.85}$$

$$\leq 16\delta^3 \tag{3.86}$$

where in equation (3.85) we used the fact that $n \geq m$ and equation (3.86) holds for $\delta \leq 1/4$. The lemma follows by applying equation (3.1). $\square$

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

Lemma 3.16. *For $n \geq m$ and $\delta \leq 1/4$,*

$$\sum_{x\geq 3}\sum_{\vec{a}\in A_x}\left\|\mathrm{poi}(\vec{\lambda}^+(\vec{a})) - \mathrm{poi}(\vec{\lambda}^-(\vec{a}))\right\|_1 \leq 16\delta^3 . \tag{3.87}$$

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 (3.55) and (3.56) (and the fact that $|x-y| \leq \max\{x,y\}$ for every positive real numbers x,y),

$$\begin{aligned}
\sum_{x\geq 3}\sum_{\vec{a}\in A_x}\left|\vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a})\right| &\leq \sum_{x\geq 3}\sum_{\vec{a}\in A_x}\frac{n}{2}\prod_{i=1}^{m}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^{a_i} && (3.88)\\
&= \sum_{x\geq 3}\sum_{\vec{a}\in A_x}\frac{n}{2}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^{\sum_{i=1}^m a_i} && (3.89)\\
&\leq \sum_{x=3}^{\infty} m^{x-1}\cdot\frac{n}{2}\left(\frac{2\delta}{n^{1/3}m^{2/3}}\right)^{x} && (3.90)\\
&= \frac{n}{2m}\sum_{x=3}^{\infty}\left(\frac{2\delta m^{1/3}}{n^{1/3}}\right)^{x} && (3.91)\\
&= 4\delta^3\sum_{x=0}^{\infty}\left(\frac{2\delta m^{1/3}}{n^{1/3}}\right)^{x} && (3.92)\\
&\leq \frac{4\delta^3}{1-2\delta} && (3.93)\\
&\leq 8\delta^3 && (3.94)
\end{aligned}$$

where in equation (3.93) we used the fact that $n \geq m$ and equation (3.94) holds for $\delta \leq 1/4$. The lemma follows by applying equation (3.1). $\square$

We are now ready to finalize the proof of Theorem 3.1.

Proof of Theorem 3.1. Let $\mathcal{D}^+$ and $\mathcal{D}^-$ be as defined in equations (3.48) and (3.49), respectively, and recall that $\kappa = \delta \cdot n^{2/3}/m^{2/3}$ (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 $1/(n^{2/3}m^{1/3}) = \delta/(\kappa\cdot m)$, as required by Theorem 3.5. By Lemma 3.11, $\mathcal{D}^-$ is $(1/20)$-far from $\mathcal{P}^{\mathrm{eq}}$. Therefore, it remains to establish that equation (3.29) holds for $\mathcal{D}^+$ and $\mathcal{D}^-$. Consider the following partition of $\mathbb{N}^m$:

$$\left\{\{\vec{a}\}_{\vec{a}\in S_1}, A_2, \bigcup_{x=2}^{m/2-1} S_x, \{\vec{a}\}_{\vec{a}\in\bigcup_{x\geq m/2} S_x}, \{\vec{a}\}_{\vec{a}\in\bigcup_{x\geq 3} A_x}\right\}, \tag{3.95}$$

where $\{\vec{a}\}_{\vec{a}\in T}$ denotes the list of all singletons of elements in T. By Lemma 3.2 it follows that

$$\begin{aligned}
\left\|\mathrm{poi}(\vec{\lambda}^+)-\mathrm{poi}(\vec{\lambda}^-)\right\|_1 \;\leq\; & \sum_{\vec{a}\in S_1}\left\|\mathrm{poi}(\vec{\lambda}^+(\vec{a})-\mathrm{poi}(\vec{\lambda}^-(\vec{a}))\right\|_1 && (3.96)\\
& + \left\|\mathrm{poi}(\vec{\lambda}^+(A_2)-\mathrm{poi}(\vec{\lambda}^-(A_2))\right\|_1 && (3.97)\\
& + \left\|\mathrm{poi}(\vec{\lambda}^+(\bigcup_{x=2}^{m/2-1}S_x))-\mathrm{poi}(\vec{\lambda}^-(\bigcup_{x=2}^{m/2-1}S_x))\right\|_1 && (3.98)\\
& + \sum_{x\geq m/2}\sum_{\vec{a}\in S_x}\left\|\mathrm{poi}(\vec{\lambda}^+(\vec{a})-\mathrm{poi}(\vec{\lambda}^-(\vec{a}))\right\|_1 && (3.99)\\
& + \sum_{x\geq 3}\sum_{\vec{a}\in A_x}\left\|\mathrm{poi}(\vec{\lambda}^+(\vec{a})-\mathrm{poi}(\vec{\lambda}^-(\vec{a}))\right\|_1 . && (3.100)
\end{aligned}$$

For $\delta < 1/16$ we get by Lemmas 3.12–3.16 that

$$\left\|\mathrm{poi}(\vec{\lambda}^+)-\mathrm{poi}(\vec{\lambda}^-)\right\|_1 \;\leq\; 35\delta+48\delta^3\,, \tag{3.101}$$

which is less than

$$\frac{31}{60}-\frac{352\delta}{5}$$

for $\delta = 1/200$. Hence, the theorem follows from Theorem 3.5. □

3.4 A lower bound for testing independence

In this subsection we show that Theorem 3.1 implies a lower bound on testing independence of joint distributions over $[m]\times[n]$. A tester for independence gets samples from a distribution Q over $[m]\times[n]$ and distinguishes between Q that is a *product* distribution $Q_1\times Q_2$, where Q_1 is a distribution over $[m]$ and Q_2 is a distribution over $[n]$ (i. e., $Q(i,j)=Q_1(i)\cdot Q_2(j)$ for every $(i,j)\in[m]\times[n]$) and Q that is ε-far in the ℓ_1-distance from any product distribution. 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 Q has the property of independence if $\pi_1 Q$ and $\pi_2 Q$ are independent. With a slight abuse of the terminology, we shall say in such a case that Q is *independent*.

In Lemma 3.18 we prove that the problem of testing equivalence of a collection of distributions in the sampling model reduces to the problem of testing whether a pair of random variables are independent given samples from their joint distribution. Due to this relation between the two problems we get that any lower bound on the number of samples for the problem of equivalence implies a lower bound for testing independence. Therefore, the lower bound in Theorem 3.1 holds for testing independence, as stated in Corollary 3.19. In the proof of the lemma we shall use the following proposition.

Proposition 3.17 ([10]). *Let p,q be distributions over $[m]\times[n]$. If $\|p-q\|_1\leq\varepsilon/3$ and q is independent, then $\|p-\pi_1 p\times\pi_2 p\|_1\leq\varepsilon$.*

Lemma 3.18. *If there exists an algorithm T for testing whether a joint distribution Q over $[m]\times[n]$ is independent using a sample of size $s(m,n,\varepsilon)$, then there exists an algorithm T' for testing whether $\mathcal{D}\in\mathcal{P}^{\mathrm{eq}}$ in the unknown-weights sampling model using a sample of size $s(m,n,\varepsilon/3)$.*

If T is provided with (and uses) an explicit description of the marginal distribution $\pi_1 Q$, then the claim holds for T' in the known-weights sampling model.

Proof. Given a sample $\{(i_\ell, j_\ell)\}_{\ell=1}^{s(m,n,\varepsilon/3)}$ generated according to $\mathcal{D}$ in the sampling model with a weight vector $\vec{w}=(w_1,\ldots,w_m)$, 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 i). If $D_1,\ldots,D_m$ are identical and equal to some D^*, then for each $(i,j)\in[m]\times[n]$ we have that the probability of getting (i,j) in the sample is $w_i\cdot D^*(j)$. 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 $\mathcal{P}^{\mathrm{eq}}$, that is, $\sum_{i=1}^m w_i\cdot\|D_i-D^*\|_1>\varepsilon$ for every distribution D^* over $[n]$. In such a case, in particular we have that $\sum_{i=1}^m w_i\cdot\left\|D_i-\overline{D}\right\|_1>\varepsilon$, where $\overline{D}$ is the distribution over $[n]$ such that $\overline{D}(j)=\sum_{i=1}^m w_i\cdot D_i(j)$. By Proposition 3.17, the joint distribution Q over i and j (determined by the list $\mathcal{D}$ and the sampling process) is $\varepsilon/3$-far from independent, so T (and hence T') rejects with probability greater than $2/3$. $\square$

The next corollary follows directly from Lemma 3.18 and Theorem 3.1.

Corollary 3.19. *Given a joint distribution Q over $[m]\times[n]$ it is impossible to test if Q is independent or $1/48$-far from independent using $o(n^{2/3}m^{1/3})$ samples.*

4 A lower bound for testing equivalence in the uniform sampling model

In this section we prove the following theorem.

Theorem 4.1. *Testing the property $\mathcal{P}^{\mathrm{eq}}_{m,n}$ in the uniform sampling model for every $\varepsilon\le 1/2$ and $m\ge 64$ 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 $2/n$ 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 only a single list of m distributions each of which is identical to U_n, where U_n denotes the uniform distribution over $[n]$. Thus the single list in $\mathcal{U}_n^m$ belongs to $\mathcal{P}^{\mathrm{eq}}_{m,n}$. On the other hand we show, in Lemma 4.2, that $\mathcal{H}_n^m$ contains mostly lists of distributions that are $\Omega(1)$-far from $\mathcal{P}^{\mathrm{eq}}_{m,n}$. However, we also show, in Lemma 4.3, that any tester in the uniform sampling model that takes less than $n^{1/2}m^{1/2}/6$ samples cannot distinguish between $\mathcal{D}$ that was uniformly drawn from $\mathcal{H}_n^m$ and $\mathcal{D}=(U_n,\ldots,U_n)\in\mathcal{U}_n^m$. Details follow.

Lemma 4.2. *For every $m\ge 3$, with probability at least $(1-2/\sqrt{m})$ over the choice of $\mathcal{D}\sim U_{\mathcal{H}_n^m}$, where $U_{\mathcal{H}_n^m}$ denotes the uniform distribution over $\mathcal{H}_n^m$, we have that $\mathcal{D}$ is $(1/2)$-far from $\mathcal{P}^{\mathrm{eq}}_{m,n}$.*

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

$$\frac{1}{m}\sum_{i=1}^m \|D_i-\mathbf{v}\|_1 > \frac{1}{2}\,. \tag{4.1}$$

We shall actually prove a slightly more general statement, namely, that equation (4.1) holds for *every* vector $\mathbf{v}\in\mathbb{R}^n$. We define the function $med^{\mathcal{D}}:[n]\to[0,1]$ by $med^{\mathcal{D}}(j)=\mu_{\frac{1}{2}}(D_1(j),\ldots,D_m(j))$, where $\mu_{\frac{1}{2}}(x_1,\ldots,x_m)$ denotes the median of $x_1,\ldots,x_m$ (where if m is even, it is the value in position $m/2$ in sorted non-decreasing order). The sum $\sum_{i=1}^m |x_i-c|$ is minimized when $c=\mu_{\frac{1}{2}}(x_1,\ldots,x_m)$. Therefore, for every $\mathcal{D}$ and every vector $\mathbf{v}\in\mathbb{R}^n$,

$$\sum_{i=1}^m \left\|D_i-med^{\mathcal{D}}\right\|_1 \leq \sum_{i=1}^m \|D_i-\mathbf{v}\|_1\,. \tag{4.2}$$

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

$$\Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[D_i(j)\neq med^{\mathcal{D}}(j)\right]\geq\frac{1}{2}\left(1-\frac{1}{\sqrt{m}}\right)\,. \tag{4.3}$$

Fix $(i,j)\in[m]\times[n]$, and consider selecting $\mathcal{D}$ uniformly at random from $\mathcal{H}_n^m$. Suppose we first determine the values $D_{i'}(j)$ for $i'\neq i$, and set $D_i(j)$ in the end. For each (i',j) the probability that $D_{i'}(j)=0$ is $1/2$, and the probability that $D_{i'}(j)=2/n$ is $1/2$. If at least $m/2$ of the outcomes are 0, or more than $m/2$ are $2/n$, then the value of $med^{\mathcal{D}}(j)$ is already determined. Conditioned on this we have that the probability that $D_i(j)\neq med^{\mathcal{D}}(j)$ is exactly $1/2$. On the other hand, in the complementary event, i. e., when for odd m there are $(m-1)/2$ that are 0 and $(m-1)/2$ that are $2/n$, and for even m there are $m/2-1$ that are 0 and $m/2$ that are $2/n$, it holds that $med^{\mathcal{D}}(j)=D_i(j)$. We thus bound the probability of this event. First consider the case that m is even:

$$\Pr\left[Bin\left(m,\frac{1}{2}\right)=\frac{m}{2}-1\right]<\Pr\left[Bin\left(m,\frac{1}{2}\right)=\frac{m}{2}\right]=\binom{m}{\frac{m}{2}}\cdot\frac{1}{2^m}=\frac{m!}{\frac{m}{2}!\frac{m}{2}!}\cdot\frac{1}{2^m}\,. \tag{4.4}$$

By Stirling's approximation, $m!=\sqrt{2\pi m}\left(\frac{m}{\mathrm{e}}\right)^m \mathrm{e}^{\lambda_m}$, where λ_m is a parameter that satisfies

$$\frac{1}{12m+1}<\lambda_m<\frac{1}{12m}\,,$$

thus,

$$\frac{m!}{\frac{m}{2}!\frac{m}{2}!}\cdot\frac{1}{2^m} < \frac{\sqrt{2\pi m}(\frac{m}{\mathrm{e}})^m \mathrm{e}^{\frac{1}{12m}}}{(\sqrt{2\pi m/2}(\frac{m/2}{\mathrm{e}})^{m/2}\mathrm{e}^{\frac{1}{12m/2+1}})^2}\cdot\frac{1}{2^m} \tag{4.5}$$

$$= \frac{\mathrm{e}^{\frac{1}{12m}-\frac{2}{6m+1}}}{\sqrt{\pi m/2}} \tag{4.6}$$

$$< \frac{1}{\sqrt{\pi m/2}} \tag{4.7}$$

$$< \frac{1}{\sqrt{m}}\,. \tag{4.8}$$

where Inequalities (4.7) and (4.8) hold for $m \geq 3$. In case m is odd, the probability (over the choice of $D_{i'}(j)$ for $i' \neq i$) that $med^{\mathcal{D}}(j)$ is determined by $D_i(j)$ is

$$\Pr\left[Bin\left(m,\frac{1}{2}\right)=\frac{m-1}{2}\right]=\Pr\left[Bin\left(m+1,\frac{1}{2}\right)=\frac{m+1}{2}\right]$$

which is at most $1/\sqrt{m+1}$ by equation (4.8). Hence, equation (4.3) holds for all m and we obtain the following bound on the expectation

$$\mathrm{E}_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{E}_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\left|D_i(j)-med^{\mathcal{D}}(j)\right|\right] \tag{4.9}$$

$$= m\cdot n\cdot \Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}G=}\left[D_j(i)\neq med^{\mathcal{D}}(i)\right]\cdot\frac{2}{n} \tag{4.10}$$

$$\geq m\cdot n\cdot\frac{1}{2}\left(1-\frac{1}{\sqrt{m}}\right)\cdot\frac{2}{n} \tag{4.11}$$

$$= m-\sqrt{m}\,. \tag{4.12}$$

while the maximum value is bounded as

$$\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left|D_i(j)-med^{\mathcal{D}}(j)\right| \tag{4.13}$$

$$\leq \sum_{j=1}^{n}\frac{m}{2}\cdot\frac{2}{n} \tag{4.14}$$

$$= m\,. \tag{4.15}$$

Assume for the sake of contradiction that

$$\Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1 \leq \frac{m}{2}\right] > \frac{2}{\sqrt{m}}\,. \tag{4.16}$$

then by equation (4.15) we have,

$$\begin{aligned}
& \mathrm{E}_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1\right] && (4.17)\\
\leq\ & \Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1 \leq \frac{m}{2}\right]\cdot\frac{m}{2} && (4.18)\\
& +\left(1-\Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1 \leq \frac{m}{2}\right]\right)\cdot m && (4.19)\\
=\ & m-\Pr_{\mathcal{D}\sim U_{\mathcal{H}_n^m}}\left[\sum_{i=1}^{m}\left\|D_i - med^{\mathcal{D}}\right\|_1 \leq \frac{m}{2}\right]\cdot\frac{m}{2} && (4.20)\\
<\ & m-\sqrt{m}\,, && (4.21)
\end{aligned}$$

which contradicts equation (4.12). □

Recall that for an element $j \in [n]$ and a distribution D_i, $i \in [m]$, we let $\alpha_{i,j}$ denote the number of times the pair (i, j) appears in the sample (when the sample is selected in the uniform sampling model). Thus $(a_{1,j}, \ldots, a_{m,j})$ is the *sample histogram* of the element j. 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 $\mathbb{N}^{m\times n}$.

Lemma 4.3. *Let $\mathcal{U}$ be the distribution of the histogram of q samples taken from the uniform distribution over $[m]\times[n]$, 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,*

$$\|\mathcal{U}-\mathcal{H}\|_1 \leq \frac{4q^2}{mn}\,. \tag{4.22}$$

Proof. For every matrix $M \in \mathbb{N}^{m\times n}$, let A_M be the event of getting the histogram M i. e., $M[i, j] = x$ if element j is chosen exactly x times from distribution D_i in the sample; For every $\vec{x} = (x_1, \ldots, x_m) \in \mathbb{N}^m$, let $B_{\vec{x}}$ be the event of getting a histogram M with exactly x_i samples from distribution D_i i. e., such that for every $i \in [m]$, $\sum_{j\in[n]} M[i, j] = x_i$; Let C be the event of getting a histogram M such that there exists $(i, j) \in [m]\times[n]$ such that $M[i, j] \geq 2$; Let

$$V = \{B_{\vec{x}} : \Pr_{\mathcal{H}}\left(B_{\vec{x}} \cap \overline{C}\right) > 0\}$$

(where $\overline{C}$ denotes the event complementary to C). Recall that if we take strictly less than $n/2$ samples then conditioned on the event that there are no collisions in the sample, i. e., each element is drawn at most once, a sample from U_n and a sample from a random distribution in $\mathcal{H}_n$ are distributed exactly the same. We extend this observation for samples from $\mathcal{H}$ and $\mathcal{U}$ as follows. For every $B_{\vec{x}} \in V$, given the occurrence of $B_{\vec{x}} \cap \overline{C}$, i. e., given the histogram projected on the first coordinate and given that there were no collisions, samples from $\mathcal{H}$ and $\mathcal{U}$ are distributed the same. We shall use this fact to bound the statistical distance between $\mathcal{H}$ and $\mathcal{U}$. We next formalize this.

$$\begin{aligned} \|\mathcal{U}-\mathcal{H}\|_1 &= \sum_{A_M \subseteq C} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| + \sum_{A_M \subseteq \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| && (4.23) \\ &\leq \Pr_{\mathcal{U}}(C) + \Pr_{\mathcal{H}}(C) + \sum_{A_M \subseteq \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| . && (4.24) \end{aligned}$$

We start by bounding the third term in equation (4.24).

$$\begin{aligned} \sum_{A_M \subseteq \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| &= \sum_{B_{\vec{x}}} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| && (4.25) \\ &= \sum_{B_{\vec{x}} \in V} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| && (4.26) \\ &\quad + \sum_{B_{\vec{x}} \in \overline{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| . && (4.27) \end{aligned}$$

We next bound the expression in equation (4.26).

$$\begin{aligned} & \sum_{B_{\vec{x}} \in V} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| && (4.28) \\ &= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \Pr_{\mathcal{U}}\left(A_M \mid B_{\vec{x}} \cap \overline{C}\right) \cdot \left| \Pr_{\mathcal{U}}\left(\overline{C} \mid B_{\vec{x}}\right) - \Pr_{\mathcal{H}}\left(\overline{C} \mid B_{\vec{x}}\right) \right| && (4.29) \\ &= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) \left| \Pr_{\mathcal{U}}\left(\overline{C} \mid B_{\vec{x}}\right) - \Pr_{\mathcal{H}}\left(\overline{C} \mid B_{\vec{x}}\right) \right| && (4.30) \\ &= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) \left| \left(1 - \Pr_{\mathcal{U}}(C \mid B_{\vec{x}})\right) - \left(1 - \Pr_{\mathcal{H}}(C \mid B_{\vec{x}})\right) \right| && (4.31) \\ &= \sum_{B_{\vec{x}} \in V} \Pr_{\mathcal{U}}(B_{\vec{x}}) \left| \Pr_{\mathcal{U}}(C \mid B_{\vec{x}}) - \Pr_{\mathcal{H}}(C \mid B_{\vec{x}}) \right| && (4.32) \\ &\leq \Pr_{\mathcal{U}}(C) + \Pr_{\mathcal{H}}(C) , && (4.33) \end{aligned}$$

where in equation (4.29) we used the fact that for every $B_{\vec{x}} \in V, M \in \mathbb{N}^{m \times n}$, $\Pr_{\mathcal{U}}(B_{\vec{x}}) = \Pr_{\mathcal{H}}(B_{\vec{x}})$ and

$\Pr_{\mathcal{U}}\left(A_M \mid B_{\vec{x}} \cap \overline{C}\right) = \Pr_{\mathcal{H}}\left(A_M \mid B_{\vec{x}} \cap \overline{C}\right)$. Turning to the expression in equation (4.27),

$$\sum_{B_{\vec{x}} \in \overline{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \left| \Pr_{\mathcal{U}}(A_M) - \Pr_{\mathcal{H}}(A_M) \right| = \sum_{B_{\vec{x}} \in \overline{V}} \sum_{A_M \subseteq B_{\vec{x}} \cap \overline{C}} \Pr_{\mathcal{U}}(A_M) \tag{4.34}$$

$$\leq \sum_{B_{\vec{x}} \in \overline{V}} \Pr_{\mathcal{U}}(B_{\vec{x}}) \tag{4.35}$$

$$= \sum_{B_{\vec{x}} \in \overline{V}} \Pr_{\mathcal{H}}(B_{\vec{x}}) \tag{4.36}$$

$$= \sum_{B_{\vec{x}} \in \overline{V}} \Pr_{\mathcal{H}}(B_{\vec{x}} \cap C) \tag{4.37}$$

$$\leq \Pr_{\mathcal{H}}(C)\,, \tag{4.38}$$

where the first equality follows from the fact that $B_{\vec{x}} \in \overline{V}$, hence by the definition of V we get that $\Pr_{\mathcal{H}}(A_M) = 0$. 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

$$\frac{1}{\ell} + \frac{2}{\ell} + \cdots + \frac{q-1}{\ell}$$

which is $q^2/(2\ell)$. 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 4.1. Assume there is a tester, T, for the property $\mathcal{P}^{\mathrm{eq}}_{m,n}$ in the uniform sampling model, which takes $q \leq m^{1/2}n^{1/2}/6$ samples. By Lemma 4.2,

$$\Pr_{\mathcal{D} \sim U_{\mathcal{H}^m_n}}[A \text{ accepts } \mathcal{D}] \leq \frac{2}{\sqrt{m}} \cdot 1 + \left(1 - \frac{2}{\sqrt{m}}\right) \cdot \frac{1}{3} \tag{4.39}$$

$$= \frac{1}{3}\left(1 + \frac{4}{\sqrt{m}}\right) \tag{4.40}$$

$$\leq \frac{1}{2}\,, \tag{4.41}$$

where the last inequality holds for $m \geq 64$. By Lemma 4.3, 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 (4.41),

$$\left(\Pr_{\mathcal{D} \sim U_{\mathcal{H}^m_n}}[\text{A accepts } \mathcal{D}] - \Pr_{\mathcal{D} \sim U_{\mathcal{H}^m_n}}[\text{A accepts } \mathcal{D}] \right) \geq \frac{2}{3} - \frac{1}{2} > \frac{1}{18}\,. \qquad \square$$

5 Algorithms for testing equivalence in the sampling model

In this section we state our two main theorems (Theorems 5.1 and 5.2) regarding testing Equivalence in the sampling model. We prove Theorem 5.1 in this section. In Section 6 we prove a stronger version of Theorem 5.2 (Theorem 6.11) as well as a stronger version of Theorem 5.1 (Theorem 6.13). We have chosen to give the proof of Theorem 5.1, in addition to the proof of Theorem 6.13, because it is simpler than the latter.

Theorem 5.1. *Let $\mathcal{D}$ be a list of m distributions over $[n]$. It is possible to test whether $\mathcal{D} \in \mathcal{P}^{\mathrm{eq}}$ in the unknown-weights sampling model using a sample of size $\widetilde{O}((n^{2/3}m^{1/3}+m)\cdot \mathrm{poly}(1/\varepsilon))$.*

Theorem 5.2. *Let $\mathcal{D}$ be a list of m distributions over $[n]$. It is possible to test whether $\mathcal{D} \in \mathcal{P}^{\mathrm{eq}}$ in the known-weights sampling model using a sample of size $\widetilde{O}((n^{1/2}m^{1/2}+n)\cdot \mathrm{poly}(1/\varepsilon))$.*

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 $\widetilde{O}(\min\{n^{2/3}m^{1/3}+m, n^{1/2}m^{1/2}+n\}\cdot \mathrm{poly}(1/\varepsilon))$. Namely, as long as $n \geq m$, the complexity (in terms of n and m) grows as $\widetilde{O}(n^{2/3}m^{1/3})$, and when $m \geq n$, it grows as $\widetilde{O}(n^{1/2}m^{1/2})$.

5.1 Proof of Theorem 5.1

By Lemma 3.18, in order to prove Theorem 5.1 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 [10]. However, there was a certain flaw in one of the claims on which their analysis built (Theorem 15 in [10], which is attributed to [11]), and hence we fix the flaw next (building on [12], which is the full version of [11]).

Given a sampling access to a pair of distributions p and q and bounds on their ℓ_∞-norm b_p and b_q, respectively, the algorithm **Bounded-ℓ_∞-Closeness-Test** (Algorithm 1) tests the equivalence of p and q. The sample complexity of the algorithm depends on b_p and b_q, as described in the next theorem.

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

Theorem 5.3. *Let p and q be two distributions over the same finite domain R. Suppose that $\|p\|_\infty \leq b_p$ and $\|q\|_\infty \leq b_q$ where $b_q \geq b_p$. For every $\varepsilon \leq 1/4$, Algorithm* **Bounded-ℓ_∞-Closeness-Test**$(p,q,b_p,b_q,|R|,\varepsilon)$ *has the following properties.*

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

The algorithm takes $O\left(|R|\cdot b_p^{1/2}/\varepsilon^2 + |R|^2\cdot b_q\cdot b_p/\varepsilon^4\right)$ sample points from each distribution.

Bounded-ℓ_∞-Closeness-Test (sampling access to p, sampling access to q, b_p, b_q, $|R|$, $\varepsilon > 0$)

1. Take samples F_p^1 and F_p^2 from p, each of size t, where $t = O\left(|R| \cdot b_p^{1/2}/\varepsilon^2 + |R|^2 \cdot b_q \cdot b_p/\varepsilon^4\right)$
2. Take samples F_q^2 and F_q^2 from q, each of size t
3. Let $r_p = \mathrm{coll}(F_p^1)$ be the number of *collisions* in F_p^1
4. Let $r_q = \mathrm{coll}(F_q^1)$ be the number of *collisions* in F_q^1
5. Let $s_{p,q} = \sum_{j\in R}(\mathrm{occ}(j,F_p^2)\cdot\mathrm{occ}(j,F_q^2))$ be the number of collisions between F_p^2 and F_q^2
6. If $r_q > (7/4)\binom{t}{2} b_p$ then REJECT
7. Define $\delta \stackrel{\text{def}}{=} \varepsilon/|R|^{1/2}$, $r \stackrel{\text{def}}{=} \frac{2t}{t-1}(r_p + r_q)$ and $s \stackrel{\text{def}}{=} 2s_{p,q}$
8. If $r - s > t^2\delta^2/2$ then REJECT, otherwise ACCEPT

Algorithm 1: The algorithm for testing ℓ_1 distance when ℓ_∞ is bounded.

Proof. Following the analysis of [11, Lemma 5], we have that

$$\mathrm{Exp}[r-s] = t^2\|p-q\|_2^2\,, \tag{5.1}$$

and we have the following bounds on the variances of r_p, r_q and s (for some constant c):

$$\mathrm{Var}[s] \le ct^2\sum_{\ell\in R} p(\ell)q(\ell) + ct^3\sum_{\ell\in R}(p(\ell)q(\ell)^2 + p(\ell)^2q(\ell))\,, \tag{5.2}$$

$$\mathrm{Var}[r_p] \le ct^2\sum_{\ell\in R} p(\ell)^2 + ct^3\sum_{\ell\in R} p(\ell)^3\,, \tag{5.3}$$

and

$$\mathrm{Var}[r_q] \le ct^2\sum_{\ell\in R} q(\ell)^2 + ct^3\sum_{\ell\in R} q(\ell)^3\,. \tag{5.4}$$

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

$$\mathrm{Var}[s] \le ct^2\|p\|_\infty + ct^3(\|p\|_\infty\|q\|_2^2 + \|p\|_\infty^2) \le ct^2b_p + ct^3(b_p\|q\|_2^2 + b_p^2)\,, \tag{5.5}$$

$$\mathrm{Var}[r_p] \le ct^2\|p\|_2^2 + ct^3\|p\|_\infty\|p\|_2^2 \le ct^2\|p\|_\infty + ct^3\|p\|_\infty^2 \le ct^2b_p + ct^3b_p^2\,, \tag{5.6}$$

and

$$\mathrm{Var}[r_q] \le ct^2\|q\|_2^2 + ct^3\|q\|_\infty\|q\|_2^2 \le ct^2\|q\|_2^2 + ct^3b_q\|q\|_2^2\,. \tag{5.7}$$

By equations (5.5) and (5.7), a tighter bound on $\|q\|_2^2$ will imply a tighter bound on $\mathrm{Var}[s]$ and $\mathrm{Var}[r_q]$. To this end, the check in Step 6 in the algorithm was added to the original ℓ_2-**Distance-Test** of [11]. This check is beneficial in achieving a tighter bound on the sample complexity. First, prove that the tester distinguishes with high constant probability between the case that $\|q\|_2^2 > 2b_p$ and the case that $\|q\|_2^2 \leq (3/2)b_p$ by rejecting (with high probability) when $r_q > (7/4)\binom{t}{2}b_p$. Notice that by the triangle inequality $\|p-q\|_2 \geq \|q\|_2 - \|p\|_2$. Thus, if $\|q\|_2^2 > (3/2)b_p$ and $\|p\|_2^2 \leq b_p$ then it follows that

$$\|p-q\|_2 \geq \sqrt{(3/2)}b_p^{1/2} - b_p^{1/2}\,.$$

Therefore, by the fact that $b_p \geq 1/|R|$, we obtain that

$$\|p-q\|_1 \geq \|p-q\|_2 \geq \left(\sqrt{(3/2)}-1\right)/|R|^{1/2}$$

which is greater than $\varepsilon/(2|R|^{1/2})$ for $\varepsilon \leq 1/4$. Consider first the case that $\|q\|_2^2 > 2b_p$, so that $\mathrm{Exp}[r_q] > 2\binom{t}{2}b_p$. Then we can bound the probability that the tester accepts, that is, that $r_q \leq (7/4)\binom{t}{2}b_p$, by the probability that $r_q < (7/8)\mathrm{Exp}[r_q]$. In the case that $\|q\|_2^2 \leq (3/2)b_p$, so that $\mathrm{Exp}[r_q] \leq (3/2)\binom{t}{2}b_p$, we can bound the probability that the tester rejects, that is, that $r_q > (7/4)\binom{t}{2}b_p$, by the probability that $r_q > (7/6)\mathrm{Exp}[r_q]$. Then the probability to accept when $\|q\|_2^2 > 2b_p$ and reject when $\|q\|_2^2 \leq b_p$ is upper bounded by

$$\Pr[|r_q - \mathrm{Exp}[r_q]| > \mathrm{Exp}[r_q]/8\,.$$

Now, using the upper bound on the variance of r_q that we have (the first bound in equation (5.7)), the fact that for every distribution q over R, $\|q\|_2^2 \leq 1/|R|$ and $\mathrm{Exp}[r_q] = \binom{t}{2}\|q\|_2^2$, we have that

$$\begin{aligned}
\Pr[|r_q - \mathrm{Exp}[r_q]| > \mathrm{Exp}[r_q]/8] &\leq \frac{64\mathrm{Var}[r_q]}{\mathrm{Exp}^2[r_q]} && (5.8)\\
&\leq \frac{c\cdot(t^2\|q\|_2^2 + t^3\|q\|_\infty\|q\|_2^2)}{t^4\|q\|_2^4} && (5.9)\\
&= \frac{c}{t^2\|q\|_2^2} + \frac{c\|q\|_\infty}{t\|q\|_2^2} && (5.10)\\
&\leq \frac{c|R|}{t^2} + \frac{c|R|\|q\|_\infty}{t}\,. && (5.11)
\end{aligned}$$

To make this a small constant, we choose t so that

$$t = \Omega\left(|R|^{1/2} + |R|b_q\right)\,. \tag{5.12}$$

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

$$\Pr[|A - \mathrm{Exp}[A]| > \rho] \leq \mathrm{Var}[A]/\rho^2\,,$$

and so, for the case $\|p-q\|_2 > \delta$ (i. e., $\mathrm{Exp}[r-s] > t^2\delta^2$) we have that

$$\begin{aligned}
\Pr[r - s < t^2\delta^2/2] &\leq \Pr[|(r-s) - \mathrm{Exp}[r-s]| < t^2\delta^2/2] && (5.13)\\
&\leq \frac{4\mathrm{Var}[r-s]}{t^4\delta^4}\,, && (5.14)
\end{aligned}$$

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

$$\begin{aligned}
\Pr[r-s \ge t^2\delta^2/2] &\le \Pr[|(r-s)-\mathrm{Exp}[r-s]| < t^2\delta^2/4] && (5.15)\\
&\le \frac{16\mathrm{Var}[r-s]}{t^4\delta^4}\,. && (5.16)
\end{aligned}$$

That is, we want

$$\frac{\mathrm{Var}[r-s]}{t^4\delta^4}$$

which is of the order of

$$\frac{\mathrm{Var}[r-s]\cdot|R|^2}{t^4\varepsilon^4}$$

to be a small constant. If we use

$$\mathrm{Var}[r-s] = \frac{4t^2}{(t-1)^2}\left(\mathrm{Var}[r_p]+\mathrm{Var}[r_q]\right)+\mathrm{Var}[s]\,,$$

then we need to ensure that each of

$$\frac{\mathrm{Var}[r_p]\cdot|R|^2}{t^4\varepsilon^4}\,,\quad \frac{\mathrm{Var}[r_q]\cdot|R|^2}{t^4\varepsilon^4}\quad\text{and}\quad \frac{\mathrm{Var}[s]\cdot|R|^2}{t^4\varepsilon^4}$$

is a small constant, which by equations (5.5), (5.6), (5.7), and the premise that $\|q\|_2^2 \le 2b_p$, holds when

$$t = \Omega\left(|R|\cdot b_p^{1/2}/\varepsilon^2 + |R|^2\cdot b_q\cdot b_p/\varepsilon^4\right)\,, \tag{5.17}$$

since both $b_p, b_q \ge 1/|R|$, the bound in equation (5.17) dominates the bound in equation (5.12). Hence, the bound on the sample complexity follows. □

As a corollary of Theorem 5.3 we obtain:

Theorem 5.4. *Let Q be a distribution over $[m]\times[n]$ such that Q satisfies: $\|\pi_1 Q\|_\infty \le b_1$, $\|\pi_2 Q\|_\infty \le b_2$ and $b_2 \le b_1$. There is a test that takes $O(nmb_1^{1/2}b_2^{1/2}/\varepsilon^2 + n^2m^2b_1b_2^2/\varepsilon^4)$ 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 $\|Q\|_\infty \le b_2$ and that $\|\pi_1 Q\times\pi_2 Q\|_\infty \le b_1\cdot b_2$. Applying Theorem 5.3 we can test if Q is identical to $\pi_1 Q\times\pi_2 Q$ using sample of size $O(nmb_1^{1/2}b_2^{1/2}/\varepsilon^2 + n^2m^2b_1b_2^2/\varepsilon^4)$ from[3] 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 $2/3$. □

[3] We 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 $\pi_1 Q\times\pi_2 Q$.

Applying Theorem 5.4 with $b_1 = 1/m$, $b_2 = 1/n^{2/3}m^{1/3}$, and combining that in the sample analysis of **TestLightIndependence** [10], the following theorem is obtained:[4]

Theorem 5.5 ([10]). *There is an algorithm that given a distribution Q over $[m] \times [n]$ and an $\varepsilon > 0$,*

- *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 $\widetilde{O}\big((n^{2/3}m^{1/3}+m)\mathrm{poly}(\varepsilon^{-1})\big)$ samples.

Finally, Theorem 5.1 follows by combining Theorem 5.5 with Lemma 3.18.

6 Algorithms for tolerant testing of equivalence in the sampling model

Given a list of distributions $\mathcal{D}$, a *tolerant* equivalence tester is guaranteed to accept, with high probability, if the distributions in $\mathcal{D}$ are close (and not necessarily identical), and reject $\mathcal{D}$, with high probability, if the distributions in $\mathcal{D}$ are far. In this section we prove Theorems 6.11 and 6.13. Theorem 6.11 states that there is a tolerant equivalence tester taking $\tilde{O}(n^{1/2}m^{1/2}+n)$ samples in the known-weights sampling model. Theorem 6.13 states that there is a tolerant equivalence tester taking $\tilde{O}(n^{2/3}m^{1/3}+m)$ samples in the unknown-weights sampling model. A tolerant equivalence tester is also a non-tolerant equivalence tester, so Theorems 6.11 and 6.13 are stronger versions of Theorems 5.2 and 5.1, respectively.

6.1 An algorithm for tolerant testing of identity in the sampling model

Consider the problem where given sample access to a distribution p and an explicit description of a distribution q, the algorithm should accept ,with high probability, if p and q are identical, and should reject, with high probability, if p and q are far. This is called Identity Testing and is defined in [10]. If the algorithm is guaranteed to accept p and q that are close, and not necessarily identical, we refer to it as a tolerant identity test. We will use the tolerant identity test as a subroutine in the algorithms for tolerant testing of equivalence.

We next present and prove Theorem 6.3, which states that there is a tolerant identity tester taking $\tilde{O}(\sqrt{n})$ samples. The theorem is a restatement of theorems in [49] and [10]. The specific tolerance of Theorem 6.3 is somewhat complex and in order to state it we introduce the following new definitions.

Definition 6.1. For two parameters $\alpha, \beta \in (0,1)$, we say that a distribution p is an (α,β)*-multiplicative approximation* of a distribution q (over the same domain R) if the following holds.

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

[4] The procedure **TestLightIndependence** [10] is called by the TestIndependence [10] Algorithm. In Step (5) of **TestLightIndependence** use Algorithm 1 to obtain the desired result.

Definition 6.2. For $\alpha \in (0,1)$, we say that a distribution p is an *α-additive approximation* of a distribution q (over the same domain R) if for every $i \in R$, $|p(i) - q(i)| \leq \alpha$.

Theorem 6.3 (Adapted from [49], [10])**.** *Given sample access to p, a black-box distribution over a finite domain R, and q, an explicitly specified distribution over R, for every $0 < \varepsilon \leq 1/3$, algorithm* **Tolerant-Identity-Test**(p,q,n,ε) *has the following properties.*

1. *If $\|p-q\|_1 > 13\varepsilon$, the algorithm rejects with high constant probability.*
2. *If q is an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of some q' such that*

$$\left\|p-q'\right\|_1 \leq \frac{72\varepsilon^2}{\ell\sqrt{n}},$$

where $\ell = \log(n/\varepsilon)/\log(1+\varepsilon)$, the algorithm accepts with high constant probability (in particular, if q is an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of p or if

$$\|p-q\|_1 \leq \frac{72\varepsilon^2}{\ell\sqrt{n}},$$

the test accepts with high constant probability).

The algorithm takes $\tilde{O}(\sqrt{n}\,\mathrm{poly}(\varepsilon^{-1}))$ samples from p.

Let p be a distribution over some finite domain R, and let R' be a subset of R such that $p(R') > 0$ where $p(R') = \sum_{i \in R'} p(i)$. Denote by $p_{|R'}$ the restriction of p to R', i. e., $p_{|R'}$ is a distribution over R' such that for every $i \in R'$, $p_{|R'}(i) = p(i)/p(R')$. In the proof of Theorem 6.3 we shall use the following definitions and lemmas.

Definition 6.4 ([10])**.** Given an explicit distribution p over R, *Bucket*(p,R,α,β) is the partition $\{R_0,\ldots,R_\ell\}$ of R with $\ell = \log(1/\alpha)/\log(1+\beta)$, $R_0 = \{i : p(i) \leq \alpha\}$, such that for all j in $[\ell]$,

$$R_j = \left\{i : \alpha(1+\beta)^{j-1} < p(i) \leq \alpha(1+\beta)^j\right\}. \tag{6.1}$$

Definition 6.5 ([10])**.** Given a distribution p over R, and a partition $\mathcal{R} = \{R_1,\ldots,R_\ell\}$ of R, the coarsening $p_{\langle\mathcal{R}\rangle}$ is the distribution over $[\ell]$ with distribution $p_{\langle\mathcal{R}\rangle}(i) = p(R_i)$.

Theorem 6.6 ([10])**.** *Let p be a black-box distribution over a finite domain R and let S be a sample set from p.* $\mathrm{coll}(S)/\binom{|S|}{2}$ *approximates $\|p\|_2^2$ to within a factor of $(1 \pm \varepsilon)$, with probability at least $1-\delta$, provided that $|S| \geq c\sqrt{|R|}\varepsilon^{-2}\log(1/\delta)$ for some sufficiently large constant c.*

Lemma 6.7 ([10])**.** *Let p,q be distributions over R and let $R' \subseteq R$, then $\|p_{|R'} - q_{|R'}\|_1 \leq 2\|p-q\|_1/p(R')$.*

Lemma 6.8 ([10])**.** *For any distribution p over R, $\|p\|_2^2 - \|U_R\|_2^2 = \|p-U_R\|_2^2$.*

Lemma 6.9 (Based on [10])**.** *Let p,q be distributions over R and let $R' \subseteq R$, then*

$$\sum_{i \in R'} |p(i) - q(i)| \leq |p(R') - q(R')| + q(R')\|p_{|R'} - q_{|R'}\|_1.$$

Proof.

$$\sum_{i\in R'}|p(i)-q(i)| \le \sum_{i\in R'}\left|\frac{p(i)(p(R')-q(R'))}{p(R')}\right| + \sum_{i\in R'}\left|\frac{p(i)q(R')}{p(R')}-q(i)\right| \tag{6.2}$$

$$= |p(R')-q(R')| + \sum_{i\in R'}\left|\frac{p(i)q(R')}{p(R')}-q(i)\right| \tag{6.3}$$

$$= |p(R')-q(R')| + \sum_{i\in R'} q(R')\cdot\left|\frac{p(i)}{p(R')}-\frac{q(i)}{q(R')}\right| \tag{6.4}$$

$$= |p(R')-q(R')| + q(R')\cdot\left\|p_{|R'}-q_{|R'}\right\|_1 , \tag{6.5}$$

and the lemma is established. □

Lemma 6.10. *Let p,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*

$$q(i)\in p(i)\cdot[1-\varepsilon,1+\varepsilon] , \tag{6.6}$$

Then for every $i\in R'$,

$$q_{|R'}(i)\in p_{|R'}(i)\cdot\left[\frac{(1-\varepsilon)}{(1+\varepsilon)},\frac{(1+\varepsilon)}{(1-\varepsilon)}\right] \tag{6.7}$$

Proof. Equation (6.6) implies that $q(R')\in p(R')\,[1-\varepsilon,1+\varepsilon]$ and therefore

$$\frac{p(R')}{q(R')}\in\left[\frac{1}{1+\varepsilon},\frac{1}{1-\varepsilon}\right] .$$

Thus, we obtain that

$$\frac{q(i)}{q(R')}\in\frac{p(i)}{p(R')}\cdot\left[\frac{(1-\varepsilon)}{(1+\varepsilon)},\frac{(1+\varepsilon)}{(1-\varepsilon)}\right] ,$$

and the lemma follows. □

Proof of Theorem 6.3. Let E_1 be the event that for every i in $[\ell]$ we have that m_i approximates $\|p_{|R_i}\|_2^2$ to within a factor of $(1\pm\varepsilon^2)$. By Theorem 6.6 and the union bound, if S_i is such that $|S_i|\ge c\sqrt{n}\varepsilon^{-4}\log\ell$ 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 $|(|S_i|/|S|)-p(R_i)|\le\varepsilon/(2\ell)$. By Hoeffding's inequality E_2 occurs with probability at least $8/9$ for $|S|=\tilde{\Omega}(\ell^2\varepsilon^{-2})$. Let E_3 be the event that $\widetilde{p}_{\langle\mathcal{R}\rangle}$ and $\widetilde{q}_{\langle\mathcal{R}\rangle}$ are $\varepsilon/(2\ell)$-additive approximations of $p_{\langle\mathcal{R}\rangle}$ and $q_{\langle\mathcal{R}\rangle}$, respectively. By taking $\Theta(\varepsilon^{-2}\ell^2\log\ell)$ samples, E_3 occurs with probability at least $8/9$.

Let p and q be as described in Case 1, i. e., $\|p-q\|_1>13\varepsilon$. Suppose the algorithm accepts p and q. Conditioned on $E_1\cap E_2$, this implies that for each partition R_i for which Steps 4a - 4b were performed, which are those for which $q(R_i)\ge\varepsilon/\ell$, we have

$$\|p_{|R_i}\|_2^2\le\frac{(1+\varepsilon^2)^2}{|R_i|}\cdot\frac{1}{1-\varepsilon^2} ,$$

Tolerant-Identity-Test (sampling access to p, explicit description of q, n, $\varepsilon > 0$)

1. $\mathcal{R} \stackrel{\text{def}}{=} \{R_0, \cdots, R_\ell\} = Bucket(q, n, \varepsilon/n, \varepsilon/24)$
2. Let S be a set of $\tilde{\Theta}(\sqrt{n}\varepsilon^{-5}\log n)$ samples from p
3. Let H be the set of all x such that $q(x) > \varepsilon(1+\varepsilon)/n$
4. For each $R_i \subseteq H$ such that $q(R_i) \geq \varepsilon/\ell$
   (a) If $|S_i| < c\sqrt{n}\varepsilon^{-4}\log \ell$, where $S_i = S \cap R_i$ and c is the constant from Theorem 6.6, then REJECT
   (b) If $m_i > \frac{(1+\varepsilon^2)^2}{|R_i|}$, where $m_i = \text{coll}(S_i)/\binom{|S_i|}{2}$, then REJECT
5. Take $\Theta(\varepsilon^{-2}\ell\log\ell)$ samples and obtain a $\varepsilon/(4\ell)$-additive approximations $\widetilde{p}_{\langle\mathcal{R}\rangle}$ and $\widetilde{q}_{\langle\mathcal{R}\rangle}$ of $p_{\langle\mathcal{R}\rangle}$ and $q_{\langle\mathcal{R}\rangle}$, respectively
6. If $\|\widetilde{p}_{\langle\mathcal{R}\rangle} - \widetilde{q}_{\langle\mathcal{R}\rangle}\|_1 > 3\varepsilon/2$ then REJECT
7. ACCEPT

Algorithm 2: The algorithm for tolerant identity testing.

which is at most $\frac{1+4\varepsilon^2}{|R_i|}$ for $0 < \varepsilon \leq 1/3$. Thus, by Lemma 6.8 it follows that

$$\|p_{|R_i} - U_{|R_i}\|_2^2 = \|p_{|R_i}\|_2^2 - \|U_{|R_i}\|_2^2 \leq \frac{4\varepsilon^2}{|R_i|}\,. \tag{6.8}$$

For every $i \in [\ell]$,

$$\|q_{|R_i} - U_{|R_i}\|_2^2 = \sum_{j\in R_i}\left|\frac{q(j)}{q(R_i)} - \frac{q(R_i)/|R_i|}{q(R_i)}\right|^2 \leq \frac{1}{q(R_i)^2}\sum_{j\in R_i}\left|\varepsilon\frac{q(R_i)}{|R_i|}\right|^2 \leq \frac{\varepsilon^2}{|R_i|}\,, \tag{6.9}$$

where the second inequality follows from the bucketing definition. By the triangle inequality we obtain from equations (6.8) and (6.9) that

$$\|p_{|R_i} - q_{|R_i}\|_2^2 \leq \frac{9\varepsilon^2}{|R_i|}$$

and thus $\|p_{|R_i} - q_{|R_i}\|_1 \leq 3\varepsilon$. We also have that the sum of $q(R_i)$ over all R_i for which Steps 4a - 4b were not preformed is at most $\ell\cdot(\varepsilon/\ell) + n\cdot(\varepsilon(1+\varepsilon)^2/n) < 4\varepsilon$. For those R_i we use the trivial bound $\|p_{|R_i} - q_{|R_i}\|_1 \leq 2$. Also, $\|p_{\langle\mathcal{R}\rangle} - q_{\langle\mathcal{R}\rangle}\|_1 \leq 2\varepsilon$ by Step 6. So by Lemma 6.9 we get that $\|p-q\|_1 \leq 13\varepsilon$ in contradiction to our assumption. Therefore, the test accepts p and 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 q is an $(\varepsilon/n, (\varepsilon/24))$-multiplicative approximation of some q' such that p is $(72\varepsilon^2/\ell\sqrt{n})$-close to q'. Conditioned on E_2, every R_i that enters

Step 4a also passes this step, since otherwise we get, in contradiction to our assumption, that $q(R_i) \geq \varepsilon/\ell$ while $p(R_i) \leq 2\varepsilon/(3\ell)$. From the bucketing definition we have that for every $i \in [\ell]$ and for every $x \in R_i$,

$$\max(y \in R_i) \cdot \frac{1}{(1+(\varepsilon/24))} \leq q(x) \leq \min(y \in R_i) \cdot (1+(\varepsilon/24)),$$

implying that

$$q(x) \in \frac{q(R_i)}{|R_i|} \cdot \left[\frac{1}{(1+(\varepsilon/24))}, (1+(\varepsilon/24))\right]. \tag{6.10}$$

Since q is an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of q', we get by Lemma 6.10 that for every $R_i \subseteq H$ and every $x \in H$,

$$\frac{q(x)}{q(R_i)} \in \frac{q'(x)}{q'(R_i)} \cdot \left[\frac{(1-(\varepsilon/24))}{(1+(\varepsilon/24))}, \frac{(1+(\varepsilon/24))}{(1-(\varepsilon/24))}\right]. \tag{6.11}$$

Combining equations (6.10) and (6.11) we get that

$$q'(x) \in \frac{q'(R_i)}{|R_i|} \cdot \left[\frac{(1-(\varepsilon/24))}{(1+(\varepsilon/24))^2}, \frac{(1+(\varepsilon/24))^2}{(1-(\varepsilon/24))}\right], \tag{6.12}$$

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

$$\frac{q'(x)}{q'(R_i)} \in \left[\frac{(1-(\varepsilon/2))}{|R_i|}, \frac{(1+(\varepsilon/2))}{|R_i|}\right]. \tag{6.13}$$

By equation (6.13) we obtain that for every $R_i \subseteq H$

$$\|q'_{|R_i} - U_{|R_i}\|_2 \leq \varepsilon/(2\sqrt{|R_i|}). \tag{6.14}$$

For all subsets $R_i \subseteq H$ with $q(R_i) \geq \varepsilon/\ell$ we have that $q'(R_i) \geq \varepsilon/((1+\varepsilon)\ell)$, combined with the fact that

$$\|p - q'\|_1 \leq \frac{72\varepsilon^2}{\ell\sqrt{n}}$$

we get by Lemma 6.7 (for sufficiently large n) that

$$\|p_{|R_i} - q'_{|R_i}\|_1 \leq \varepsilon/(2\sqrt{n}). \tag{6.15}$$

This implies that

$$\|p_{|R_i} - q'_{|R_i}\|_2 \leq \|p_{|R_i} - q'_{|R_i}\|_1 \leq \varepsilon/(2\sqrt{n}) < \varepsilon/(2\sqrt{|R_i|}). \tag{6.16}$$

By the triangle inequality we get that

$$\|p_{|R_i} - U_{|R_i}\|_2 \leq \|p_{|R_i} - q'_{|R_i}\|_2 + \|q'_{|R_i} - U_{|R_i}\|_2 \leq \varepsilon/\sqrt{|R_i|}. \tag{6.17}$$

Therefore, by Lemma 6.8 it follows that

$$\|p_{|R_i}\|_2^2 = \|p_{|R_i} - U_{|R_i}\|_2^2 + \|U_{|R_i}\|_2^2 \leq (1+\varepsilon^2)/|R_i|. \tag{6.18}$$

Therefore, conditioned on $E_1 \cap E_2$ all such subsets will pass Step 4b. Since q is $\varepsilon/2$-close to q', by the triangle inequality p is ε-close to q and thus conditioned on E_3 the algorithm will pass Step 6 as well. Thus the algorithm accepts with probability at least $2/3$.

Finally, the sample complexity is $\tilde{O}(\sqrt{n}\varepsilon^{-5})$ from Step 2, which dominates the sample complexity of Step 5. □

6.2 An algorithm for tolerant testing of equivalence in the known-weights sampling model

In this section we prove Theorem 6.11. 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 6.11. *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 $[m]\times[n]$ such that $Q^{\mathcal{D},\vec{w}}(i,j) = w_i \cdot D_i(j)$. There is a test that works in the known-weights sampling model, which takes $\widetilde{O}((n^{1/2}m^{1/2}+n)\mathrm{poly}(1/\varepsilon))$ samples from $\mathcal{D}$, and whose output satisfies the following:*

- *If $\mathcal{D}$ is $24\varepsilon^2(1-3\varepsilon)/(\ell\sqrt{n})$-close to being in $\mathcal{P}^{\mathrm{eq}}$, where $\ell = \log(n/\varepsilon)/\log(1+\varepsilon)$, or if $Q^{\mathcal{D},\vec{w}}$ is an $(\varepsilon/n,\ \varepsilon/120)$-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 $\mathcal{P}^{\mathrm{eq}}$, then the test rejects with probability at least $2/3$.*

In the proof of Theorem 6.11 we shall use the following lemma:

Lemma 6.12. *Let Q be a joint distribution over $[m]\times[n]$. Let $\widetilde{Q}^1$ be a (α_1,β_1)-multiplicative approximation of $\pi_1 Q$. Let $\widetilde{Q}^2$ be a (α_2,β_2)-multiplicative approximation of $\pi_2 Q$. Denote by A_1 the set of all $i\in[m]$ such that $\widetilde{Q}^1(i) \ge \alpha_1(1+\beta_1)$. Denote by A_2 the set of all $j\in[n]$ such that $\widetilde{Q}^2(j) \ge \alpha_2(1+\beta_2)$. For every $\mathrm{B}_1 \subseteq A_1$ and every $\mathrm{B}_2 \subseteq A_2$, $\left(\widetilde{Q}^1 \times \widetilde{Q}^2\right)_{|B_1\times B_2}$ is a $\left(0, 2(\beta_1+\beta_2)/((1-\beta_1)\cdot(1-\beta_2))\right)$-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

$$\widetilde{Q}^1(i)\cdot\widetilde{Q}^2(j) \in \pi_1 Q(i)\cdot \pi_2 Q(j)\cdot[(1-\beta_1)\cdot(1-\beta_2),(1+\beta_1)\cdot(1+\beta_2)]\ . \tag{6.19}$$

From the facts that

$$\frac{(1+\beta_1)\cdot(1+\beta_2)}{(1-\beta_1)\cdot(1-\beta_2)} = 1+\frac{2(\beta_1+\beta_2)}{(1-\beta_1)\cdot(1-\beta_2)} \quad\text{and}\quad \frac{(1-\beta_1)\cdot(1-\beta_2)}{(1+\beta_1)\cdot(1+\beta_2)} > 1-\frac{2(\beta_1+\beta_2)}{(1-\beta_1)\cdot(1-\beta_2)},$$

and from Lemma 6.10 the lemma follows. □

Proof of Theorem 6.11. The test referred to in the statement of the theorem is Algorithm 3. Let E_1 be the event that $\widetilde{Q}^2$ is an $(\varepsilon/n,\ \varepsilon/120)$-multiplicative approximation of $\pi_2 Q$, as defined in Definition 6.1. 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 6.12, conditioned on E_1, we have that

$$\left(\vec{w}\times\widetilde{Q}^2\right)_{|[m]\times H}$$

is a $(0,\varepsilon/24)$-multiplicative approximation of

$$(\pi_1 Q\times \pi_2 Q)_{|[m]\times H}\ .$$

Equivalence-Known-Weights-Tolerant-Test (sampling access to a list of m distributions, $\mathcal{D}$, over $[n]$, in the known-weights sampling model, $0 < \varepsilon \le 1/3$)

1. Let Q denote $Q^{\mathcal{D},\vec{w}}$
2. Take $\Theta(\varepsilon^{-3} n \log n)$ samples and obtain a $(\varepsilon/n,\, \varepsilon/120)$-multiplicative approximation, $\widetilde{Q}^2$, of $\pi_2 Q$
3. Let H be the set of all $j \in [n]$ such that $\widetilde{Q}^2(j) > \varepsilon(1+\varepsilon)/n$ and let L be $[n] \setminus H$
4. If **Tolerant-Identity-Test** with parameters: $Q_{[m]\times H}$, $\left(\vec{w} \times \widetilde{Q}^2\right)_{|[m]\times H}$, $|H| \cdot m$, ε and confidence $1/9$, rejects then REJECT
5. $\mathcal{J} \stackrel{\text{def}}{=} \{[m] \times H, [m] \times L\}$
6. Take $\Theta(\varepsilon^{-2})$ samples and obtain a $(\varepsilon/2)$-additive approximations $\widetilde{Q}^{1\times 2}_{\langle \mathcal{J} \rangle}$ and $\widetilde{Q}_{\langle \mathcal{J} \rangle}$ of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{J} \rangle}$ and $Q_{\langle \mathcal{J} \rangle}$, respectively
7. If $\left\| \widetilde{Q}^{1\times 2}_{\langle \mathcal{J} \rangle} - \widetilde{Q}_{\langle \mathcal{J} \rangle} \right\|_1 > 2\varepsilon$ then REJECT
8. ACCEPT

Algorithm 3: The algorithm for tolerant testing of equivalence in the known-weights sampling model.

Thus,

$$\left\| \left(\vec{w} \times \widetilde{Q}^2\right)_{|[m]\times H} - (\vec{w} \times \pi_2 Q)_{|[m]\times H} \right\|_1 \le \varepsilon/24\,.$$

Let E_2 be the event that the application of **Tolerant-Identity-Test** returned a correct answer, as defined by Theorem 6.3. We run the amplified version of **Tolerant-Identity-Test**, therefore the additional parameter, which is the confidence parameter, is set to $1/9$, i. e., E_2 occurs with probability at least $8/9$.

Let $\mathcal{D}$ be 19ε-far from being in $\mathcal{P}^{\text{eq}}$ and assume the test accepts. Conditioned on E_2 this implies that

$$\left\| Q_{|[m]\times H} - \left(\vec{w} \times \widetilde{Q}^2\right)_{|[m]\times H} \right\|_1 \le 13\varepsilon\,.$$

By the triangle inequality, we obtain that conditioned on $E_1 \cap E_2$,

$$\left\| Q_{|[m]\times H} - (\vec{w} \times \pi_2 Q)_{|[m]\times H} \right\|_1 \le \varepsilon/24 + 13\varepsilon < 14\varepsilon\,. \tag{6.20}$$

Conditioned on E_1 we have that $Q([m] \times L) \le 3\varepsilon$, and therefore

$$Q([m] \times L) \cdot \left\| Q_{|[m]\times L} - (\vec{w} \times \pi_2 Q)_{|[m]\times L} \right\|_1 \le 6\varepsilon\,. \tag{6.21}$$

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

$$\left\|(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{J} \rangle} - Q_{\langle \mathcal{J} \rangle}\right\|_1 \le 3\varepsilon\,. \tag{6.22}$$

Combining equations (6.20) - (6.22), by Lemma 6.9, we have that

$$\|(\pi_1 Q \times \pi_2 Q) - Q\|_1 \le 3\varepsilon + 14\varepsilon + 6\varepsilon = 23\varepsilon\,. \tag{6.23}$$

Hence $\mathcal{D}$ is 23ε-close to being in $\mathcal{P}^{\mathrm{eq}}$, in contradiction to our assumption. It follows that the test accepts with probability at most $1/3$.

On the other hand, consider the case that either $\mathcal{D}$ is $24\varepsilon^2(1-3\varepsilon)/(\ell\sqrt{n})$-close to being in $\mathcal{P}^{\mathrm{eq}}$, or that $\pi_1 Q^{\mathcal{D},\vec{w}} \times \pi_2 Q^{\mathcal{D},\vec{w}}$ is an $(\varepsilon/n,\, \varepsilon/120)$-multiplicative approximation of $Q^{\mathcal{D},\vec{w}}$, and assume that the test rejects. For the sake of simplicity, denote $\left(\vec{w} \times \widetilde{Q}^2\right)$ by A and $(\vec{w} \times \pi_2 Q)$ by B. In case the test rejects in Step (4) then conditioned on E_2, we get by Theorem 6.3 that $A_{|[m]\times H}$ is not an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of any q' such that

$$\left\|Q_{|[m]\times H} - q'\right\|_1 \le \frac{72\varepsilon^2}{\ell\sqrt{n}}\,.$$

Conditioned on E_1, we have that $A_{|[m]\times H}$ is an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of $B_{|[m]\times H}$. Thus, conditioned on $E_1 \cap E_2$, we obtain that

$$\left\|Q_{|[m]\times H} - B_{|[m]\times H}\right\|_1 > \frac{72\varepsilon^2}{\ell\sqrt{n}}\,.$$

Based on the fact that $Q([m]\times H) = B([m]\times H)$, it follows, conditioned $E_1 \cap E_2$, that

$$\|Q - B\|_1 > Q([m]\times H)\cdot \left\|Q_{|[m]\times H} - B_{|[m]\times H}\right\|_1 > \frac{72\varepsilon^2(1-3\varepsilon)}{\ell\sqrt{n}}\,.$$

By Proposition 3.17 this implies that $\mathcal{D}$ is $24\varepsilon^2(1-3\varepsilon)/(\ell\sqrt{n})$-far from being in $\mathcal{P}^{\mathrm{eq}}$. By setting $q' = Q_{|[m]\times H}$ we also have that $A_{|[m]\times H}$ is not an $(\varepsilon/n, \varepsilon/24)$-multiplicative approximation of $Q_{|[m]\times H}$. Hence, there exists $(i, j) \in [m]\times H$ that satisfies either

$$A_{|[m]\times H}(i, j) > (1 + (\varepsilon/24))Q_{|[m]\times H}(i, j) \tag{6.24}$$

or

$$A_{|[m]\times H}(i, j) < (1 - (\varepsilon/24))Q_{|[m]\times H}(i, j)\,. \tag{6.25}$$

By Lemma 6.12, we get that $A_{|[m]\times H}$ is a $(0, \varepsilon/30)$-multiplicative approximation of $B_{|[m]\times H}$. Therefore, by equations (6.24) and (6.24), either it holds that

$$Q_{|[m]\times H}(i, j) < \frac{1 + (\varepsilon/30)}{1 + (\varepsilon/24)} B_{|[m]\times H}(i, j) \tag{6.26}$$

or that

$$Q_{|[m]\times H}(i, j) > \frac{1 - (\varepsilon/30)}{1 - (\varepsilon/24)} B_{|[m]\times H}(i, j)\,. \tag{6.27}$$

Since $Q([m]\times H) = B([m]\times H)$, we obtain from equations (6.26) and (6.27) that either

$$Q(i,j) < \frac{1+(\varepsilon/30)}{1+(\varepsilon/24)}B(i,j) \quad \text{or} \quad Q(i,j) > \frac{1-(\varepsilon/30)}{1-(\varepsilon/24)}B(i,j),$$

which by a simple calculation implies that Q is not a $(\varepsilon/n, \varepsilon/120)$-multiplicative approximation of $\vec{w}\times\pi_2 Q$.

Alternatively, in case the test rejects in 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 $\tilde{O}(n^{1/2}m^{1/2}\mathrm{poly}(\varepsilon^{-1}))$ so the overall sample complexity is $\tilde{O}((n^{1/2}m^{1/2}+n)\mathrm{poly}(\varepsilon^{-1}))$. □

6.3 An algorithm for tolerant testing of equivalence in the unknown-weights sampling model

In this section we prove the following theorem:

Theorem 6.13. *Let $\mathcal{D}$ be a list of m distributions over $[n]$. It is possible to distinguish between the case that $\mathcal{D}$ is $36\varepsilon^3/(\ell\sqrt{n})$-close to being in $\mathcal{P}^{\mathrm{eq}}$, where $\ell = \log(n/\varepsilon)/\log(1+\varepsilon)$ and the case that $\mathcal{D}$ is 25ε-far from being in $\mathcal{P}^{\mathrm{eq}}$ in the unknown-weights sampling model using a sample of size $\tilde{O}((n^{2/3}m^{1/3}+m)\cdot\mathrm{poly}(1/\varepsilon))$.*

Proof of Theorem 6.13. The algorithm referred to in the statement of the theorem is Algorithm 4. We note that we run the amplified version of **Tolerant-Identity-Test** and **Bounded-ℓ_∞-Closeness-Test** and that the additional parameter in the application of **Tolerant-Identity-Test** and **Bounded-ℓ_∞-Closeness-Test** is the confidence parameter. Let E_1 be the event that $\widetilde{Q}^1$ is an $(\varepsilon/m,\ \varepsilon/250)$-multiplicative approximation of $\pi_1 Q$. By taking a sample of size $\Theta(\varepsilon^{-3}m\log m)$, E_1 occurs with probability at least $20/21$. Let E_2 be the event that $\widetilde{Q}^2$ is an $(\varepsilon/n^{2/3}m^{1/3},\ \varepsilon/250)$-multiplicative approximation of $\pi_2 Q$. For a sample of size $\Theta(\varepsilon^{-3}n^{2/3}m^{1/3}\log n)$, we get, by Chernoff's inequality, that E_2 occurs with probability at least $20/21$. By Lemma 6.12, for every $0<\varepsilon\le 1/3$, we get, condition on $E_1\cap E_2$, that $\left(\widetilde{Q}^1\times\widetilde{Q}^2\right)_{|H_1\times H_2}$ is a $(0,\varepsilon/24)$-multiplicative approximation of $(\pi_1 Q\times\pi_2 Q)_{|H_1\times H_2}$. Thus, conditioned on $E_1\cap E_2$, we have that

$$\left\|\left(\widetilde{Q}^1\times\widetilde{Q}^2\right)_{|H_1\times H_2} - (\pi_1 Q\times\pi_2 Q)_{|H_1\times H_2}\right\|_1 \le \varepsilon/24. \tag{6.28}$$

Let E_3 be the event that the application of **Tolerant-Identity-Test** returned a correct answer, as defined by Theorem 6.3. E_3 occurs with probability at least $20/21$.

Let $\mathcal{D}$ be 25ε-far from being in $\mathcal{P}^{\mathrm{eq}}$ and assume the algorithm accepts. Then either **Tolerant-Identity-Test** returns accept or $\gamma < 3\varepsilon/2$. Consider the case that **Tolerant-Identity-Test** returns accept. Conditioned on E_3, by Theorem 6.3, we have that

$$\left\|\left(\widetilde{Q}^1\times\widetilde{Q}^2\right)_{|H_1\times H_2} - Q_{|H_1\times H_2}\right\|_1 \le 13\varepsilon.$$

Equivalence-Unknown-Weights-Tolerant-Test (sampling access to a list of m distributions, $\mathcal{D}$, over $[n]$, in the unknown-weights sampling model, $0 < \varepsilon \leq 1/8$)

1. Let Q denote $Q^{\mathcal{D},\vec{w}}$
2. Take $\Theta(\varepsilon^{-3} m \log m)$ samples and obtain an $(\varepsilon/m,\ \varepsilon/250)$-multiplicative approximation $\widetilde{Q}^1$ of $\pi_1 Q$
3. $\mathcal{R} \stackrel{\text{def}}{=} \{R_0, \cdots, R_\ell\} = Bucket(\widetilde{Q}^1, m, (1+\varepsilon)\varepsilon/m, \varepsilon)$
4. Let $L_1 = R_0$ and let $H_1 = [m] \setminus L_1$
5. Take $\Theta(\varepsilon^{-3} n^{2/3} m^{1/3} \log n)$ samples and obtain an $(\varepsilon/(n^{2/3}m^{1/3}),\ \varepsilon/250)$-multiplicative approximation $\widetilde{Q}^2$ of $\pi_2 Q$
6. Let H_2 be the set of all $j \in [n]$ such that $\widetilde{Q}^2(j) > \varepsilon(1+\varepsilon)/(n^{2/3}m^{1/3})$ and let $L_2 = [n] \setminus H_2$
7. Take $\Theta(\varepsilon^{-2})$ samples and let γ be the fraction of samples in $H_1 \times H_2$. If $\gamma < 3\varepsilon/2$ then skip the next step.
8. If **Tolerant-Identity-Test** with parameters: $Q_{|H_1 \times H_2}$, $(\widetilde{Q}^1 \times \widetilde{Q}^2)_{|H_1 \times H_2}$, $|H_1| \cdot |H_2|$,ε and confidence $1/21$ rejects then REJECT
9. Let S be a set of $\tilde{\Theta}(\ell^2 \varepsilon^{-2})$ independent samples. For each R_i such that $|S_i|/|S| \geq \varepsilon/\ell$, where $S_i = S \cap (R_i \times L_2)$, if **Bounded-$\ell_\infty$-Closeness-Test** with parameters: $(\pi_1 Q \times \pi_2 Q)_{|R_i \times L_2}$, $Q_{|R_i \times L_2}$, $4\ell/(\varepsilon n^{2/3} m^{1/3} |R_i|)$, $2\ell/(\varepsilon n^{2/3} m^{1/3})$, $|L_2| \cdot |R_i|$, ε and confidence $1/(21\ell)$, rejects then REJECT
10. $\mathcal{I} \stackrel{\text{def}}{=} \{H_1 \times H_2, L_1 \times H_2, R_0 \times L_2, \cdots, R_\ell \times L_2\}$
11. Take $\Theta(\varepsilon^{-2} \ell^2 \log \ell)$ samples and obtain an $\varepsilon/(2\ell)$-additive approximations $\widetilde{Q}^{1\times 2}_{\langle \mathcal{I} \rangle}$ and $\widetilde{Q}_{\langle \mathcal{I} \rangle}$ of $(\pi_1 Q \times \pi_2 Q)_{\langle \mathcal{I} \rangle}$ and $Q_{\langle \mathcal{I} \rangle}$, respectively
12. If $\left\| \widetilde{Q}^{1\times 2}_{\langle \mathcal{I} \rangle} - \widetilde{Q}_{\langle \mathcal{I} \rangle} \right\|_1 > 2\varepsilon$ then REJECT
13. ACCEPT

Algorithm 4: The algorithm for tolerant testing of equivalence in the unknown-weights sampling model.

By the triangle inequality and equation (6.28) we obtain that

$$\left\|(\pi_1 Q\times\pi_2 Q)_{|H_1\times H_2} - Q_{|H_1\times H_2}\right\|_1 \le 13\varepsilon+\varepsilon/24 < 14\varepsilon\,. \tag{6.29}$$

Consider the case $\gamma < 3\varepsilon/2$. Let E_4 be the event that $|\gamma - Q(H_1\times H_2)| \le \varepsilon/2$. By taking $\Theta(\varepsilon^{-2})$ samples, E_4 occurs with probability at least $20/21$. Then we have that

$$Q(H_1\times H_2)\le 2\varepsilon\,. \tag{6.30}$$

Let E_5 be the event that all applications of **Bounded-ℓ_∞-Closeness-Test** returned a correct answer, as defined by Theorem 5.3. By the union bound, E_5 occurs with probability at least $20/21$. Conditioned on E_5, we obtain that every R_i that passes Step 9 satisfies the following

$$\left\|(\pi_1 Q\times\pi_2 Q)_{|R_i\times L_2} - Q_{|R_i\times L_2}\right\|_1 \le \varepsilon\,. \tag{6.31}$$

Let E_6 to be the event that for every i in $[\ell]$ we have that $|(|S_i|/|S|) - Q(L_2\times R_i)| \le \varepsilon/(2\ell)$. By Hoeffding's inequality E_6 occurs with probability at least $20/21$ for $|S| = \tilde{\Omega}(\ell^2\varepsilon^{-2})$. From the fact that for every R_i that doesn't enter Step 9 we have that $|S_i|/|S| < \varepsilon/\ell$, we obtain, conditioned on E_6, that

$$Q(R_i\times L_2)\le 3\varepsilon/(2\ell)\,. \tag{6.32}$$

Let E_7 be the event that $\widetilde{Q}^{1\times 2}_{\langle\mathcal{I}\rangle}$ and $\widetilde{Q}_{\langle\mathcal{I}\rangle}$ are $\varepsilon/(2\ell)$-additive approximations of $(\pi_1 Q\times\pi_2 Q)_{\langle\mathcal{I}\rangle}$ and $Q_{\langle\mathcal{I}\rangle}$, respectively. By taking $\Theta(\varepsilon^{-2}\ell^2\log\ell)$ samples, E_7 occurs with probability at least $20/21$. Since we assume that the algorithm accepts $\mathcal{D}$ then, in particular, $\mathcal{D}$ passes Step 12. Therefore, conditioned on E_7, we have that

$$\left\|(\pi_1 Q\times\pi_2 Q)_{\langle\mathcal{I}\rangle} - Q_{\langle\mathcal{I}\rangle}\right\|_1 \le 3\varepsilon\,. \tag{6.33}$$

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

$$Q(L_1\times H_2)\le 3\varepsilon/2\,. \tag{6.34}$$

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

$$\left\|(\pi_1 Q\times\pi_2 Q)_{|I} - Q_{|I}\right\|_1 \le 2\,. \tag{6.35}$$

Combining equations (6.29) - (6.35), by Lemma 6.9, we have that

$$\|(\pi_1 Q\times\pi_2 Q) - Q\|_1 \le 3\varepsilon+14\varepsilon+2\varepsilon+\ell\cdot 3\varepsilon/(2\ell)\cdot 2+3\varepsilon/2\cdot 2 = 25\varepsilon\,. \tag{6.36}$$

Therefore, $\mathcal{D}$ is 25ε-close to being in $\mathcal{P}^{\mathrm{eq}}$ 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 $36\varepsilon^3/(\ell\sqrt{n})$-close to being in $\mathcal{P}^{\mathrm{eq}}$ and assume the algorithm rejects. Conditioned on $E_1\cap E_2$, we have that $(\widetilde{Q}^1\times\widetilde{Q}^2)_{|H_1\times H_2}$ is a $(0,\varepsilon/24)$-multiplicative approximation of $(\pi_1 Q\times\pi_2 Q)_{|H_1\times H_2}$. Therefore, conditioned on $E_1\cap E_2\cap E_3\cap E_4$, if we reject in Step 8, then we obtain by Theorem 6.3 that

$$\left\|Q_{|H_1\times H_2} - \big(\pi_1 Q\times\pi_2 Q\big)_{|H_1\times H_2}\right\|_1 > 72\cdot\frac{\varepsilon^2}{\ell\sqrt{n}}\,. \tag{6.37}$$

It follows, by Lemma 6.7, that

$$\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \frac{\pi_1 Q(H_1)\cdot \pi_2 Q(H_2)}{2}\cdot 72 \cdot \frac{\varepsilon^2}{\ell\sqrt{n}} \geq \frac{36\varepsilon^3}{\ell\sqrt{n}}\,.$$

If we reject in Step 9, then conditioned on $E_5 \cap E_6$, there is R_i such that $Q(R_i \times L_2) \geq \varepsilon/\ell$ in which the following holds,

$$\left\|(\pi_1 Q \times \pi_2 Q)_{|R_i\times L_2} - Q_{|R_i\times L_2}\right\|_1 > \varepsilon/(2\sqrt{n})\,. \tag{6.38}$$

Thus, by Lemma 6.7,

$$\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \frac{Q(R_i \times L_2)}{2}\cdot \varepsilon/(2\sqrt{n}) \geq \varepsilon^2/(4\ell\sqrt{n})\,.$$

If we reject in Step 12, then conditioned on E_7 it follows that $\|\pi_1 Q \times \pi_2 Q - Q\|_1 > \varepsilon$. Thus we get a contradiction to our assumption (that the algorithm rejects), which implies that the algorithm accepts $\mathcal{D}$ with probability at least $2/3$. 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 ($H_1 \times H_2$ in **Tolerant-Identity-Test** and $R_i \times L_2$ in **Bounded-ℓ_∞-Closeness-Test**). This affects the sample complexity only by a factor of $\mathrm{poly}(1/\varepsilon, \log n)$. For every R_i that enters Step 9, the number of required samples from the domain $R_i \times L_2$ in that step is bounded by

$$\tilde{O}((n^{2/3}\cdot |R_i|^{1/2}/m^{1/6} + n^{2/3}\cdot |R_i|/m^{2/3})\cdot \mathrm{poly}(1/\varepsilon))\,.$$

Thus, since ℓ is logarithmic in n and $1/\varepsilon$, the number of samples required by all the applications of **Bounded-ℓ_∞-Closeness-Test** is bounded by $\tilde{O}(n^{2/3}m^{2/3}\cdot \mathrm{poly}(1/\varepsilon))$. Therefore, the sample complexity is $\tilde{O}((n^{2/3}m^{1/3} + m)\cdot \mathrm{poly}(1/\varepsilon))$ as required. □

7 Testing (k,β)-clusterability in the query model

In this section we consider an extension of the property $\mathcal{P}^{\mathrm{eq}}_{m,n}$ 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 7.1. 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 $\mathcal{P}^{\mathrm{eq}}_{m,n}$. 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, for any ε and β (by adapting the algorithm we give below), to obtain a tester that distinguishes collections that are (k,β)-clusterable from collections that are ε-far from being (k,β)-clusterable. The complexity of this tester is $\tilde{O}(n\cdot k \cdot \mathrm{poly}(1/\varepsilon)))$. 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 [47]. We also note that for $\beta = 0$ the dependence on n must be at least $\Omega(n^{2/3})$ (for $m = 2$ and $k = 1$) [47]. Our algorithm works for $\beta = 0$ and slightly more generally, for $\beta = O(\varepsilon/\sqrt{n})$, has no dependence on m, has almost linear dependence on k, and its dependence on n grows like $\tilde{O}(n^{2/3})$.

Theorem 7.2. *Algorithm 5* **Clusterability-Test** *is a testing algorithm for (k,β)-clusterability of a list of distributions in the query model, which works for every $\varepsilon > 8\beta n^{1/2}$, and performs $\tilde{O}(n^{2/3} \cdot k \cdot \mathrm{poly}(1/\varepsilon))$ sampling queries.*

We build on the following theorem.

Theorem 7.3 ([11])**.** *Given parameter δ, and sampling access to distributions p, q over $[n]$, there is a test, ℓ_1-**Distance-Test**(p,q,ε,δ), which takes $O(\varepsilon^{-4} n^{2/3} \log n \log \delta^{-1})$ samples from each distribution and for which the following holds.*

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

Algorithm 5 is an adaptation of the diameter-clustering tester of [3], which applies to clustering vectors in $\mathbb{R}^d$. While often clustering algorithms rely on a method of evaluating distances between the objects that they cluster, the algorithm from [11] 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 [3] to construct a good distribution (k,b)-clusterability tester.

Clusterability-Test (query access to a list, $\mathcal{D}$, of m distributions over $[n]$ in the query model, k, β, $\varepsilon > 0$)

1. Pick the representative of the first cluster, rep_1, uniformly from $\mathcal{D}$
2. For $i := 1, \ldots, k$
   (a) Uniformly and independently select a set, $\mathcal{D}'$, of $2\ln(6(k+1))/\varepsilon$ distributions from $\mathcal{D}$
   (b) If there exists $D \in \mathcal{D}'$ such that ℓ_1-**Distance-Test**$(D, \mathrm{rep}_\ell, \beta/2, \varepsilon/12(k+1)\ln(6(k+1)))$ rejects for every $1 \le \ell \le i$, then set $\mathrm{rep}_{i+1} = D$, otherwise ACCEPT
3. REJECT

Algorithm 5: The algorithm for testing clusterability.

Proof of Theorem 7.2. Assume all applications of ℓ_1-Distance-Test returned a correct answer, as defined by Theorem 7.3. 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 $(\varepsilon/2)/(4n^{1/2}) \geq \beta$. 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, $\mathsf{rep}_1,\ldots,\mathsf{rep}_t$, the number of distributions $D_i \in \mathcal{D}$ such that $\|D_i - \mathsf{rep}_\ell\|_1 > \varepsilon/2$ for every $1 \leq \ell \leq t$, is at least $\varepsilon m/2$. To verify this, assume for the sake of contradiction that there are less than $\varepsilon m/2$ such distributions. Then the sum of the distances of each of these distributions to an arbitrary representative plus the sum over the other distributions of the distance between each distribution and the closest representative is less than εm, implying that the collection is ε-close to being $(k,0)$-clusterable.

Since in each iteration of the while loop, there are less than $k+1$ representative distributions, at least $\varepsilon m/2$ of the distributions in $\mathcal{D}$ are $\varepsilon/2$-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

$$(1-\varepsilon/2)^{2\ln(6(k+1))/\varepsilon} < \mathrm{e}^{\ln(6(k+1))} = \frac{1}{6(k+1)}\,.$$

By applying the union bound, the algorithm rejects $\mathcal{D}$ with probability greater than $2/3$. Since there are $O(\log k/\varepsilon)$ iterations, and in each there is a single application of the ℓ_1-distance test, by Theorem 7.3 the total number of samples used is as stated. □

Acknowledgements

We would like to thank the two anonymous referees for providing us with constructive comments.

References

[1] MICHAŁ ADAMASZEK, ARTUR CZUMAJ, AND CHRISTIAN SOHLER: Testing monotone continuous distributions on high-dimensional real cubes. In *Proc. 21st Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'10)*, pp. 56–65. ACM Press, 2010. Short version in Property Testing. [ACM:1873601.1873607] 299

[2] NOGA ALON, ALEXANDR ANDONI, TALI KAUFMAN, KEVIN MATULEF, RONITT RUBINFELD, AND NING XIE: Testing k-wise and almost k-wise independence. In *Proc. 39th STOC*, pp. 496–505. ACM Press, 2007. [doi:10.1145/1250790.1250863] 299

[3] NOGA ALON, SEANNIE DAR, MICHAL PARNAS, AND DANA RON: Testing of clustering. *SIAM J. Discrete Math.*, 16(3):393–417, 2003. Preliminary version in FOCS'00. [doi:10.1137/S0895480102410973] 298, 341

[4] NOGA ALON, YOSSI MATIAS, AND MARIO SZEGEDY: The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1):137–147, 1999. Preliminary version in STOC'96. [doi:10.1006/jcss.1997.1545] 296

[5] ALEXANDR ANDONI, PIOTR INDYK, KRZYSZTOF ONAK, AND RONITT RUBINFELD: External sampling. In *Proc. 36th Internat. Colloq. on Automata, Languages and Programming (ICALP'09)*, pp. 83–94. Springer, 2009. [doi:10.1007/978-3-642-02927-1_9] 299

[6] KHANH DO BA, HUY L. NGUYEN, HUY N. NGUYEN, AND RONITT RUBINFELD: Sublinear time algorithms for Earth Mover's Distance. *Theory Comput. Syst.*, 48(2):428–442, 2011. Preprint available at arXiv. [doi:10.1007/s00224-010-9265-8] 299

[7] ZIV BAR-YOSSEF, T. S. JAYRAM, RAVI KUMAR, D. SIVAKUMAR, AND LUCA TREVISAN: Counting distinct elements in a data stream. In *Proc. 6th Internat. Workshop on Randomization and Computation (RANDOM'02)*, pp. 1–10. Springer, 2002. [doi:10.1007/3-540-45726-7_1] 296

[8] TUĞKAN BATU: *Testing properties of distributions*. Ph. D. thesis, Computer Science department, Cornell University, 2001. `http://www.maths.lse.ac.uk/Personal/batu/papers/t.ps`. [ACM:934334] 299

[9] TUĞKAN BATU, SANJOY DASGUPTA, RAVI KUMAR, AND RONITT RUBINFELD: The complexity of approximating the entropy. *SIAM J. Comput.*, 35(1):132–150, 2005. Preliminary versions in CCC'02 and STOC'02. [doi:10.1137/S0097539702403645] 296, 299

[10] TUĞKAN BATU, ELDAR FISCHER, LANCE FORTNOW, RAVI KUMAR, RONITT RUBINFELD, AND PATRICK WHITE: Testing random variables for independence and identity. In *Proc. 42nd FOCS*, pp. 442–451. IEEE Comp. Soc. Press, 2001. [doi:10.1109/SFCS.2001.959920] 296, 298, 299, 318, 325, 329, 330

[11] TUĞKAN BATU, LANCE FORTNOW, RONITT RUBINFELD, WARREN D. SMITH, AND PATRICK WHITE: Testing that distributions are close. In *Proc. 41st FOCS*, pp. 259–269. IEEE Comp. Soc. Press, 2000. [doi:10.1109/SFCS.2000.892113] 296, 298, 310, 325, 326, 327, 341

[12] TUĞKAN BATU, LANCE FORTNOW, RONITT RUBINFELD, WARREN D. SMITH, AND PATRICK WHITE: Testing closeness of discrete distributions. Technical report, 2010. Preliminary version in FOCS'00. Accepted for publication in JACM. [arXiv:1009.5397] 296, 297, 298, 299, 304, 325

[13] TUĞKAN BATU, RAVI KUMAR, AND RONITT RUBINFELD: Sublinear algorithms for testing monotone and unimodal distributions. In *Proc. 36th STOC*, pp. 381–390. ACM Press, 2004. [doi:10.1145/1007352.1007414] 299

[14] VLADIMIR BRAVERMAN AND RAFAIL OSTROVSKY: Measuring k-wise independence of streaming data. Technical report, 2008. [arXiv:0806.4790v1] 296

[15] VLADIMIR BRAVERMAN AND RAFAIL OSTROVSKY: Measuring independence of datasets. In *Proc. 42nd STOC*, pp. 271–280. ACM Press, 2010. [doi:10.1145/1806689.1806728] 296

[16] VLADIMIR BRAVERMAN AND RAFAIL OSTROVSKY: Zero-one frequency laws. In *Proc. 42nd STOC*, pp. 281–290. ACM Press, 2010. [doi:10.1145/1806689.1806729] 296

[17] DON COPPERSMITH AND RAVI KUMAR: An improved data stream algorithm for frequency moments. In *Proc. 15th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'04)*, pp. 151–156. ACM Press, 2004. [ACM:982815] 296

[18] GRAHAM CORMODE, MAYUR DATAR, PIOTR INDYK, AND S. MUTHUKRISHNAN: Comparing data streams using Hamming norms (how to zero in). *IEEE Trans. Knowl. Data Eng.*, 15(3):529–540, 2003. Preliminary version in VLDB'02. [doi:10.1109/TKDE.2003.1198388] 296

[19] IMRE CSISZÁR: Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967. 296

[20] ARTUR CZUMAJ AND CHRISTIAN SOHLER: Testing expansion in bounded-degree graphs. *Combinatorics, Probability & Computing*, 19(5-6):693–709, 2010. Preliminary version in FOCS'07. [doi:10.1017/S096354831000012X] 296

[21] JOAN FEIGENBAUM, SAMPATH KANNAN, MARTIN J. STRAUSS, AND MAHESH VISWANATHAN: An approximate L^1-difference algorithm for massive data streams. *SIAM J. Comput.*, 32(1):131–151, January 2003. Preliminary verison in FOCS'99. [doi:10.1137/S0097539799361701] 296

[22] WILLIAM FELLER: *An introduction to probability theory and its applications*. Volume 1. Wiley, 3rd edition, 1967. 306

[23] ELDAR FISCHER AND ARIE MATSLIAH: Testing graph isomorphism. *SIAM J. Comput.*, 38(1):207–225, 2008. Preliminary version in SODA'06. [doi:10.1137/070680795] 296

[24] JESSICA H. FONG AND MARTIN STRAUSS: An approximate L^p difference algorithm for massive data streams. *Discrete Mathematics & Theoretical Computer Science*, 4(2):301–322, 2001. DMTCS. Preliminary version in STACS'00. 296

[25] ODED GOLDREICH AND DANA RON: On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography*, pp. 68–75. Springer, 2011. Preliminary version in ECCC. [doi:10.1007/978-3-642-22670-0_9] 296

[26] SUDIPTO GUHA, ANDREW MCGREGOR, AND SURESH VENKATASUBRAMANIAN: Sublinear estimation of entropy and information distances. *ACM Trans. Algorithms*, 5(4):35, 2009. Preliminary version in SODA'06. [doi:10.1145/1597036.1597038] 296, 299

[27] BERNARD HARRIS: The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis János Bolyai*, 16:323–355, 1975. DTIC Online. 296

[28] PIOTR INDYK AND ANDREW MCGREGOR: Declaring independence via the sketching of sketches. In *Proc. 19th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'08)*, pp. 737–745. ACM Press, 2008. [ACM:1347163] 296

[29] YUVAL ISHAI, EYAL KUSHILEVITZ, RAFAIL OSTROVSKY, AND AMIT SAHAI: Extracting correlations. In *Proc. 50th FOCS*, pp. 261–270. IEEE Comp. Soc. Press, 2009. [doi:10.1109/FOCS.2009.56] 296

[30] SATYEN KALE AND COMANDUR SESHADHRI: An expansion tester for bounded degree graphs. *SIAM J. Comput.*, 40(3):709–720, 2011. Preliminary versions in ICALP'08 and ECCC. [doi:10.1137/100802980] 296

[31] DONALD KNUTH: *The Art of Computer Programming: Seminumerical Algorithms*. Volume 2. Addison Wesley, Phillipines, 1969. 296

[32] REUT LEVI, DANA RON, AND RONITT RUBINFELD: Testing properties of collections of distributions. In *Proc. 2nd Symp. Innovations in Comput. Sci. (ICS'11)*, pp. 179–194, 2011. ICS'11. 295

[33] SHANG-KENG MA: Calculation of entropy from data of motion. *J. Statistical Physics*, 26(2):221–240, 1981. [doi:10.1007/BF01013169] 296

[34] DRAGOSLAV S. MITRINOVIĆ AND PETAR M. VASIĆ: *Analytic Inequalities*. Springer, 1970. 309

[35] ASAF NACHMIAS AND ASAF SHAPIRA: Testing the expansion of a graph. *Inform. and Comput.*, 208(4):309–314, 2010. Preliminary version in ECCC. [doi:10.1016/j.ic.2009.09.002] 296

[36] ILYA NEMENMAN, WILLIAM BIALEK, AND ROB DE RUYTER VAN STEVENINCK: Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E*, 69(5):056111, 2004. Preprint available at arXiv. [doi:10.1103/PhysRevE.69.056111] 296

[37] LIAM PANINSKI: Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. [doi:10.1162/089976603321780272] 296

[38] LIAM PANINSKI: Estimating entropy on *m* bins given fewer than *m* samples. *IEEE Trans. Inform. Theory*, 50(9):2200–2203, 2004. [doi:10.1109/TIT.2004.833360] 296

[39] LIAM PANINSKI: A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008. [doi:10.1109/TIT.2008.928987] 296

[40] SOFYA RASKHODNIKOVA, DANA RON, RONITT RUBINFELD, AND ADAM SMITH: Sublinear algorithms for approximating string compressibility. *Algorithmica*, 65(3):685–709, 2013. Preliminary version in RANDOM'07. [doi:10.1007/s00453-012-9618-6] 296

[41] SOFYA RASKHODNIKOVA, DANA RON, AMIR SHPILKA, AND ADAM SMITH: Strong lower bonds for approximating distributions support size and the distinct elements problem. *SIAM J. Comput.*, 39(3):813–842, 2009. Preliminary version in FOCS'07. [doi:10.1137/070701649] 296, 301

[42] BERO ROOS: On the rate of multivariate Poisson convergence. *J. Multivar. Anal.*, 69(1):120–134, 1999. [doi:10.1006/jmva.1998.1789] 305, 306

[43] RONITT RUBINFELD AND ROCCO A. SERVEDIO: Testing monotone high-dimensional distributions. *Random Structures & Algorithms*, 34(1):24–44, 2009. Preliminary version in STOC'05. [doi:10.1002/rsa.20247] 299

[44] RONITT RUBINFELD AND NING XIE: Robust characterizations of k-wise independence over product spaces and related testing results. *Random Structures & Algorithms*, 2012 (online). Preliminary version in ICALP 2010. [doi:10.1002/rsa.20423] 299

[45] STEVE P. STRONG, ROLAND KOBERLE, ROB R. DE RUYTER VAN STEVENINCK, AND WILLIAM BIALEK: Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197–200, 1998. [doi:10.1103/PhysRevLett.80.197] 296

[46] WOJCIECH SZPANKOWSKI: *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001. [ACM:517168] 301, 305

[47] PAUL VALIANT: *Testing symmetric properties of distributions*. Ph. D. thesis, CSAIL, MIT, 2008. DSpace@MIT. 296, 297, 298, 299, 301, 341

[48] PAUL VALIANT: Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011. Preliminary version in STOC'08. [doi:10.1137/080734066] 296, 301, 305, 310

[49] PATRICK WHITE: Testing random variables for independence and identity. Unpublished manuscript, 2002. 329, 330

[50] DAVID H. WOLPERT AND DAVID R. WOLF: Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995. [doi:10.1103/PhysRevE.52.6841] 296

[51] KENJI YAMANISHI: Probably almost discriminative learning. *Machine Learning*, 18(1):23–50, 1995. Preliminary version in COLT'92. [doi:10.1007/BF00993820] 296

AUTHORS

Reut Levi
Ph. D. candidate
School of Computer Science, Tel-Aviv University
reuti.levi@gmail.com

Dana Ron
Professor
School of Computer Science, Tel-Aviv University
danar@eng.tau.ac.il
http://www.eng.tau.ac.il/~danar

Ronitt Rubinfeld
Professor
School of Computer Science, Tel-Aviv University and CSAIL, MIT
ronitt@csail.mit.edu
http://people.csail.mit.edu/ronitt/

ABOUT THE AUTHORS

REUT LEVI received her M. Sc. from Tel-Aviv University and now is a Ph. D. candidate under the supervision of Dana Ron and Ronitt Rubinfeld. Her research focuses on testing properties of distributions.

DANA RON received her Ph. D. from the Hebrew University, Jerusalem, in 1995. Between 1995 and 1997 she was an NSF Postdoc at MIT. During the academic year 1997-98 she was a science scholar at the Mary Ingraham Bunting Institute (Radcliffe College) and at MIT. Since 1998 she has been a faculty member at the Faculty of Engineering, Tel Aviv University. During the academic year 2003-04 she was a fellow at the Radcliffe Institute for Advanced Study, Harvard University, Her research focuses on sublinear approximation algorithms and in particular on property testing.

RONITT RUBINFELD received her undergraduate degree at the University of Michigan and Ph. D. at the University of California, Berkeley under the direction of Manuel Blum. She held postdoctoral positions at DIMACS and the Hebrew University at Jerusalem. After several years as a faculty member at Cornell University and NEC Research Institute, she currently is on the faculties of MIT and Tel Aviv University. She has been a recipient of the ONR Young Investigator award, NSF Career Award, Sloan Fellowship, Cornell Association for Computer Science Undergraduates Faculty of the Year award and a Cornell College of Engineering Teaching award. She was an invited speaker at the International Congress of Mathematicians in 2006. Her research focuses on sub-linear time algorithms for big datasets.