Innovations in Computer Science 2011

# Testing Properties of Collections of Distributions

Reut Levi[1*] Dana Ron[2†] Ronitt Rubinfeld[3‡]
[1]School of Computer Science, Tel Aviv University
[2]School of Electrical Engineering, Tel Aviv University
[3]CSAIL, MIT and School of Computer Science, Tel Aviv University
reuti.levi@gmail.com danar@eng.tau.ac.il ronitt@csail.mit.edu

**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.

**Keywords:** property testing, distributions.

## 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 $\ell_1$ 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 [26, 40] or similar to another known distribution [13], and testing whether a joint distribution is independent [11]. Related tasks concern sublinear estimation of various measures of a distribution, such as its entropy [10, 27] or its support size [42]. Recently, general techniques have been designed to obtain nearly tight lower bounds on such testing and estimation problems [48, 49].

These types of questions have arisen in several disparate areas, including physics [34, 37, 46], cryptography and pseudorandom number generation [32], statistics [20, 28, 38-40, 50], learning theory [51], property testing of graphs and sequences (e.g., [21, 24, 26, 31, 36, 41]) and streaming algorithms (e.g., [4, 8, 15-19, 22, 25, 27, 29, 30]). 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 [12, 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

*Research supported by the Israel Science Foundation grant nos. 1147/09 and 246/08.

†Research supported the Israel Science Foundation grant number 246/08.

‡Research 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.

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 samples. In the first model, referred to as the *sampling model*, at each time step, the algorithm receives a pair of the form $(i, j)$ where $i \in [n]$ is distributed according to $D_j$ and $j$ is selected uniformly in $[m]$. In the second model, referred to as the *query model*, at each time step, the algorithm is allowed to specify $j \in [m]$ and receives $i$ that is distributed according to $D_j$. It is immediate that any algorithm in the sampling model can also be used in the query model. On the other hand, as is implied by our results, there are property testing problems which have a significantly larger sample complexity in the sampling model than in the query model.

In both models the task is to distinguish between the case that the tested distributions have the property and the case that they are $\epsilon$-far from having the property, for a given distance parameter $\epsilon$. Distance to the property is measured in terms of the average $\ell_1$-distance between the tested distributions and the closest collection of distributions that have the property. In all of our results, the dependence of the algorithms on the distance parameter $\epsilon$ is (inverse) polynomial. Hence, for the sake of succinctness, in all that follows we do not mention this dependence explicitly. We note that the sampling model can be extended to allow the choice of the distribution (that is, the index $j$) to be non-uniform (i.e., be determined by a weight $w_j$) and the distance measure is adapted accordingly.

### 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 $\ell_1$-distance). In [13], an algorithm is given that uses $\tilde{O}(n^{2/3})$ samples and distinguishes between the case that the two distributions are $\epsilon$-far and the case that they are $O(\epsilon/\sqrt{n})$-close. This algorithm has been shown to be nearly tight (in terms of the dependence on $n$) by Valiant [49]. Valiant also shows that in order to distinguish between the case that the distributions are $\epsilon$-far and the case that they are $\beta$-close, for two constants $\epsilon$ and $\beta$, 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 $\frac{1}{m}\sum_{i=1}^{m} \|D_i - D^*\|_1 \leq \text{poly}(\epsilon)/\sqrt{n}$), from the case in which there is no distribution $D^*$ for which $\frac{1}{m}\sum_{i=1}^{m} \|D_i - D^*\|_1 \leq \epsilon$. To solve this problem in the (uniform) sampling model with sample complexity $\tilde{O}(n^{2/3}m)$ (which ensures with high probability that each distribution is sampled $\tilde{\Omega}(n^{2/3}\log m)$ times), one can make $m-1$ calls to the algorithm of [13] to check that every distribution is close to $D_1$.

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

Both of our algorithms build on the close relation between testing equivalence and testing independence of a joint distribution over $[n] \times [m]$ which was studied in [11]. The $\tilde{O}(n^{2/3}m^{1/3}+m)$ algorithm follows from [11] after we fill in a certain gap in the analysis of their algorithm due to an imprecision of a claim given

in [12]. The $\tilde{O}(n^{1/2}m^{1/2}+n)$ algorithm exploits the fact that $j$ is selected uniformly (or, more generally, according to a known weight $w_j$) to improve on the $\tilde{O}(n^{2/3}m^{1/3}+m)$ algorithm (in the case that $m \geq n$).

Almost matching lower bounds. We show that the behavior of the upper bound on the sample complexity of the problem is not just an artifact of our algorithms, but rather (almost) captures the complexity of the problem. Namely, we give almost matching lower bounds of $\Omega(n^{2/3}m^{1/3})$ for $n = \Omega(m \log m)$ and $\Omega(n^{1/2}m^{1/2})$ (for every $n$ and $m$). The latter lower bound can be viewed as a generalization of a lower bound given in [13], 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 [49] 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 [11] for testing independence). As in [49], 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. Namely, we ask whether the distributions can be partitioned into at most $k$ subsets (clusters), such that within in cluster the distance between every two distributions is (very) small. We study this property in the query model, and give an algorithm whose complexity does not depend on the number of distributions and for which the dependence on $n$ is $\tilde{O}(n^{2/3})$. The dependence on $k$ is almost linear. The 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 closeness of distributions tester of [13]. Note that the results of [49] imply that this is tight to within polylogarithmic factors in $n$.

### 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 [49]). 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 [13]: Let $Q$ be a distribution over pairs of elements drawn from $[n] \times [m]$ (without loss of generality, assume $n \geq m$); what is the sample complexity in terms of $m$ and $n$ required to distinguish independent joint distributions, from those that are far from the nearest independent joint distribution (in term of $\ell_1$ distance)? The lower bound claimed in [11], contains a known gap in the proof. Similar gaps in the lower bounds of [13] for testing the closeness of distributions and of [10] for estimating the entropy of a distribution were settled by the work of [49], which applies to symmetric properties. Since independence is not a symmetric property, the work of [49] 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 [11], 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 [1, 2, 6, 7, 9, 14, 27, 44, 45].

## 1.3 Open problems and further research

There are many interesting directions to pursue concerning the testing of properties of collections of distributions, and because of the applicability of the model to a wide range of circumstances, we expect that new directions will present themselves. Here we give a few examples: One natural extension of our results is to

give algorithms for testing the property of clusterability for $k > 1$ in the sampling model. One may also consider testing properties of collections of distributions that are defined by certain measures of distributions, and may be less sensitive to the exact form of the distributions. For example, a very basic measure is the mean (expected value) of the distribution, when we view the domain $[n]$ as integers instead of element names, or when we consider other domains. Given this measure, we may consider testing whether the distributions all have similar means (or whether they should be modified significantly so that this holds). It is not hard to verify that this property can be quite easily tested in the query model by selecting $\Theta(1/\epsilon)$ distributions uniformly and estimating the mean of each. On the other hand, in the sampling model an $\Omega(\sqrt{m})$ lower bound is quite immediate even for $n = 2$ (and a constant $\epsilon$). 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

In this extended abstract we focus on one result: the lower bound of $\Omega(n^{2/3}m^{1/3})$ for testing equivalence. We give all the details for this result, where the more technical parts can be found in the appendix. All other results are provided in the full version of this paper [33].

## 2 Preliminaries

Let $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$, and let $\mathcal{D} = (D_1, \ldots, D_m)$ be a list of $m$ distributions, where $D_j : [n] \to [0, 1]$ and $\sum_{i=1}^{n} D_j(i) = 1$ for every $1 \leq j \leq m$. For a vector $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, let $\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$ denote the $L_1$ norm of the vector $v$.

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

Given a distance parameter $\epsilon$, 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 $\epsilon$-far from $\mathcal{P}$. We consider two models within which this task is performed.

1. **The Query Model.** In this model the testing algorithm may indicate an index $1 \leq j \leq m$ of its choice and it gets a sample $i$ distributed according to $D_j(i)$.

2. **The Sampling Model.** In this model the algorithm cannot select ("query") a distribution of its choice. Rather, it may obtain a pair $(i, j)$ where $j$ is selected uniformly (we refer to this as the *Uniform* sampling model) and $i$ is distributed according to $D_j(i)$.

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

## 3 A lower bound of $\Omega(\mathbf{n^{2/3}m^{1/3}})$ for testing equivalence in the uniform sampling model when $\mathbf{n = \Omega(m \log m)}$

In this section we prove the following theorem:

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

The proof of Theorem 1 consists of two parts. The first is a general theorem (Theorem 2) concerning testing symmetric properties of lists of distributions. This theorem extends a lemma of Valiant [49, Lem. 4.5.4] (which leads to what Valiant refers to as the "Wishful Thinking Theorem"). The second part is a construction of two lists of distributions to which Theorem 2 is applied. Our analysis uses a technique called *Poissonization* [47] (which was used in the past in the context of lower bounds for testing and estimating properties of distributions in [42, 48, 49]), and hence we first introduce some preliminaries concerning Poisson distributions. We later provide some intuition regard-

ing the benefits of Poissonization. All missing details of the analysis can be found in the appendix.

## 3.1 Preliminaries concerning Poisson distributions

For a positive real number $\lambda$, the Poisson distribution $\mathrm{poi}(\lambda)$ takes the value $x \in \mathbb{N}$ (where $\mathbb{N} = \{0, 1, 2, \ldots\}$) with probability $\mathrm{poi}(x; \lambda) = e^{-\lambda}\lambda^x/x!$. The expectation and variance of $\mathrm{poi}(\lambda)$ are both $\lambda$. For $\lambda_1$ and $\lambda_2$ we shall use the following bound on the $\ell_1$ distance between the corresponding Poisson distributions (for a proof see for example [42, Claim A.2]):

$$\|\mathrm{poi}(\lambda_1) - \mathrm{poi}(\lambda_2)\|_1 \leq 2|\lambda_1 - \lambda_2| \,. \tag{1}$$

For a vector $\vec{\lambda} = (\lambda_1, \ldots, \lambda_d)$ of positive real numbers, the corresponding *multivariate* Poisson distribution $\mathrm{poi}(\vec{\lambda})$ is the product distribution $\mathrm{poi}(\lambda_1) \times \ldots \times \mathrm{poi}(\lambda_d)$. That is, $\mathrm{poi}(\vec{\lambda})$ assigns each vector $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{N}^d$ 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 $\vec{a} = (a_1, \ldots, a_m) \in \mathbb{N}^m$, 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 $\vec{\lambda}^+ = (\lambda_1^+, \ldots, \lambda_d^+)$ and $\vec{\lambda}^- = (\lambda_1^-, \ldots, \lambda_d^-)$ of positive real values, we get from the proof of [49, Lemma 4.5.3] that,

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

for our purposes we shall use the following generalized lemma.

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

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

We shall also make use of the following Lemma.

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

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

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

$$p^{\mathcal{D},\kappa}(i; \vec{a}) \stackrel{\mathrm{def}}{=} \prod_{j=1}^{m} \mathrm{poi}(a_j; \kappa \cdot D_j(i)) \,. \tag{2}$$

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

$$\vec{\lambda}^{\mathcal{D},\kappa}(\vec{a}) = \sum_{i=1}^{n} p^{\mathcal{D},\kappa}(i; \vec{a}) \,. \tag{3}$$

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

## 3.2 Testability of symmetric properties of lists of distributions

In this subsection we state an important building block in the proof of prove the Theorem 1.

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

*If*

$$\left\|\mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^+,\kappa}\right) - \mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^-,\kappa}\right)\right\|_1 < \frac{16}{30} - \frac{352\delta}{5} \,, \tag{4}$$

*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 2. For an element $i \in [n]$ and a distribution $D_j$, $j \in [m]$, let $\alpha_{i,j}$ be the number of times the pair $(i,j)$ appears in the sample (when the sample is selected according to some sampling model). Thus $(\alpha_{i,1}, \ldots, \alpha_{i,m})$ is the *sample histogram* of the element $i$. The histogram of the elements' histograms is called the *fingerprint* of the sample. That is, the fingerprint indicates, for every $\vec{a} \in \mathbb{N}^m$, the number of elements $i$ such that $(\alpha_{i,1}, \ldots, \alpha_{i,m}) = \vec{a}$. As shown in [13], 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 $\vec{0} = (0, \ldots, 0)$ entry.

For example, consider the distributions $D_1$ and $D_2$ over $\{1,2,3\}$ such that $D_1[i] = 1/3$ for every $i \in \{1,2,3\}$, $D_2[1] = D_2[2] = 1/2$ 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 $(1,1), (1,2), (2,2), (3,1)$, where the first coordinate denotes the element and the second coordinate denotes the distribution. 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 $(0,1,1,0,1,0,0,\ldots)$ for the following order of histograms: $((0,0),(0,1),(1,0),(2,0)(1,1),(0,2),(3,0),\ldots)$.

In order to prove Theorem 2, we would like to show that the distributions of the fingerprints when the sample is generated according to $\mathcal{D}^+$ and when it is generated according to $\mathcal{D}^-$ are similar, for a sample size that is below the lower bound stated in the theorem. For each choice of element $i \in [n]$ and a distribution $D_j$, 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 can't 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* [47]. In the Poissonized process the number of samples we take from each distribution $D_j$, denoted by $\kappa'_j$, 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 $\kappa'_j > \kappa_j$, for every $j$, where $\kappa_j$ is the number of samples taken from $D_j$. 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 $i$ 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 fortunately for us has been studied by Roos [43] (the connection is due to Valiant [48]). See details in Appendix B.

## 3.3 Proof of Theorem 1

In this subsection we show how to apply Theorem 2 to two lists of distributions, $\mathcal{D}^+$ and $\mathcal{D}^-$, which we will define shortly, where $\mathcal{D}^+ \in \mathcal{P}^{\text{eq}} = \mathcal{P}^{\text{eq}}_{m,n}$ while $\mathcal{D}^-$ is $(1/20)$-far from $\mathcal{P}^{\text{eq}}$. Recall that by the premise of Theorem 1, $n \geq cm \log m$ for some sufficiently large constant $c > 1$. In the proof it will be convenient to assume that $m$ is even and that $n$ (which corresponds in the lemma 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** *For every two even integers $t$ and $m$, there exists a $0/1$-valued matrix $M$ with $t$ rows and $m$ columns for which the following holds:*

1. *In each row and each column of $M$, exactly half of the elements are 1 and the other half are 0.*

2. *For every integer* $2 \leq x < m/2$*, and for every subset* $S \subseteq [m]$ *of size* $x$*, the number of rows* $i$ *such that* $M[i,j] = 1$ *for every* $j \in S$ *is at 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* $t \cdot \left(\frac{1}{2^x} + \sqrt{\frac{2x \ln m}{t}}\right)$.

Lemma 3 is proved using the probabilistic method. Namely, we describe how to construct a matrix $M$ in a certain random manner, and then prove that the conditions in the lemma hold with non-zero probability. Specifically. Consider selecting $M$ randomly as follows: Denote the first $t/2$ rows of $M$ by $F$. For each row in $F$, pick, independently from the other $t/2-1$ rows in $F$, a random half of its elements to be 1, and the other half of the elements to be 0. Rows $t/2+1, \ldots, t$ are the negations of rows $1, \ldots, t/2$, respectively. Thus, in each row and each column of $M$, exactly half of the elements are 1 and the other half are 0. Proving that the second condition in the lemma holds with non-zero probability is deferred to Appendix C.

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

$$D_j^+(i) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2} \\ \frac{1}{n} & \text{if } \frac{n}{2} < i \leq n \\ 0 & o.w. \end{cases} \tag{5}$$

We now turn to defining $\mathcal{D}^-$. Let $M$ be a matrix as in Lemma 3 for $t = n/2$. For every $j \in [m]$:

$$D_j^-(i) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{n^{2/3}m^{1/3}} & \text{if } 1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2} \\ \frac{2}{n} & \text{if } \frac{n}{2} < i \leq n \\ & \text{and } M[i - n/2, j] = 1 \\ 0 & o.w. \end{cases} \tag{6}$$

For both $\mathcal{D}^+$ and $\mathcal{D}^-$, we refer to the elements $1 \leq i \leq \frac{n^{2/3}m^{1/3}}{2}$ as the *heavy* elements, and to the elements $\frac{n}{2} \leq i \leq n$, as the *light* elements. Observe that each heavy element has exactly the same probability weight, $\frac{1}{n^{2/3}m^{1/3}}$, in all distributions $D_j^+$ and $D_j^-$. On the other hand, for each light element $i$, while $D_j^+(i) = \frac{1}{n}$ (for every $j$), in $\mathcal{D}^-$ we have that $D_j^+(i) = \frac{2}{n}$ for half of the distributions, the distributions selected by the $M$, and $D_j^+(i) = 0$ for half of the distributions, the distributions which are not selected by $M$. We later use the properties of $M$ to bound the $\ell_1$ 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 [12]). 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 won't include (with high probability) any light element $i$ such that $\alpha_{i,1} = a_1$ and $\alpha_{i,2} = a_2$ (for both $\mathcal{D}^+$ and $\mathcal{D}^-$). Moreover, the expected number of elements $i$ such that $(\alpha_{i,1}, \alpha_{i,2}) = (1,0)$ is the same in $\mathcal{D}^+$ and $\mathcal{D}^-$, as well as the variance (from symmetry, the same applies to $(0,1)$). Thus, most of the difference between the fingerprints' distributions is due to the numbers of elements $i$ such that $(\alpha_{i,1}, \alpha_{i,2}) \in \{(1,1), (2,0), (0,2)\}$. For these settings of $\vec{a}$ we do expect to see a non-negligble difference for light elements between $\mathcal{D}^+$ and $\mathcal{D}^-$ (in particular, we can't 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 $(a_{i,1}, a_{i,2}) = (1,1)$ (and similarly for $(2,0)$ and $(0,2)$), 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)) in the coordinates corresponding to $\vec{a}$ such that $a_1 + a_2 = 2$. We can then apply Lemmas 1 and 2 to obtain Equation (4) in Theorem 2.

Turning to $m > 2$, it is no longer true that in a sample of size $o(n^{2/3}m^{1/3})$ we won't get histogram vectors $\vec{a}$ such that $\sum_{j=1}^{m} a_i > 2$ 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 $\sum_{j=1}^{m} a_j$ and depending on whether all $a_j's$ are in $\{0,1\}$, or there exists a least one $a_j$ such that $a_j \geq 2$. 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 $\sum_{j=1}^{m} a_i = 2$ 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.

We next introduce some more notation, which will be used throughout the remainder of the proof of Theorem 1. Let $S_x$ be the set of vectors that contain exactly $x$ coordinates that are 1, and all the rest are 0 (which corresponds to an element that was sampled once or 0 times by each distribution). Let $A_x$ be the set of vector 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 $\mathbb{N}^m$:

$$S_x \stackrel{\text{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m \ : \ \begin{array}{r} \sum_{j=1}^{m} a_j = x \text{ and} \\ \forall j \in [m], a_j < 2 \end{array} \right\} ,$$

and

$$A_x \stackrel{\text{def}}{=} \left\{ \vec{a} \in \mathbb{N}^m \ : \ \begin{array}{r} \sum_{j=1}^{m} a_j = x \text{ and} \\ \exists j \in [m], a_j \geq 2 \end{array} \right\}$$

For $\vec{a} \in \mathbb{N}^m$, let $\sup(\vec{a}) \stackrel{\text{def}}{=} \{j : a_j \neq 0\}$ denote the *support* of $\vec{a}$, and let $I_M(\vec{a}) \stackrel{\text{def}}{=} \left\{ i : D_j^-(i) = \frac{2}{n} \quad \forall j \in \sup(\vec{a}) \right\}$. Note that in terms of the matrix $M$ (based on which $\mathcal{D}^-$ is defined), $I_M(\vec{a})$ consists of the rows in $M$ whose restriction to the support of $\vec{a}$ contains only 1's. In terms of the $\mathcal{D}^-$, it corresponds to the set of light elements that might have a sample histogram of $\vec{a}$ (when sampling according to $\mathcal{D}^-$).

The proof of the next lemma appears in Appendix C.

**Lemma 4** *$\mathcal{D}^-$ is (1/20)-far from $\mathcal{P}^{\text{eq}}$ for every $m > 5$ and $n \geq c \ln m$ where $c$ is a sufficiently large constant.*

In what follows we work towards establishing that Equation (4) in Theorem 2 holds for $\mathcal{D}^+$ and $\mathcal{D}^-$. Set $\kappa = \delta \cdot \frac{n^{2/3}}{m^{2/3}}$, where $\delta$ 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)). By the definition of $\vec{\lambda}^+$, for each $\vec{a} \in \mathbb{N}^m$,

$$\begin{aligned}
&\vec{\lambda}^+(\vec{a}) \\
&= \sum_{i=1}^{n} \prod_{j=1}^{m} \frac{(\kappa \cdot D_j^+(i))^{a_j}}{e^{\kappa \cdot D_j^+(i)} \cdot a_j!} \\
&= \sum_{i=1}^{n^{2/3}m^{1/3}/2} \prod_{j=1}^{m} \frac{(\delta/m)^{a_j}}{e^{\delta/m} \cdot a_j!} \\
&\quad + \sum_{i=n/2+1}^{n} \prod_{j=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{e^{\delta/(n^{1/3}m^{2/3})} \cdot a_j!} \\
&= \frac{n^{2/3}m^{1/3}}{2e^{\delta}} \prod_{j=1}^{m} \frac{(\delta/m)^{a_j}}{a_j!} \\
&\quad + \frac{n}{2e^{\delta(m/n)^{1/3}}} \prod_{j=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} \ . \qquad (7)
\end{aligned}$$

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

$$\begin{aligned}
&\vec{\lambda}^-(\vec{a}) \\
&= \frac{n^{2/3}m^{1/3}}{2e^{\delta}} \prod_{j=1}^{m} \frac{(\delta/m)^{a_j}}{a_j!} \\
&\quad + \frac{1}{e^{\delta(m/n)^{1/3}}} \sum_{i \in I_M(\vec{a})} \prod_{j=1}^{m} \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} . \qquad (8)
\end{aligned}$$

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 (7) and (8) demonstrate why we choose $M$ with the specific properties defined in Lemma 3. First of all, in order for every $D_j^-$ to be a probability distribution, we want each column of $M$ to sum up to exactly $n/2$. We also want each row of $M$ to sum up to exactly $m/2$, in order to get

$\prod_{j=1}^{m} e^{-\kappa \cdot D_j^+(i)} = \prod_{j=1}^{m} e^{-\kappa \cdot D_j^-(i)}$. Finally, we would have liked $|I_M(\vec{a})| \cdot \prod_{j=1}^{m} 2^{a_j}$ 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 ensures that $|I_M(\vec{a})|$ is sufficiently close to $\frac{n}{2} \cdot \frac{1}{2^x}$. 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 (4) of histogram vectors $\vec{a} \in S_1$ (i.e., vectors of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$) which correspond to the number of elements that are sampled only by one distribution, once. We prove that in the Poissonized 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 5**

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

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

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

By Equations (7) and (8), 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 (1). □

We now turn to bounding the contribution to Equation (4) of histogram vectors $\vec{a} \in A_2$ (i.e., vectors of the form $(0, \ldots, 0, 2, 0, \ldots, 0)$ which correspond to the number of elements that are sampled only by one distribution, twice.

**Lemma 6**

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

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

$$\sum_{i \in I_M(\vec{a})} \prod_{j=1}^{m} \frac{(2\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} = n \prod_{j=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!}. \tag{9}$$

By Equations (7), (8) and (9) it follows that

$$\begin{aligned} &\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a}) \\ &= \frac{n}{2e^{\delta(m/n)^{1/3}}} \prod_{j=1}^{m} \frac{(\delta/(n^{1/3}m^{2/3}))^{a_j}}{a_j!} \\ &= \frac{n^{1/3}\delta^2}{4e^{\delta(m/n)^{1/3}}m^{4/3}} \ , \end{aligned} \tag{10}$$

and that

$$\begin{aligned} \vec{\lambda}^-(\vec{a}) &\geq \frac{n^{2/3}m^{1/3}}{2e^{\delta}} \prod_{j=1}^{m} \frac{(\delta/m)^{a_j}}{a_j!} \\ &= \frac{n^{2/3}\delta^2}{4e^{\delta}m^{5/3}}. \end{aligned} \tag{11}$$

By Equations (10) and (11) we have that

$$\begin{aligned} \frac{\left(\vec{\lambda}^-(\vec{a}) - \vec{\lambda}^+(\vec{a})\right)^2}{\vec{\lambda}^-(\vec{a})} &\leq \frac{e^{\delta - 2\delta(m/n)^{1/3}}\delta^2}{4m} \\ &\leq \frac{\delta^2}{m} \ . \end{aligned} \tag{12}$$

By Equation (12) and the fact that $|A_2| = m$ we get

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

The lemma follows by applying Lemma 2. □

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. The proofs of the next lemmas appear in Appendix C.

**Lemma 7** *For $m \geq 16$, $n \geq cm \ln m$ (where $c$ is a sufficiently large constant) and for $\delta \leq 1/16$*

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

**Lemma 8** *For $n \geq m$, $m \geq 12$ and $\delta \leq 1/4$,*

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

We finally turn to the contribution of $\vec{a} \in A_x$ such that $x \geq 3$, where the proof of the next lemma is similar to the proof of Lemma 8.

**Lemma 9** *For $n \geq m$ and $\delta \leq 1/4$,*

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

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

**Proof of Theorem 1:** Let $\mathcal{D}^+$ and $\mathcal{D}^-$ be as defined in Equations (5) and (6), respectively, and recall that $\kappa = \delta \cdot \frac{n^{2/3}}{m^{2/3}}$ (where $\delta$ will be set subsequently). By the definition of the distributions in $\mathcal{D}^+$ and $\mathcal{D}^-$, the probability weight assigned to each element is at most $\frac{1}{n^{2/3}m^{1/3}} = \frac{\delta}{\kappa \cdot m}$, as required by Theorem 2. By Lemma 4, $\mathcal{D}^-$ is (1/20)-far from $\mathcal{P}^{\mathrm{eq}}$. Therefore, it remains to establish that Equation (4) holds for $\mathcal{D}^+$ and $\mathcal{D}^-$. Consider the following partition of $\mathbb{N}^m$:

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

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

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

For $\delta < 1/16$ we get by Lemmas 5–9 that

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

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

## References


[1] M. Adamaszek, A. Czumaj, and C. Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 56-65, 2010.

[2] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing $k$-wise and almost $k$-wise independence. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 496-505, 2007.

[3] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. *SIAM Journal on Discrete Math*, 16(3): 393-417, 2003.

[4] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58, 1999.

[5] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley, New York, 1992.

[6] A. Andoni, P. Indyk, K. Onak, and R. Rubinfeld. External sampling. In *Automata, Languages and Programming: Thirty-Sixth International Colloquium (ICALP)*, pages 83-94, 2009.

[7] K. D. Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld. Sublinear time algorithms for earth mover's distance. In *CoRR abs/0904.0292*, 2009.

[8] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Proceedings of RANDOM*, 2002.

[9] T. Batu. *Testing properties of distributions*. PhD thesis, Computer Science department, Cornell University, 2001.

[10] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132-150, 2005.

[11] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science (FOCS)*, pages 442-451, 2001.

[12] T. Batu, L. Fortnow, R. Rubinfeld, W. Smith, and P. White. Testing that distributions are close. In *Proceedings of the Forty-First Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259-269, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *CoRR*, abs/1009.5397, 2010. This is a long version of [12].

[14] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 381-390, 2004.

[15] V. Braverman and R. Ostrovsky. Measuring $k$-wise independence of streaming data. In *CoRR abs/0806.4790*, 2008.

[16] V. Braverman and R. Ostrovsky. Measuring independence of datasets. In *Proceedings of the Fourty-Second Annual ACM Symposium on the Theory of Computing (STOC)*, pages 271-280, 2010.

[17] V. Braverman and R. Ostrovsky. Zero-one frequency laws. In *Proceedings of the Fourty-Second Annual ACM Symposium on the Theory of Computing (STOC)*, pages 281-290, 2010.

[18] D. Coppersmith and R. Kumar. An improved data stream algorithm for frequency moments. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 151-156, 2004.

[19] G. Cormode, M.Datar, P. Indyk, and S. Muthukrishnan. Comparing data stream using hamming norms (how to zero in). *IEEE Trans. Knowl. Data Eng.*, 15(3): 529-540, 2003.

[20] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299-318, 1967.

[21] A. Czumaj and C. Sohler. Testing expansion in bounded-degree graphs. In *Proceedings of the Forty-Eighth Annual Symposium on Foundations of Computer Science (FOCS)*, pages 570-578, 2007.

[22] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate $L^1$ -difference algorithm for massive data streams (extended abstract). In *Proceedings of the Fortieth Annual Symposium on Foundations of Computer Science (FOCS)*, 1999.

[23] W. Feller. *An introduction to probability theory and its applications/William Feller.* Wiley, New York; Sydney :, 3rd ed. edition, 1967.

[24] E. Fischer and A. Matsliah. Testing graph isomorphism. *SIAM Journal on Computing*, 38(1): 207-225, 2008.

[25] J. Fong and M. Strauss. An approximate $L^p$-difference algorithm for massive data streams. In *Annual Symposium on Theoretical Aspects of Computer Science*, 2000.

[26] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloqium on Computational Complexity*, 7(20), 2000.

[27] S. Guha, A. McGregor, and S. Venkatasubramanian. Sub-linear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5, 2009.

[28] B. Harris. The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis János Bolyai*, 16:323-355, 1975. Topics in Information Theory.

[29] P. Indyk and A. McGregor. Declaring independence via the sketching of sketches. In *Proceedings of the Ninteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 737-745, 2008.

[30] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai. Extracting correlations. In *Proceedings of the Fiftieth Annual Symposium on Foundations of Computer Science (FOCS)*, pages 261-270, 2009.

[31] S. Kale and C. Seshadhri. Testing expansion in bounded degree graphs. In *Automata, Languages and Programming: Thirty-Fifth International Colloquium (ICALP)*, pages 527-538, 2008. A preliminary version appeared in ECCC, TR07-076.

[32] D. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison Wesley, Phillipines, 1969.

[33] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. Technical Report TR10-157, Electronic Colloquium on Computational Complexity (ECCC), 2010.

[34] S. Ma. Calculation of entropy from data of motion. *J. of Statistical Physics*, 26(2): 221-240, 1981.

[35] D. S. Mitrinovic and P. M. Vasic. *Analytic inequalities/D.S. Mitrinovi; in cooperation with P.M. Vasic.* Springer-Verlag, Berlin; New York:, 1970.

[36] A. Nachmias and A. Shapira. Testing the expansion of a graph. Technical Report TR07-118, Electronic Colloquium on Computational Complexity (ECCC), 2007.

[37] I. Nemenman, W. Bialek, and R. Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E*, 69(056111), 2004.

[38] L. Paninski. Estimation of information-theoretic quantities and discrete distributions. *Neural Computation*, 15: 1191-1254, 2003.

[39] L. Paninski. Estimating entropy on bins given fewer than $m$ samples. *IEEE Transactions on Information Theory*, 50(9): 2200-2203, 2004.

[40] L. Paninski. Testing for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750-4755, 2008.

[41] S. Raskhodnikova, D. Ron, R. Rubinfeld, and A. Smith. Sublinear algorithms for approximating string compressibility. In *Proceedings of the Eleventh International Workshop on Randomization and Computation (RANDOM)*, pages 609-623, 2007.

[42] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bonds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3): 813-842, 2009.

[43] B. Roos. On the rate of multivariate poisson convergence. *J. Multivar. Anal.*, 69(1):120-134, 1999.

[44] R. Rubinfeld and R. Servedio. Testing monotone high-dimensional distributions. Manuscript, 2004.

[45] R. Rubinfeld and N. Xie. Testing non-uniform k-wise independent distributions over product spaces. In *Automata, Languages and Programming: Thirty-Seventh International Colloquium (ICALP)*, pages 565-581, 2010.

[46] S. P. Strong, R. Koberle, R. Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197-200, 1998.

[47] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley Sons, Inc., New York, 2001.

[48] P. Valiant. Testing symmetric properties of distributions. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 383-392, 2008.

[49] P. Valiant. *Testing symmetric properties of distributions*. PhD thesis, CSAIL, MIT, 2008.

[50] D. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. Part I. Bayes estimators and the Shannon entropy. *Physical Review E*, 52(6):6841-6854, 1995.

[51] K. Yamanishi. Probably almost discriminative learning. *Machine Learning*, 18(1): 23-50, 1995.

## A Missing details for Subsection 3.1

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

$$\begin{aligned}
&\left|\mathrm{poi}(\vec{i};\, \vec{\lambda}^{+}) - \mathrm{poi}(\vec{i};\, \vec{\lambda}^{-})\right| \\
&= \Big| \prod_{j\in[d]} \mathrm{poi}(i_j;\lambda_j^{+}) - \prod_{j\in[d]} \mathrm{poi}(i_j;\lambda_j^{-})\Big| \\
&\le \Big| \prod_{j\in[d]} \mathrm{poi}(i_j;\lambda_j^{+}) \\
&\quad - \prod_{j\in[d]\setminus I_k} \mathrm{poi}(i_j;\lambda_j^{+}) \prod_{j\in I_k} \mathrm{poi}(i_j;\lambda_j^{-})\Big| \\
&\quad + \Big| \prod_{j\in[d]\setminus I_k} \mathrm{poi}(i_j;\lambda_j^{+}) \prod_{j\in I_k} \mathrm{poi}(i_j;\lambda_j^{-}) \\
&\quad - \prod_{j\in[d]} \mathrm{poi}(i_j;\lambda_j^{-})\Big| .
\end{aligned}$$

Hence, we obtain that

$$\begin{aligned}
&\left\|\mathrm{poi}(\vec{\lambda}^{+}) - \mathrm{poi}(\vec{\lambda}^{-})\right\|_1 \\
&\quad = \sum_{\vec{i}\in\mathbb{N}^d} \left|\mathrm{poi}(\vec{i};\, \vec{\lambda}^{+}) - \mathrm{poi}(\vec{i};\, \vec{\lambda}^{-})\right| \\
&\quad \le \left\|\mathrm{poi}(\vec{\lambda}^{+}(I_k)) - \mathrm{poi}(\vec{\lambda}^{-}(I_k))\right\|_1 \\
&\qquad + \left\|\mathrm{poi}(\vec{\lambda}^{+}([d]\setminus I_k)) - \mathrm{poi}(\vec{\lambda}^{-}([d]\setminus I_k))\right\|_1 .
\end{aligned}$$

Thus, the lemma follows by induction on $\ell$. □

**Proof of Lemm 2:** In order to prove the lemma we shall use the *KL-divergence* between distributions.

Namely, for two distributions $p_1$ and $p_2$ over a domain $X$, $D_{\mathrm{KL}}(p_1\|p_2) \stackrel{\mathrm{def}}{=} \sum_{x\in X} p_1(x)\cdot \ln\frac{p_1(x)}{p_2(x)}$. Let $\vec{\lambda}^+ = (\lambda_1^+\ldots,\lambda_d^+)$, $\vec{\lambda}^- = (\lambda_1^-\ldots,\lambda_d^-)$ and let $\vec{i}$ denote $(i_1,\ldots i_d)$. We have that

$$\begin{aligned}
&\ln\frac{\mathrm{poi}(\vec{i};\,\vec{\lambda}^+)}{\mathrm{poi}(\vec{i};\,\vec{\lambda}^-)}\\
&=\sum_{j=1}^{d}\ln\left(e^{\lambda_j^- - \lambda_j^+}\left(\lambda_j^+/\lambda_j^-\right)^{i_j}\right)\\
&=\sum_{j=1}^{d}\left((\lambda_j^- - \lambda_j^+) + i_j\cdot\ln(\lambda_j^+/\lambda_j^-)\right)\\
&\le\sum_{j=1}^{d}\left((\lambda_j^- - \lambda_j^+) + i_j\cdot(\lambda_j^+/\lambda_j^- - 1)\right),
\end{aligned}$$

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

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mathrm{poi}(\vec{\lambda}^+)\|\mathrm{poi}(\vec{\lambda}^-)\right)\\
&=\sum_{\vec{i}\in\mathbb{N}^d}\mathrm{poi}(\vec{i};\,\vec{\lambda}^+)\cdot\ln\frac{\mathrm{poi}(\vec{i};\,\vec{\lambda}^+)}{\mathrm{poi}(\vec{i};\,\vec{\lambda}^-)}\\
&\le\sum_{j=1}^{d}\left((\lambda_j^- - \lambda_j^+) + \lambda_j^+\cdot(\lambda_j^+/\lambda_j^- - 1)\right) \qquad (13)\\
&=\sum_{j=1}^{d}\frac{(\lambda_j^- - \lambda_j^+)^2}{\lambda_j^-},
\end{aligned}$$

where in Equation (13) 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 $\ell_1$ distance is related to the KL-divergence by $\|D-D'\|_1 \le 2\sqrt{2D_{\mathrm{KL}}\left(D\|D'\right)}$ and thus we obtain the lemma. □

The next lemma will be used in the proof of Theorem 2.

**Lemma 10** *Let $X\sim\mathrm{poi}(\lambda)$, then,*

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

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

$$\begin{aligned}
&\frac{\mathrm{poi}(j+\lambda/2;\lambda)}{\mathrm{poi}(j;\lambda)}\\
&=\frac{e^{-\lambda}\cdot\lambda^{j+\lambda/2}/(j+\lambda/2)!}{e^{-\lambda}\cdot\lambda^{j}/j!}\\
&=\frac{\lambda^{\lambda/2}}{(j+\lambda/2)(j+\lambda/2-1)\cdots(j+1)}\\
&=\frac{\lambda}{j+\lambda/2}\cdot\frac{\lambda}{j+\lambda/2-1}\cdot\cdots\cdot\frac{\lambda}{j+1}\\
&\ge\frac{\lambda}{\lambda-1}\cdot\frac{\lambda}{\lambda-2}\cdot\cdots\cdot\frac{\lambda}{\lambda/2}\\
&>\left(\frac{\lambda}{(3/4)\lambda}\right)^{\lambda/4}\\
&=(4/3)^{\lambda/4}
\end{aligned}$$

This implies that

$$\begin{aligned}
&\Pr[X<\lambda/2]\\
&=\frac{\Pr[X<\lambda/2]}{\Pr[\lambda/2\le X<\lambda]}\cdot\Pr[\lambda/2\le X<\lambda]\\
&<\frac{\Pr[X<\lambda/2]}{\Pr[\lambda/2\le X<\lambda]} < (3/4)^{\lambda/4},
\end{aligned}$$

and the proof is completed. □

## B Missing details for Subsection 3.2

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

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

**Lemma 11** *Assume that there exists a tester $T$ in the uniform sampling model for a property $\mathcal{P}$ of lists of $m$ distributions, that takes a sample of size $s=\kappa m$ where $\kappa \ge c\log m$ for some sufficiently large constant $c$, and works for every $\epsilon\ge\epsilon_0$ where $\epsilon_0$ 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 $4\kappa$, that works for every $\epsilon\ge\epsilon_0$ and whose success probability is at least $\frac{19}{30}$.*

**Proof:** Roughly speaking, the tester $T'$ tries to simulate $T$ if it has a sufficiently large sample, and otherwise it guesses the answer. More precisely, let $\mathcal{D} = (D_1, \ldots, D_m)$ be a list of $m$ distributions. For each $j \in [m]$ let $\kappa_j$ denote the random variable that equals the number of samples that are selected according to $D_j$ in the uniform sampling model, when the total number of samples is $\kappa m$. Thus, $\kappa_j \sim Bin(\kappa m, \frac{1}{m})$. By [5, Thm. A.12], for each $j \in [m]$,

$$\Pr[\kappa_i \geq 2\kappa] < (e/4)^{\kappa} .$$

Now consider a tester $T'$ that receives $\kappa'_j$ samples from each $D_j$ where $\kappa'_j \sim \text{poi}(4\kappa)$. By Lemma (10), for each $j$ we have that,

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

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

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

Given Lemma 11 it suffices to consider samples that are generated in the Poissonized uniform sampling model. The process for generating a sample $\{\alpha_{i,1}, \ldots, \alpha_{i,m}\}_{i \in [n]}$ (recall that $\alpha_{i,j}$ is the number of times that element $i$ was selected by distribution $D_j$) in the $\kappa$-Poissonized model is equivalent to the following process: For each $i \in [n]$ and $j \in [m]$, independently select $\alpha_{i,j}$ according to $\text{poi}(\kappa \cdot D_j(i))$ (see [23], p. 216). Thus the probability of getting a particular histogram $\vec{a}_i = (a_{i,1}, \ldots, a_{i,m})$ for element $i$ is $p^{\mathcal{D},\kappa}(i; \vec{a}_i)$ (as defined in Equation (2)). We can represent the event that the histogram of element $i$ is $\vec{a}_i$ by a Bernoulli random vector $\vec{b}_i$ that is indexed by all $\vec{a} \in \mathbb{N}^m$, is 1 in the coordinate corresponding to $\vec{a}_i$, and is 0 elsewhere. Given this representation, the fingerprint of the sample corresponds to $\sum_{i=1}^{n} \vec{b}_i$. In fact, we would like $\vec{b}_i$ to be of finite dimension, so we have to consider only a finite number (sufficiently large) of possible histograms. Under this relaxation, $\vec{b}_i = (0, \ldots, 0)$ would correspond to the case that the sample histogram of element $i$ is not in the set of histograms we consider. 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 $\vec{b}_i$ that are indexed directly, that is $\vec{b}_i = (b_{i,1}, \ldots, b_{i,h})$.

**Theorem 3 ([43])** *Let $D^{S_n}$ be the distribution of the sum $S_n$ of $n$ independent Bernoulli random vectors $\vec{b}_1, \ldots, \vec{b}_n$ in $\mathbb{R}^h$ where $\Pr\left[\vec{b}_i = \vec{e}_\ell\right] = p_{i,\ell}$ and $\Pr\left[\vec{b}_i = (0, \ldots, 0)\right] = 1 - \sum_{\ell=1}^{h} p_{i,\ell}$ (here $\vec{e}_\ell$ satisfies $e_{i,\ell} = 1$ and $e_{i,\ell'} = 0$ for every $\ell' \neq \ell$). Suppose we define an $h$-dimensional vector $\vec{\lambda} = (\lambda_1, \ldots, \lambda_h)$ as follows: $\lambda_\ell = \sum_{i=1}^{n} p_{i,\ell}$. Then*

$$\left\| D^{S_n} - \text{poi}(\vec{\lambda}) \right\|_1 \leq \frac{88}{5} \sum_{\ell=1}^{h} \frac{\sum_{i=1}^{n} p_{i,\ell}^2}{\sum_{i=1}^{n} p_{i,\ell}} . \tag{14}$$

We next show how to obtain a bound on sums of the form given in Equation (14) under appropriate conditions.

**Lemma 12** *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 $i \in [n]$ and for all $j \in [m]$, $D_j(i) \leq \frac{\delta}{m \cdot \kappa}$ for some integer $\kappa$, we have that*

$$\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{i=1}^{n} p^{\mathcal{D},\kappa}(i; \vec{a})^2}{\sum_{i=1}^{n} p^{\mathcal{D},\kappa}(i; \vec{a})} \leq 2\delta . \tag{15}$$

**Proof:**

$$\begin{aligned}
&\sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \frac{\sum_{i=1}^{n} p^{\mathcal{D},\kappa}(i; \vec{a})^2}{\sum_{i=1}^{n} p^{\mathcal{D},\kappa}(i; \vec{a})} \\
&\leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_i \left( p^{\mathcal{D}}(i; \vec{a}) \right) \\
&= \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \max_i \left( \prod_{j=1}^{m} \text{poi}(a_j; \kappa \cdot D_j(i)) \right) \qquad (16) \\
&\leq \sum_{\vec{a} \in \mathbb{N}^m \setminus \vec{0}} \left( \frac{\delta}{m} \right)^{a_1 + \cdots + a_m} \\
&\leq \sum_{a=1}^{\infty} m^a \left( \frac{\delta}{m} \right)^{a} \\
&\leq 2\delta ,
\end{aligned}$$

where the inequality in Equation (16) holds for $\delta \leq$

1/2 and the inequality in Equation (16) follows from:

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

and the proof is completed. □

**Proof of Theorem 2:** By the first premise of the theorem, $D_j^+(i), D_j^+(i) \leq \frac{\delta}{\kappa m}$ for every $i \in [n]$ and $j \in [m]$. By Lemma 12 this implies that Equation (15) holds both for $\mathcal{D} = \mathcal{D}^+$ and for $\mathcal{D} = \mathcal{D}^-$. Combining this with Theorem 3 we get that the $\ell_1$ distance between the fingerprint distribution when the sample is generated according to $\mathcal{D}^+$ (in the $\kappa$-Poissonized model, see Definition 1) and the distribution $\mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^+,\kappa}\right)$ 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 (4) (concerning the $\ell_1$ distance between $\mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^+,\kappa}\right)$ and $\mathrm{poi}\left(\vec{\lambda}^{\mathcal{D}^-,\kappa}\right)$) and the triangle inequality, we get that the $\ell_1$ distance between the two fingerprint distributions is smaller than $2\cdot\frac{176}{5}\delta + \frac{16}{30} - \frac{352\delta}{5} = \frac{16}{30}$, which implies that the statistical difference is smaller than $\frac{8}{30}$, and thus it is not possible to distinguish between $\mathcal{D}^+$ and $\mathcal{D}^-$ in the $\kappa$-Poissonized model with success probability at least $\frac{19}{30}$. By Lemma 11 we get the desired result. □

## C Missing details for Subsection 3.3

**Proof of Lemma 3:** Recall that we consider selecting a matrix $M$ randomly as follows: Denote the first $t/2$ rows of $M$ by $F$. For each row in $F$, pick, independently from the other $t/2-1$ rows in $F$, a random half of its elements to be 1, and the other half of the elements to be 0. Rows $t/2+1,\ldots,t$ are the negations of rows $1,\ldots,t/2$, respectively. Thus, in each row and each column of $M$, exactly half of the elements are 1 and the other half are 0.

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

$$\begin{aligned}&\Pr[I_{S,i,b}=1]\\ &= \frac{1}{2}\cdot\left(\frac{1}{2}-\frac{1}{m}\right)\cdot\ \cdots\ \cdot\left(\frac{1}{2}-\frac{x-1}{m}\right) .\end{aligned}$$

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

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

where the last inequality is due to Bernoulli's inequality which states that $(1+x)^n > 1+nx$, for every real number $x > -1 \neq 0$ and an integer $n > 1$ ([35]).

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

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

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

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

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

**Proof of Lemma 4:** Consider any $\vec{a} \in S_2$. By Lemma 3, setting $t = n/2$, the size of $I_M(\vec{a})$, i.e. the number of light elements $\ell$ such that $D_j^-[\ell] = \frac{2}{n}$ for every $j \in \mathrm{sup}(\vec{a})$, is at most $\frac{n}{2}\left(\frac{1}{4}+\sqrt{\frac{8\ln m}{n}}\right)$. The same lower bound holds for the number of light elements $\ell$ such that $D_j^-[\ell] = 0$ for every $j \in \mathrm{sup}(\vec{a})$. This implies that for every $j \neq j'$ in $[m]$, for at least $\frac{n}{2} - n\left(\frac{1}{4}+\sqrt{\frac{8\ln m}{n}}\right)$ of the light elements, $\ell$, we have that $D_j^-[\ell] = \frac{2}{n}$ while $D_{j'}^-[\ell] = 0$, or that $D_{j'}^-[\ell] = \frac{2}{n}$ while $D_j^-[\ell] = 0$. Therefore, $\|D_j^- - D_{j'}^-\|_1 \geq \frac{1}{2} - 2\sqrt{\frac{8\ln m}{n}}$, which for

$n \geq c \ln m$ and a sufficiently large constant $c$, is at least $\frac{1}{8}$. Thus, by the triangle inequality we have that for every $D^*$, $\sum_{j=1}^{m} \|D_j^- - D^*\|_1 \geq \lfloor \frac{m}{2} \rfloor \cdot \frac{1}{8}$,

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

**Proof of Lemma 9:** 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 (7) and (8) (and the fact that $|x-y| \leq \max\{x,y\}$ for every positive real numbers $x$,$y$),

$$\begin{aligned}
&\sum_{x\geq 3} \sum_{\vec{a}\in A_x} \left| \vec{\lambda}^+(\vec{a}) - \vec{\lambda}^-(\vec{a}) \right| \\
&\leq \sum_{x\geq 3} \sum_{\vec{a}\in A_x} \frac{n}{2} \prod_{j=1}^{m} \left( \frac{2\delta}{n^{1/3} m^{2/3}} \right)^{a_j} \\
&= \sum_{x\geq 3} \sum_{\vec{a}\in A_x} \frac{n}{2} \left( \frac{2\delta}{n^{1/3} m^{2/3}} \right)^{\sum_{j=1}^{m} a_j} \\
&\leq \sum_{x=3}^{\infty} m^{x-1} \cdot \frac{n}{2} \left( \frac{2\delta}{n^{1/3} m^{2/3}} \right)^{x} \\
&= \frac{n}{2m} \sum_{x=3}^{\infty} \left( \frac{2\delta m^{1/3}}{n^{1/3}} \right)^{x} \\
&= 4\delta^3 \sum_{x=0}^{\infty} \left( \frac{2\delta m^{1/3}}{n^{1/3}} \right)^{x} \\
&\leq \frac{4\delta^3}{1-2\delta} \qquad (17) \\
&\leq 8\delta^3 \qquad (18)
\end{aligned}$$

where in Equation (17) we used the fact that $n \geq m$ and Equation (18) holds for $\delta \leq 1/4$. The lemma follows by applying Equation (1). □