Probability–Revealing Samples

Krzysztof Onak
IBM Research

Xiaorui Sun
Microsoft Research

Abstract

In the most popular distribution testing and parameter estimation model, one can obtain information about an underlying distribution $\mathcal{D}$ via independent samples from $\mathcal{D}$. We introduce a model in which every sample comes with the information about the probability of selecting it. In this setting, we give algorithms for problems such as testing if two distributions are (approximately) identical, estimating the total variation distance between distributions, and estimating the support size. The sample complexity of all of our algorithms is optimal up to a constant factor for sufficiently large support size. The running times of our algorithms are near-linear in the number of samples collected. Additionally, our algorithms are robust to small multiplicative errors in probability estimates.

The complexity of our model lies strictly between the complexity of the model where only independent samples are provided and the complexity of the model where additionally arbitrary probability queries are allowed.

Our model finds applications where once a given element is sampled, it is easier to estimate its probability. We describe two scenarios in which all occurrences of each element are easy to explore once at least one copy of the element is detected.

1 INTRODUCTION

Testing properties and estimating parameters of probability distributions have been research directions in statistics and probability theory. Two seminal papers of Batu et. al [3, 2] instigated the modern research on these questions through the theoretical computer science lens.

In the most popular setting, an algorithm has sample access to one or more discrete distributions on $[n]$.[1] For a given distribution $\mathcal{D}$, it can obtain an arbitrarily long sequence of independent samples from $\mathcal{D}$. Over the last two decades, a number of distribution testing and parameter estimation problems have been considered in this model (e.g., [3, 2, 11, 12, 13, 15, 8]).

In the big data age, even the sample-efficient algorithms in the model described above may become inefficient. It makes therefore sense to investigate the question whether there exist more efficient algorithms, provided some additional information about distributions is available. This kind of question has been investigated in a few papers [1, 7, 5, 6], which allowed for various types of additional queries.

Probability–Revealing Samples. In this paper, we propose a new sample model for property testing and parameter estimation problems. In our model, instead of simply receiving an independent sample x from $\mathcal{D}$, an algorithm receives a pair (x, p_x), where p_x is the probability of selecting x from $\mathcal{D}$. We call such a pair a *probability–revealing sample* (or a *PR-sample* in short).

Additionally, we consider a more relaxed version of the model, where the probability estimate may be off by a small multiplicative factor. In this case, a pair (x, p'_x) is an *α-approximate probability–revealing sample* (or in short, an *α-PR-sample*) from a distribution $\mathcal{D}$ if x is selected from the distribution $\mathcal{D}$ and $p_x/\alpha \leq p'_x \leq \alpha p_x$, where p_x is the probability of x in $\mathcal{D}$ and $\alpha \geq 1$.

We sometimes refer to PR-samples as *exact PR-samples* when we want to stress that no approximation of probabilities is allowed. Conversly, when α can be larger than 1, but its exact value is either not important or is clear from the context, we may refer to α-PR-samples as *approximate PR-samples*.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

[1] We write $[n]$ to denote $\{1, \ldots, n\}$.

Problems. In this paper, we assume that all distributions are on $[n]$, where n is a parameter known to the algorithm. We consider the following problems, where d_{TV} denotes the total variation distance between probability distributions.

- **Identity Testing:** Accept if an unknown distribution $\mathcal{D}$ equals a known distribution $\mathcal{D}_\star$ and reject if $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_\star) > \epsilon$ for $\epsilon \in (0, 1)$.
- **Distance to a Known Distribution:** For an unknown distribution $\mathcal{D}$ and a known distribution $\mathcal{D}_\star$, approximate $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_\star)$ up to an additive ϵ for $\epsilon \in (0, 1)$.
- **Equality Testing:** Accept if two unknown distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are identical and reject if $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) > \epsilon$ for $\epsilon \in (0, 1)$.
- **Distance between Unknown Distributions:** For two unknown distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, approximate $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)$ up to an additive ϵ for $\epsilon \in (0, 1)$.
- **Distinct Elements:** For a positive integer $n > 0$, an unknown distribution $\mathcal{D}$ such that every element with non-zero probability has probability at least $1/n$, and a parameter $\epsilon \in (0, 1)$, approximate the number of elements with non-zero probability in $\mathcal{D}$ up to an additive ϵn.

In this work, we only require that for any input distribution or pair of distributions, the algorithm output a correct answer with probability 2/3. It follows directly from standard concentration bounds, such as the Chernoff bound, that it suffices to run the algorithm $O(\log(1/\delta))$ times—and take the median or majority of computed solutions—to improve the success probability to $1 - \delta$ for any $\delta \in (0, 1/3)$.

As our main technical results of this paper, we present algorithms for the above five problems in the probability–revealing sample model. The query complexity of them is optimal up to constant factors. Additionally, their input distributions can be on an arbitrary set of size at most n, not just $[n]$.

Theorem 1.1 (Complexity of Identity Testing)**.** *There is a polynomial time algorithm to solve the Identity Testing problem using $O(1/\epsilon)$ $(1 + \epsilon/3)$-PR-samples from $\mathcal{D}$. If $n \geq 3$, then any algorithm to solve the Identity Testing problem requires at least $\Omega(1/\epsilon)$ PR-samples.*

Theorem 1.2 (Complexity of Distance to a Known Distribution)**.** *There is a polynomial time algorithm to solve the Distance to Known Distribution problem using $O(1/\epsilon^2)$ $(1 + \epsilon/3)$-PR-samples from $\mathcal{D}$. If $\epsilon < 1/10000$ and $n = \omega(1/\epsilon^4)$, then any algorithm to solve the Distance to Known Distribution problem requires at least $\Omega(1/\epsilon^2)$ PR-samples.*

Theorem 1.3 (Complexity of Equality Testing and Distance between Unknown Distributions)**.** *For $\epsilon \in (0, 1)$, there is a polynomial time algorithm to solve Equality Testing problem and Distance between Unknown Distributions problem using $O(\max\{\sqrt{n}/\epsilon, 1/\epsilon^2\})$ $(1 + \epsilon/3)$-PR-samples. For $\epsilon < 1/100$ and $n = \omega(1/\epsilon^{10})$, any algorithm to solve the two problems requires at least $\Omega(\sqrt{n}/\epsilon)$ PR-samples.*

Theorem 1.4 (Complexity of Distinct Elements)**.** *For $\epsilon \in (0, 1)$, there is a polynomial time algorithm to solve the Distinct Elements problem using $O(1/\epsilon^2)$ $(1+\epsilon/3)$-PR-samples. If $\epsilon < 1/10000$ and $n = \omega(1/\epsilon^4)$, then any algorithm to solve the Distinct Elements problems requires at least $\Omega(1/\epsilon^2)$ PR-samples.*

Comparison to other models and related works. In the standard distribution testing model, algorithms are given access to independent samples from the unknown distributions. The five problems considered in this paper have been studied extensively in this model. See the column "Samples" of Table 1 for a summary of known results.

At least two papers [1, 6] consider a dual model, which is stronger than ours. In their model, for each distribution, apart from drawing independent samples, one can also query the probability of each element. Clearly, one can simulate PR-sampling in their model by first drawing a sample and then querying its probability. The complexity of problems in their model is presented in the column "Dual." Contrary to their paper, the algorithms we present here are robust to small multiplicative errors in probability estimates.

One can observe that the complexity of our model falls in between these two models. In particular, we avoid the lower bounds of Valiant [15], by being able to easily distinguish heavy elements from light elements. This difficulty in the standard model was at heart of his lower bound constructions.

Techniques. Our algorithm are very simple and are based on evaluating very simple estimators. We apply Hoeffding's inequality or Chebyshev's inequality to show the proper concentration of the estimate output by each algorithm.

For lower bounds we construct two distributions with different desired behavior of the algorithm and show that a large number of samples is required to distinguish them. Some of our techniques are similar to those of Canetti, Even, and Goldreich [4].

Due to the space constraints, we omit some of our proofs and make the full paper available on arXiv.

Table 1: Comparison Between Different Models

	Samples	PR-Samples [here]	Dual [6]
Identity Testing	$\Theta\left(\frac{n^{1/2}}{\epsilon^2}\right)$ [2, 11, 14]	$\Theta(1/\epsilon)$ for $n \geq 3$ 1 for $n = 2$	$\Theta(1/\epsilon)$
Distance to Known	$\Omega\left(\frac{n^{1/2}}{\epsilon^2}\right)$ [11]	$O(1/\epsilon^2)$ $\Omega(\frac{1}{\epsilon^2})$ for $n = \omega(\epsilon^{-4})$	$\Theta(1/\epsilon^2)$
Equality Testing	$\Theta(\max\{\frac{n^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{n}}{\epsilon^2}\})$ [3, 8]	$O(\max\{\frac{n^{1/2}}{\epsilon}, \frac{1}{\epsilon^2}\})$ $\Omega(\frac{n^{1/2}}{\epsilon})$ for $n = \omega(\epsilon^{-10})$	$\Theta(1/\epsilon)$
Distance between Unknown	$\Theta(\frac{n}{\log n})$ for $\epsilon = \Theta(1)$ [13]	$O(\max\{\frac{n^{1/2}}{\epsilon}, \frac{1}{\epsilon^2}\})$ $\Omega(\frac{n^{1/2}}{\epsilon})$ for $n = \omega(\epsilon^{-10})$	$\Theta(1/\epsilon^2)$
Distinct Elements	$\Theta(\frac{n}{\log n})$ for $\epsilon = \Theta(1)$ [12, 13]	$\Theta(1/\epsilon^2)$ $\Omega(\frac{1}{\epsilon^2})$ for $n = \omega(\epsilon^{-4})$	$\Theta(1/\epsilon^2)$

2 APPLICATIONS

We now describe two applications of our model. In both cases, we take advantage of the fact that once a given element is detected by sampling, it is easier to compute its probability.

2.1 Grouped Identical Labels

Consider a set of records stored in an array $T[1 \dots n]$, organized in such a way that records with the same label appear in consecutive entries of the array (see Figure 1a). This could be a result of collecting the records from a hash table with all elements in each entry of the hash table sorted according to their label.

A PR-sample can be generated in two steps. First, we select a random entry $T[i]$ in the array. Clearly, the label of $T[i]$ is selected from the empirical distribution of labels in the array. Second, in order to compute the probability of selecting this label, we use binary search to determine the first index $j_0 \leq i$ and the last index $j_1 \geq i$ such that all entries $T[i']$ for $j_0 \leq i' \leq j_1$ have the same label as $T[i]$. The probability of selecting this label is $(j_1 - j_0 + 1)/n$. Therefore, generating a PR-sample requires no more than $O(\log n)$ queries to the array.

All algorithms that we design for PR-samples can be applied to this setting. Their query complexity becomes the sample complexity of the original PR-sample algorithm times an additional factor of $O(\log n)$.

2.2 Geographic Distributions

The setting we consider now models the following situation where the area on which a given element of the distribution occurs induces a connected component in an underlying graph. In the simplest version, we have a $\sqrt{n} \times \sqrt{n}$ chessboard and each square of the board is occupied by a single element in $[n]$. If two squares (x, y) and (x', y') are occupied by the same element, there is a sequence $(x_0, y_0) = (x, y), (x_1, y_1), \dots, (x_k, y_k) = (x', y')$ of squares connecting them. More precisely, any pair of consecutive squares (x_i, y_i) and (x_{i+1}, y_{i+1}) (where $0 \leq i < k$) in the sequence is adjacent (i.e., $|x_i - x_{i+1}| + |y_i - y_{i+1}| = 1$). See Figure 1b for an example.

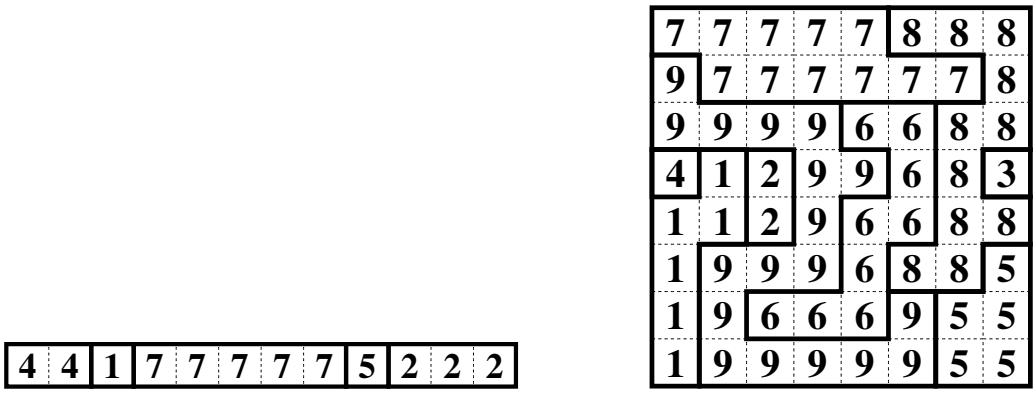
In this case, the probability of an element occupying a given square can be computed exactly in time proportional to the area occupied by the element. It suffices to run BFS exploration that moves only between adjacent squares containing the same element. We now sketch an algorithm for the Distance between Unknown Distributions problem in this setting.[2] It is based on our algorithm that uses approximate PR-samples. The algorithm learns approximate probabilities of heavy elements by sampling and employs BFS exploration for light elements.

Distance between Unknown Distributions. We first assume that $n > 1/\epsilon^2$. Let

$$t = n^{1/4}\sqrt{\epsilon^{-1}\log(1+n)}$$

be a threshold value. We want to estimate the probabilities of all elements that occupy approximately t or more squares. To this end, the algorithm samples

[2] We note that for Distinct Elements, an algorithm due to Chazelle, Rubinfeld, and Trevisan [9] solves the problem with $\tilde{O}((1/\epsilon)^2)$ queries. For Equality Testing, the standard $O(n^{2/3}/\epsilon^2)$-sample algorithm is better than the bound for Distance between Unknown Distributions that we provide here.



(a) Grouped Identical Labels (b) A Geographic Distribution

Figure 1: Sample Applications

$O\left(\frac{n}{t}\cdot\epsilon^{-2}\cdot\log(1+n)\right) = O\left(n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}\right)$ squares. Let S_1 and S_2 be the sets of elements that have probability at least t/n and $t/(2n)$, respectively. Routinely applying the Chernoff and union bounds, with high probability, the sample suffices to find a set $S_\star$ such that $S_1 \subseteq S_\star \subseteq S_2$ and additionally, it provides a multiplicative $(1+\epsilon/3)$-approximation to the probability of elements in $S_\star$. For each element not in $S_\star$, its exact probability can be computed using BFS exploration with at most $O(t)$ queries. We can therefore apply our Distance Estimation algorithm that uses approximate PR-samples. Since it uses $O\left(\max\left\{\sqrt{n}/\epsilon, 1/\epsilon^2\right\}\right)$ $(1+\epsilon/3)$-PR-samples, the query complexity of this step is bounded by $O\left(t\cdot\max\left\{\sqrt{n}/\epsilon, 1/\epsilon^2\right\}\right) = O(t\cdot\sqrt{n}/\epsilon) = O\left(n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}\right)$, where the first transition follows from the assumption that $n > 1/\epsilon^2$. Hence, the query complexity of both steps of the algorithm is bounded by $O\left(n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}\right)$.

Additionally, note that when the number of queries in the procedure described above is $\Omega(n)$ or when $n \leq 1/\epsilon^2$ (in this case $n = O\left(n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}\right)$), one can simply query all squares and compare distributions directly. This way we obtain an algorithm that works for any n and makes $O(\min\{n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}, n\})$ queries.

General Bounded-Degree Graphs. The $\sqrt{n}\times\sqrt{n}$ chessboard example above easily extends to the setting where the elements occupy nodes of a graph with maximum degree bounded by a constant. We say that a distribution of elements is *geographic* if the subgraph induced by each element is connected. As before, once we detect a specific element, its probability can be computed in $O(\bar{n}\cdot d)$ time, using BFS, where $\bar{n}$ is the number of nodes occupied by the element and $d > 0$ is a bound on the maximum degree. If $d = O(1)$, our algorithm for the Distance between Unknown Distributions problem still requires at most $O(\min\{n^{3/4}\epsilon^{-3/2}\sqrt{\log(1+n)}, n\})$ queries.

We leave it as an open question whether there is an algorithm $\mathcal{A}$ and a constant $\delta > 0$ such that for any fixed $\epsilon > 0$, $\mathcal{A}$ uses $O(n^{3/4-\delta})$ queries.

3 PRELIMINARIES

For simplicity, we assume that all our distributions are on a finite discrete domain $[n] = \{1, 2, \ldots, n\}$. Since our algorithms make no use of labels other than comparing them, this does not limit their applicability to a more general set of labels.

For a discrete distribution $\mathcal{D}$ on $[n]$, we write $\mathcal{D}[x]$ for $x \in [n]$ to denote the probability of drawing x from $\mathcal{D}$. For a set $S \subset [n]$, we write $\mathcal{D}[S]$ to denote the probability of drawing an element of S from $\mathcal{D}$.

Throughout this paper, we use the total variation distance to measure the distance between two distributions:

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \max_{S\subseteq[n]} (\mathcal{D}_1[S] - \mathcal{D}_2[s]) \\ &= \frac{1}{2}\sum_{x\in[n]} |\mathcal{D}_1[x] - \mathcal{D}_2[x]| \\ &= 1 - \sum_{x\in[n]} \min\{\mathcal{D}_1[x], \mathcal{D}_2[x]\}, \end{aligned}$$

where it is easy to verify that the three expressions of it are equivalent. In particular, we refer to $\sum_{x\in[n]} \min\{\mathcal{D}_1[x], \mathcal{D}_2[x]\}$ as the *common probability mass* of $\mathcal{D}_1$ and $\mathcal{D}_2$

We make use of both the Chernoff bound and Hoeffding's inequality.

Algorithm 1: Testing if an unknown distribution $\mathcal{D}$ is a given distribution $\mathcal{D}_\star$ with parameter $\epsilon \in (0,1)$

Let $k = \lceil 6/\epsilon \rceil$.
Collect k independent $(1+\epsilon/3)$-PR-samples (x_1, p_1), (x_2, p_2), $\ldots$, (x_k, p_k) from $\mathcal{D}$.
if $\mathcal{D}_\star[x_i] \cdot (1+\epsilon/3) < p_i$ for some $i \in [k]$ **then reject**
accept

Theorem 3.1 (Chernoff bound [10]). *Let $X = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are n independent Bernoulli random variables such that $\Pr[X_i] = p_i$. Let $\mu = E[X]$. Then, for any $0 < \delta < 1$, we have*

$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}$$

and

$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}.$$

Theorem 3.2 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ are n independent random variables such that $\Pr[a_i \leq X_i \leq b_i] = 1$ for $1 \leq i \leq n$. Let $X = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\mu = E[X]$. Then,*

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

4 IDENTITY TESTING

We now prove Theorem 1.1. We prove the upper bound as Lemma 4.1 and the lower bound as Lemma 4.2.

4.1 Upper Bound

We now show that Algorithm 1 solves the Identity Testing problem. The algorithm collects $O(1/\epsilon)$ PR-samples, searching for an element on which the unknown distribution has a significantly higher probability than the known distribution.

Lemma 4.1. *Let $\mathcal{D}_\star$ be a distribution with the probability of each element known and let $\mathcal{D}$ be a distribution from which approximate PR-samples can be drawn. For any parameter $\epsilon \in (0,1)$, Algorithm 1 collects $O(1/\epsilon)$ independent $(1+\epsilon/3)$-PR-samples from $\mathcal{D}$. If $\mathcal{D} = \mathcal{D}_\star$, the algorithm accepts with probability 1. If $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_\star) \geq \epsilon$, the algorithm rejects with probability 2/3.*

Proof. The sample complexity of the algorithm is clearly $O(1/\epsilon)$. It is also obvious that it never rejects in the case that $\mathcal{D} = \mathcal{D}_\star$, because in this case, for any $(1+\epsilon/3)$-approximate PR-sample (x, p) from $\mathcal{D}$, it holds

$$\mathcal{D}_\star[x]/(1+\epsilon/3) \leq p \leq \mathcal{D}_\star[x] \cdot (1+\epsilon/3).$$

It remains to prove that if $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_\star) \geq \epsilon$, the algorithm rejects with probability at least 2/3. Let $S = \{x \in [n] : \mathcal{D}_\star[x] \cdot (1+\epsilon/3) < \mathcal{D}[x]/(1+\epsilon/3)\}$. Clearly, if a PR-sample (x, p) such that $x \in S$ is drawn from $\mathcal{D}$ by the algorithm, the algorithm rejects. We claim that $\mathcal{D}[S] \geq \epsilon/3$ if the distance between distributions is at least ϵ. Suppose for contradiction that $\mathcal{D}[S] < \epsilon/3$. Then the common probability mass of $\mathcal{D}$ and $\mathcal{D}_\star$ is

$$\begin{aligned}
\sum_{x\in[n]} \min\{\mathcal{D}[x], \mathcal{D}_\star[x]\} &\geq \sum_{x\in[n]\setminus S} \min\{\mathcal{D}[x], \mathcal{D}_\star[x]\} \\
&\geq \sum_{x\in[n]\setminus S} \mathcal{D}[x]/(1+\epsilon/3)^2 \\
&= \mathcal{D}\big[[n] \setminus S\big]/(1+\epsilon/3)^2 \\
&= (1 - \mathcal{D}[S])/(1+\epsilon/3)^2 \\
&> (1-\epsilon/3)/(1+\epsilon/3)^2 \\
&> (1-\epsilon/3)^3 > 1 - \epsilon.
\end{aligned}$$

This implies that the distance between distributions is less than ϵ, which contradicts the assumption that the distributions are at distance at least ϵ. We therefore have $\mathcal{D}[S] \geq \epsilon/3$. The probability that the algorithm selects no PR-sample (x, p) from $\mathcal{D}$ such that $x \in S$ (in which case it rejects) is at most

$$(1-\epsilon/3)^{\lceil 6/\epsilon \rceil} \leq (1-\epsilon/3)^{6/\epsilon} \leq e^{-2} < 1/3,$$

which implies that the algorithm rejects with probability at least 2/3. □

4.2 Lower Bound

Intuitively, our lower bound is based on the observation that if two distributions differ on a single element of probability ϵ (in which case the algorithm is expected to reject), it takes $\Omega(1/\epsilon)$ samples to detect the element they differ at. Before that happens the algorithm does not know whether the distributions are identical or their total variation distance is ϵ.

Lemma 4.2. *If $n \geq 3$, then the Identity Testing problem requires $\Omega(1/\epsilon)$ PR-samples if an algorithm has to succeed with probability at least 2/3.*

Proof. We define distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ on $[n]$. The probability of drawing 1 in each of them is $1 - \epsilon$. For $\mathcal{D}_0$, the probability of drawing 2 equals ϵ and the probability of drawing any integer greater than 2 is 0. For $\mathcal{D}_1$, the probability of drawing 3 is ϵ and the probability of drawing 2 or any integer greater than 3 is 0. The total variation distance between the distributions is exactly ϵ.

Let $\mathcal{D}_\star = \mathcal{D}_0$ be the known distribution. We select the unknown distribution $\mathcal{D}$ uniformly at random from $\{\mathcal{D}_0, \mathcal{D}_1\}$. The algorithm's goal is to accept if $\mathcal{D} = \mathcal{D}_0$

Algorithm 2: Estimate Total Variation Distance to a Known Distribution $\mathcal{D}_\star$ with parameter $\epsilon \in (0,1)$

Let $k = \lceil 4/\epsilon^2 \rceil$.
Collect k independent $(1+\epsilon/3)$-PR-samples (x_1, p_1), (x_2, p_2), $\ldots$, (x_k, p_k) from $\mathcal{D}$.
return $1 - \frac{1}{k}\sum_{i\in[k]} \min\{1, \mathcal{D}_\star[x_i]/p_i\}$

and reject if $\mathcal{D} = \mathcal{D}_1$. We claim that the algorithm cannot succeed with probability at least $2/3$ if the number of samples is at most $1/(10\epsilon)$.

We say that a PR-sample (x, p) is *exposing* if $x \neq 1$. For each of $\mathcal{D}_0$ and $\mathcal{D}_1$, a PR-sample is exposing with probability exactly ϵ. If at most $1/(10\epsilon)$ independent PR-samples are drawn, the expected number of exposing samples is $1/10$. By Markov's inequality, the probability that at least one of these samples is exposing is bounded by $1/10$. Note that the probability that no exposing sample was selected is the same for $\mathcal{D}_0$ and $\mathcal{D}_1$. In this case, the sequence of PR-samples consists of pairs $(1, 1-\epsilon)$ and for a randomly chosen $\mathcal{D} \in \{\mathcal{D}_0, \mathcal{D}_1\}$, the algorithm cannot be correct with probability greater than $1/2$. Therefore, the probability of success of any algorithm that uses at most $1/(10\epsilon)$ PR-samples is bounded by $1/2 + 1/10 < 2/3$. □

5 DISTANCE TO A KNOWN DISTRIBUTION

We prove Theorem 1.3 in this section. We prove the upper bound as Lemma 5.1 and the lower bound as Lemma 5.2.

5.1 Upper Bound

We show that Algorithm 2 solves the Distance to a Known Distribution problem using $O(1/\epsilon^2)$ PR-samples.

Lemma 5.1. *For any distribution $\mathcal{D}_\star$ with the probability of each element known and for any distribution $\mathcal{D}$, Algorithm 2 collects $O(1/\epsilon^2)$ independent $(1+\epsilon/3)$-PR-samples from $\mathcal{D}$ and outputs an estimate E such that with probability 2/3, $|E - d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_\star)| \leq \epsilon$.*

Proof. Let $S = \{x \in [n] : \mathcal{D}[x] > 0\}$ be the set of elements with non-zero probability in $\mathcal{D}$. Let $\psi = \sum_{z\in S} \min\{\mathcal{D}_\star[z], \mathcal{D}[z]\}$, which can be seen as the probability mass shared by the distributions. We have $d_{\mathrm{TV}}(\mathcal{D}_\star, \mathcal{D}) = 1 - \psi$. Let $T = \sum_{i\in[k]} \min\{1, \mathcal{D}_\star[x_i]/p_i\}$. We prove that with probability $2/3$, $|T - k\psi| \leq \epsilon k$, in which case we obtain the desired bound $|(1 - T/k) - d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)| \leq \epsilon$. We rewrite the definition of T as $T = \sum_{1\leq i\leq k} T_i$, where $T_i = \min\{p_i, \mathcal{D}_\star[x_i]\}/p_i$.

We start by assuming that the samples obtained by the algorithm are exact PR-samples, i.e., for all $i \in [k]$, $p_i = \mathcal{D}[x_i]$. For a single term T_i, we have

$$\begin{aligned}\mathbb{E}[T_i] &= \sum_{z\in S} \mathcal{D}[z] \cdot \frac{\min\{\mathcal{D}[z], \mathcal{D}_\star[z]\}}{\mathcal{D}[z]} \\ &= \sum_{z\in S} \min\{\mathcal{D}_\star[z], \mathcal{D}[z]\} = \psi.\end{aligned}$$

By the linearity of expectation, $\mathbb{E}[T] = k\psi$. Since T_i's are independent and each $T_i \in [0,1]$, we can apply Hoeffding's inequality:

$$\begin{aligned}\Pr\left[|T - k\psi| \geq \epsilon k/2\right] &\leq 2\exp\left(-\frac{2(\epsilon k/2)^2}{k}\right) \\ &= 2\exp\left(-\frac{\epsilon^2 k}{2}\right) \leq 2e^{-2} < 1/3.\end{aligned}$$

We now bound the impact that approximate PR-samples can have on the output of the algorithm. In this case, $T_i = \min\{\mathcal{D}_\star[x_i], p_i\}/p_i$, and therefore, if $(1+\epsilon/3)$-PR-samples are used,

$$\begin{aligned}&\frac{1}{1+\epsilon/3} \cdot \frac{\min\{\mathcal{D}_\star[x_i], \mathcal{D}[x_i]\}}{\mathcal{D}[x_i]} \leq T_i \\ &\leq (1+\epsilon/3) \cdot \frac{\min\{\mathcal{D}_\star[x_i], \mathcal{D}[x_i]\}}{\mathcal{D}[x_i]},\end{aligned}$$

as opposed to $T_i = \min\{\mathcal{D}_\star[x_i], \mathcal{D}[x_i]\}/\mathcal{D}[x_i]$. Let T' denote the estimate obtained using exact PR-samples (with the same sequence of x_i's) as opposed to approximate PR-samples. We have

$$(1-\epsilon/3)T' \leq \frac{T'}{1+\epsilon/3} \leq T \leq (1+\epsilon/3)T'.$$

Therefore, $|T' - T| \leq \epsilon T'/3$. With probability $2/3$, $|T' - k\psi| \leq \epsilon k/2$, and hence,

$$\begin{aligned}|T - k\psi| &\leq |T - T'| + |T' - k\psi| \leq \epsilon T'/3 + \epsilon k/2 \\ &\leq (1+\epsilon/3)k \cdot \epsilon/3 + \epsilon k/2 \leq \epsilon k.\end{aligned}$$ □

5.2 Lower Bound

To prove the lower bound, we let the known distribution $\mathcal{D}_\star$ be a uniform distributions over half of the domain. We define two families of distributions: The first family contains all the uniform distributions over half of the domain with total variation distance $1/2$ to the known distribution. The second family contains all the uniform distributions over half of the domain with total variation distance $1/2 - 3\epsilon$ to the known distribution.

We prove the lower bound by showing that it is impossible to distinguish whether the unknown distribution

Algorithm 3: Estimate Total Variation Distance of Two Distributions on $[n]$ with parameter $\epsilon \in (0,1)$

Let $k = \lceil \max\{5\sqrt{n}/\epsilon, 48/\epsilon^2\} \rceil$.
Collect k independent $(1+\epsilon/3)$-PR-samples (x_1, p_1), (x_2, p_2), $\ldots$, (x_k, p_k) from $\mathcal{D}_1$.
Collect k independent $(1+\epsilon/3)$-PR-samples (y_1, q_1), (y_2, q_2), $\ldots$, (y_k, q_k) from $\mathcal{D}_2$.
Let $T = \sum_{i,j \in [k] \text{ s.t. } x_i = y_j} \min\{p_i, q_j\}/(p_i q_j)$.
return $1 - T/k^2$

is from the first family or from the second family with probability at least $2/3$ using c/ϵ^2 samples for a small constant c.

Lemma 5.2. *If $\epsilon < 1/10000$ and $n = \omega(1/\epsilon^4)$, then the Distance to a Known Distribution problem requires $\Omega(1/\epsilon^2)$ PR-samples if an algorithm has to succeed with probability at least $2/3$.*

6 EQUALITY TESTING AND DISTANCE BETWEEN UNKNOWN DISTRIBUTIONS

We prove Theorem 1.3 in this section. We prove the upper bound as Lemma 6.1 and the lower bound as Lemma 6.2.

6.1 Upper Bound

We show that Algorithm 3 estimates the total variation distance between two unknown distributions using $O(\max\{\sqrt{n}/\epsilon, 1/\epsilon^2\})$ PR-samples.

Lemma 6.1. *For two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $[n]$ and a parameter $\epsilon \in (0,1)$, Algorithm 3 collects $O(\max\{\sqrt{n}/\epsilon, 1/\epsilon^2\})$ $(1+\epsilon/3)$-PR-samples from the distributions and outputs an estimate E such that with probability 2/3, $|E - d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)| \le \epsilon$.*

Proof. Let $S = \{x \in [n] : \mathcal{D}_1[x] > 0 \wedge \mathcal{D}_2[x] > 0\}$ be the set of elements with non-zero probability in both distributions. Let $\psi = \sum_{z \in S} \min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}$ be the overlap of the distributions. We have $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = 1 - \psi$. We prove that with probability $2/3$, $|T - k^2\psi| \le \epsilon k^2$, in which case we obtain the desired bound $|(1 - T/k^2) - d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)| \le \epsilon$.

We rewrite the definition of T as $T = \sum_{1 \le i,j \le k} T_{i,j}$, where

$$T_{i,j} = \begin{cases} \min\{p_i, q_j\}/(p_i q_j) & \text{if } x_i = y_j, \\ 0 & \text{otherwise.} \end{cases}$$

We start by assuming that the samples obtained by the algorithm are exact PR-samples, i.e., for all $i \in [k]$, $p_i = \mathcal{D}[x_i]$ and $q_i = \mathcal{D}[y_i]$. For a single term $T_{i,j}$, we have

$$\begin{aligned} \mathbb{E}[T_{i,j}] &= \sum_{z \in S} \mathcal{D}_1[z] \cdot \mathcal{D}_2[z] \cdot \frac{\min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}}{\mathcal{D}_1[z]\mathcal{D}_2[z]} \\ &= \sum_{z \in S} \min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\} = \psi. \end{aligned}$$

By the linearity of expectation, $\mathbb{E}[T] = k^2\psi$. This implies that the expected value returned by the algorithm is $\mathbb{E}[1 - T/k^2] = 1 - \psi = d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)$. We now bound the variance of T: $\mathrm{Var}(T) = \mathbb{E}[T^2] - (\mathbb{E}[T])^2$. We have

$$\begin{aligned} \mathbb{E}[T^2] = &\sum_{i,j \in [k]} \mathbb{E}[T_{i,j}^2] + \sum_{\substack{i,j,j' \in [k] \\ j \ne j'}} \mathbb{E}[T_{i,j} T_{i,j'}] \\ &+ \sum_{\substack{i,i',j \in [k] \\ i \ne i'}} \mathbb{E}[T_{i,j} T_{i',j}] + \sum_{\substack{i,i',j,j' \in [k] \\ i \ne i' \\ j \ne j'}} \mathbb{E}[T_{i,j} T_{i',j'}]. \end{aligned}$$

We now bound each of the terms in the above equation. First,

$$\begin{aligned} \sum_{i,j \in [k]} \mathbb{E}[T_{i,j}^2] &= k^2 \sum_{z \in S} \mathcal{D}_1[z]\mathcal{D}_2[z] \left(\frac{\min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}}{\mathcal{D}_1[z]\mathcal{D}_2[z]} \right)^2 \\ &= k^2 \sum_{z \in S} \frac{\min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}}{\max\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}} \le k^2 n. \end{aligned}$$

Second,

$$\begin{aligned} &\sum_{\substack{i,j,j' \in [k] \\ j \ne j'}} \mathbb{E}[T_{i,j} T_{i,j'}] \\ &= k^2(k-1) \sum_{z \in S} \mathcal{D}_1[z](\mathcal{D}_2[z])^2 \left(\frac{\min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}}{\mathcal{D}_1[z]\mathcal{D}_2[z]} \right)^2 \\ &\le k^2(k-1) \sum_{z \in S} \min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\} = k^2(k-1)\psi. \end{aligned}$$

By symmetry, $\sum_{\substack{i,i',j \in [k] \\ i \ne i'}} \mathbb{E}[T_{i,j} T_{i',j}]$ is also at most $k^2(k-1)\psi$. Finally,

$$\begin{aligned} &\sum_{\substack{i,i',j,j' \in [k] \\ i \ne i' \\ j \ne j'}} \mathbb{E}[T_{i,j} T_{i',j'}] \\ &= k^2(k-1)^2 \sum_{z,w \in S} \mathcal{D}_1[z]\mathcal{D}_2[z]\mathcal{D}_1[w]\mathcal{D}_2[w] \\ &\qquad \cdot \left(\frac{\min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\}}{\mathcal{D}_1[z]\mathcal{D}_2[z]} \right) \left(\frac{\min\{\mathcal{D}_1[w], \mathcal{D}_2[w]\}}{\mathcal{D}_1[w]\mathcal{D}_2[w]} \right) \\ &= k^2(k-1)^2 \left(\sum_{z \in S} \min\{\mathcal{D}_1[z], \mathcal{D}_2[z]\} \right) \\ &\qquad \cdot \left(\sum_{w \in S} \min\{\mathcal{D}_1[w], \mathcal{D}_2[w]\} \right) \\ &= k^2(k-1)^2 \psi^2. \end{aligned}$$

‘ Thus, $\mathbb{E}[T^2] \le k^2 n + 2k^3\psi + k^4\psi^2$, and $\mathrm{Var}(T) \le k^2 n + 2k^3\psi + k^4\psi^2 - (k^2\psi)^2 = k^2 n + 2k^3\psi \le k^2 n + 2k^3$. Since $k \ge \max\{5\sqrt{n}/\epsilon, 48/\epsilon^2\}$, $n \le \epsilon^2 k^2/25$ and $6 \le \epsilon^2 k/8$. We apply these inequalities:

$$\begin{aligned}\sqrt{3\,\mathrm{Var}(T)} &\le \sqrt{3k^2 n + 6k^3} \\ &\le \sqrt{3k^2 \cdot (\epsilon^2 k^2/25) + (\epsilon^2 k/8)\cdot k^3} \\ &\le \epsilon k^2/2.\end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned}&\Pr\left(|T - \mathbb{E}[T]| \ge \epsilon k^2/2\right) \\ &\le \Pr\left(|T - \mathbb{E}[T]| \ge \sqrt{3\,\mathrm{Var}(T)}\right) \le 1/3.\end{aligned}$$

Therefore, for exact PR-samples, the algorithm outputs an estimate at distance at most $\epsilon/2$ from $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)$ with probability at least 2/3.

We now bound the impact that approximate PR-samples can have on the output of the algorithm. A term $T_{i,j}$ can be non-zero only if $x_i = y_j$ (which can happen only if both $\mathcal{D}_1[x_i]$ and $\mathcal{D}_2[y_j]$ are non-zero). In this case, $T_{i,j} = \min\{p_i, q_j\}/(p_i q_j) = 1/\max\{p_i, q_j\}$, and therefore, if $(1+\epsilon/3)$-PR-samples are used,

$$\begin{aligned}&\frac{1}{1+\epsilon/3} \cdot \frac{\min\{\mathcal{D}_1[x_i], \mathcal{D}_2[y_j]\}}{\mathcal{D}_1[x_i]\mathcal{D}_2[y_j]} \le T_{i,j} \\ &\le (1+\epsilon/3) \cdot \frac{\min\{\mathcal{D}_1[x_i], \mathcal{D}_2[y_j]\}}{\mathcal{D}_1[x_i]\mathcal{D}_2[y_j]},\end{aligned}$$

as opposed to $T_{i,j} = \min\{\mathcal{D}_1[x_i], \mathcal{D}_2[y_j]\}/(\mathcal{D}_1[x_i]\mathcal{D}_2[y_j])$. Let $T_\star$ denote the estimate obtained using exact PR-samples (with the same sequence of x_i's and y_j's) as opposed to approximate PR-samples. We have

$$(1-\epsilon/3)T_\star \le \frac{T_\star}{1+\epsilon/3} \le T \le (1+\epsilon/3)T_\star.$$

Therefore, $|T_\star - T| \le \epsilon T_\star/3$. With probability 2/3, $|T_\star - k^2\psi| \le \epsilon k^2/2$, and hence,

$$\begin{aligned}|T - k^2\psi| &\le |T - T_\star| + |T_\star - k^2\psi| \le \epsilon T_\star/3 + \epsilon k^2/2 \\ &\le (1+\epsilon/2)k^2 \cdot \epsilon/3 + \epsilon k^2/2 \\ &\le (1/3 + \epsilon/6 + 1/2)\cdot \epsilon k^2 \le \epsilon k^2,\end{aligned}$$

which finishes the proof. □

6.2 Lower Bound

To prove the lower bound, we construct two families of distribution pairs, in which all the distributions are uniform over half of the domain. The first family contains all the pairs of identical distributions. The second family contains all the pairs of distributions with total variation distance ϵ.

We show that it is impossible to distinguish whether the two unknown distributions are from the first family or from the second family with probability at least 2/3 using $c\sqrt{n}/\epsilon^2$ samples for a small constant c.

Lemma 6.2. *If $\epsilon < 1/100$ and $n = \omega((1/\epsilon)^{10})$, then the Equality Testing problem requires $\Omega(\sqrt{n}/\epsilon)$ PR-samples if an algorithm has to succeed with probability at least 2/3.*

7 DISTINCT ELEMENTS

We prove Theorem 1.4 in this section. We prove the upper bound as Lemma 7.1 and the lower bound as Lemma 7.2.

7.1 Upper Bound

Algorithm 4: Estimating the Number of Distinct Elements up to an Additive ϵn

1 Let $k = \lceil 27/\epsilon^2 \rceil$.
2 Collect k independent $(1+\epsilon/3)$-PR-samples (x_1, p_1), (x_2, p_2), …, (x_k, p_k) from $\mathcal{D}$.
3 **return** $\frac{1}{k}\sum_{i\in[k]} 1/p_i$

Lemma 7.1. *Let $\epsilon \in (0,1)$ and let $\mathcal{D}$ be a discrete distribution where each element that can be drawn from $\mathcal{D}$ has probability at least $1/n$. Algorithm 4 uses $O(1/\epsilon^2)$ $(1+\epsilon/3)$-PR-samples to compute an estimate to the number of distinct elements in $\mathcal{D}$. With probability 2/3, the difference between the estimate and the number of distinct elements is bounded by ϵn.*

7.2 Lower Bound

We show the lower bound by constructing two families of distributions. The support size of every distribution in the first family is $3n/4$, and the support size of every distribution in the second family $n(3/4 - 3\epsilon)$. We prove the lower bound of Distinct Elements problem by showing that it is impossible to distinguish whether the unknown distribution is from the first family or from the second family with probability at least 2/3 using c/ϵ^2 samples for a small constant c.

Theorem 7.2. *If $\epsilon \le 1/10000$ and $n = \omega(1/\epsilon^4)$, then the Distinct Elements problem requires $\Omega(1/\epsilon^2)$ PR-samples if an algorithm has to succeed with probability at least 2/3.*

References

[1] Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM J. Comput.*, 35(1):132–150, 2005.

[2] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *FOCS*, pages 442–451, 2001.

[3] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[4] Ran Canetti, Guy Even, and Oded Goldreich. Lower bounds for sampling algorithms for estimating the average. *Inf. Process. Lett.*, 53(1):17–25, 1995.

[5] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *SODA*, pages 1174–1192, 2014.

[6] Clément L. Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *ICALP (1)*, pages 283–295, 2014.

[7] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *ITCS*, pages 561–580, 2013.

[8] Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[9] Bernard Chazelle, Ronitt Rubinfeld, and Luca Trevisan. Approximating the minimum spanning tree weight in sublinear time. *SIAM J. Comput.*, 34(6):1370–1379, 2005.

[10] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[11] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[12] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.*, 39(3):813–842, 2009.

[13] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011.

[14] Gregory Valiant and Paul Valiant. Instance-by-instance optimal identity testing. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:111, 2013.

[15] Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011.