Optimal Testing of Discrete Distributions with High Probability

Ilias Diakonikolas*
ilias@cs.wisc.edu
University of Wisconsin-Madison
USA

Themis Gouleakis†
tgouleak@mpi-inf.mpg.de
Max Planck Institute for Informatics
Germany

Daniel M. Kane‡
dakane@cs.ucsd.edu
University of California, San Diego
USA

John Peebles§
jltp@princeton.edu
Princeton University
USA

Eric Price¶
ecprice@cs.utexas.edu
UT Austin
USA

ABSTRACT

We study the problem of testing discrete distributions with a focus on the high probability regime. Specifically, given samples from one or more discrete distributions, a property $\mathcal{P}$, and parameters $0 < \epsilon, \delta < 1$, we want to distinguish *with probability at least* $1 - \delta$ whether these distributions satisfy $\mathcal{P}$ or are ϵ-far from $\mathcal{P}$ in total variation distance. Most prior work in distribution testing studied the constant confidence case (corresponding to $\delta = \Omega(1)$), and provided sample-optimal testers for a range of properties. While one can always boost the confidence probability of any such tester by black-box amplification, this generic boosting method typically leads to sub-optimal sample bounds.

Here we study the following broad question: For a given property $\mathcal{P}$, can we *characterize* the sample complexity of testing $\mathcal{P}$ as a function of all relevant problem parameters, including the error probability δ? Prior to this work, uniformity testing was the only statistical task whose sample complexity had been characterized in this setting. As our main results, we provide the first algorithms for closeness and independence testing that are sample-optimal, within constant factors, as a function of all relevant parameters. We also show matching information-theoretic lower bounds on the sample complexity of these problems in the full version of this paper. Our techniques naturally extend to give optimal testers for related problems. To illustrate the generality of our methods, we give optimal algorithms for testing collections of distributions and testing closeness with unequal sized samples.

*Supported by NSF Award CCF-1652862 (CAREER), NSF AiTF Award CCF-2006206, and a Sloan Research Fellowship.
†Some of this work was performed while the author was a postdoctoral researcher at USC.
‡Supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.
§Supported by the Swiss National Science Foundation Grant #200021_182527.
¶Supported by NSF Award CCF-1751040 (CAREER).

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3450997

CCS CONCEPTS

• **Theory of computation → Streaming, sublinear and near linear time algorithms**.

KEYWORDS

distribution property testing, sampling, high confidence

ACM Reference Format:
Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. 2021. Optimal Testing of Discrete Distributions with High Probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3406325.3450997

1 INTRODUCTION

1.1 Background and Motivation

This paper studies problems in distribution property testing [2, 3, 24], a field at the intersection of property testing [23, 34] and statistical hypothesis testing [28, 31]. The prototypical problem of this field is the following: Given sample access to a collection of unknown probability distributions and a pre-specified global property $\mathcal{P}$ of these distributions, determine whether the distributions satisfy $\mathcal{P}$ or are "far" from satisfying the property. (See Section 1.2 for a formal definition.) The main goal is to characterize the sample and computational complexity of this general question, for any given property $\mathcal{P}$ of interest, as a function of the relevant parameters. During the past two decades, distribution property testing has received significant attention within the computer science and statistics communities. The reader is referred to [5, 33] for two surveys on the topic. It should be noted that the TCS definition of distribution testing is equivalent to the minimax view of statistical hypothesis testing, pioneered in the statistics community by Ingster and coauthors (see, e.g., [26].)

The vast majority of prior research in distribution testing focused on characterizing the complexity of testing various properties of arbitrary discrete distributions in the "constant confidence regime." That is, the testing algorithm is allowed to fail with probability (say) at most 1/3. This regime is by now fairly well understood: For a range of natural and important properties (see, e.g., [1, 7, 8, 15–18, 30, 32, 36]), prior work has developed testers with provably optimal sample complexity (up to universal constant factors). More recently, a body of work has focused on leveraging a priori structure

of the underlying distributions to obtain significantly improved sample complexities [4, 6, 9–11, 17–20]. Similarly, all these results on testing structured distributions study the constant confidence regime.

Since distribution property testing is a (promise) decision problem, one can use standard amplification to boost the confidence probability of any tester to any desired value in a black-box manner. Suppose we have a testing algorithm for property $\mathcal{P}$ that guarantees confidence probability $2/3$ (failure probability $1/3$) with N samples. Using amplification, we can increase the confidence probability to $1-\delta$, for any $\delta > 0$, by increasing the sample complexity of the algorithm by a factor of $\Theta(\log(1/\delta))$. In part due to this simple fact, the initial definition of property testing [23] had set the confidence parameter δ to be constant by default. As Goldreich notes [21], "eliminating the error probability as a parameter does not allow to ask whether or not one may improve over the straightforward error reduction". Indeed, as we will see below, for a range of tasks this $\Theta(\log(1/\delta))$ multiplicative increase in the sample size is sub-optimal.

The previous paragraph leads us to the following general question:

Question 1.1. *For a given property $\mathcal{P}$, can we* characterize *the sample complexity of testing $\mathcal{P}$ as a function of all relevant problem parameters, including the error probability δ?*

We believe that Question 1.1 is of fundamental theoretical and practical interest that merits investigation in its own right. The analogous question in the context of *distribution learning* has been intensely studied in statistical learning theory (see, e.g., [12, 37]) and tight bounds are known in a range of settings.

Question 1.1 is of substantial interest in statistical hypothesis testing, where the family of distribution testing algorithms with failure probability δ for a given property $\mathcal{P}$ is equivalent to the family of minimax statistical tests whose probability of Type I error (p-value) and probability of Type II error are both at most δ. Standard techniques for addressing the problem of multiple comparisons, such as Bonferroni correction, require vanishingly small p-values. In such settings, obtaining optimal testers in the high-confidence regime might have practical implications in application areas of hypothesis testing (e.g., in biology).

It should be noted that Question 1.1 has received renewed research attention in the information theory and statistics communities. Specifically, [25, 27] focused on developing testers with improved dependence on δ for uniformity testing [25], equivalence and independence testing [27]. Prior to this work, uniformity testing—and, via Goldreich's reduction [22], identity testing—was the only statistical task whose sample complexity had been characterized in the high-confidence regime [14]. As shown in [14], all previously studied uniformity testers are in fact sub-optimal in the high-confidence regime. In other words, obtaining an optimal sample bound was not just a matter of improved analysis, but a new algorithm was required.

Most relevant to the results of this paper is the concurrent work by Kim, Balakrishnan, and Wasserman [27]. Kim *et al.* [27] give equivalence and independence testers for discrete distributions with respect to the total variation distance (i.e., in the same setting as ours) whose sample complexities beat standard amplification as a function of δ (in some parameter regimes). As we show in this paper, their sample complexity upper bounds are sub-optimal – by roughly a quadratic factor. See Section 1.4 for a detailed description of the most relevant prior work.

1.2 Our Contributions

In this work, we systematically investigate the sample complexity of distribution testing in the high-confidence regime. Our main focus is on the problems of closeness (equivalence) testing and independence testing. We develop new techniques that lead to the first sample-optimal testing algorithms for these properties. Moreover, we prove information-theoretic lower bounds showing that the sample complexity of our algorithms is optimal in all parameters (within a constant factor). Our techniques can be naturally adapted to give sample-optimal testers for other properties. To illustrate the generality of our methods, we show that our techniques lead to sample-optimal testers (and matching lower bounds in the full version of this paper) for testing properties of collections of distributions and testing closeness with unequal sized samples.

We start with a general definition of distribution property testing for tuples of distributions.

Definition 1.1 ((ϵ,δ)-testing of property $\mathcal{P}$)**.** Let $\mathcal{P}$ be a property of a k-tuple of distributions. Given parameters $0 < \epsilon, \delta < 1$, and sample access to a collection of distributions $p^{(1)}, \ldots, p^{(k)}$, we want to distinguish *with probability at least* $1-\delta$ between the following cases:

- **Completeness**: $(p^{(1)}, \ldots, p^{(k)}) \in \mathcal{P}$.
- **Soundness**: $(p^{(1)}, \ldots, p^{(k)})$ is ϵ-far from $\mathcal{P}$, in total variation distance, i.e., for every $(q^{(1)}, \ldots, q^{(k)}) \in \mathcal{P}$ the average total variation distance between $p^{(i)}$ and $q^{(i)}$, $i \in [k]$, is at least ϵ.

We call this the problem of (ϵ,δ)-*testing property* $\mathcal{P}$. An algorithm that solves this problem will be called an (ϵ,δ)-*tester* for property $\mathcal{P}$.

Here we focus on testing properties of distributions on discrete domains. Definition 1.1 captures all testing tasks we study in this paper. Our contributions are described in detail in the proceeding discussion.

The task of closeness testing (or equivalence testing) of two discrete distributions p, q supported on $[n]$ corresponds to the case $k = 2$ of Definition 1.1 and the property in question is $\mathcal{P} = \{(p, q) : p = q\}$. In other words, given samples from p and q, we want to distinguish between the cases that $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \epsilon$. For closeness testing, we show:

Theorem 1.2 (Closeness Testing). *There exists a computationally efficient (ϵ,δ)-closeness tester for discrete distributions of support size n with sample complexity*

$$\Theta\left(n^{2/3}\log^{1/3}(1/\delta)/\epsilon^{4/3} + (n^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta))/\epsilon^2\right).$$

Moreover, this sample size upper bound is information-theoretically optimal, within a universal constant factor, for all n, ϵ, δ.

The statistical task of (two-dimensional) independence testing of a discrete distribution p on the domain $[n] \times [m]$ corresponds to the case $k = 1$ of Definition 1.1, where the property of interest is $\mathcal{P} =$

$\{p : p \text{ is a product distribution}\}$. That is, we want to distinguish between the case that p is a product distribution versus ϵ-far, in total variation distance, from any product distribution. For independence testing, we show:

Theorem 1.3 (Independence Testing). *There exists a computationally efficient (ϵ, δ)-independence tester for discrete distributions on $[n] \times [m]$, where $n \geq m$, with sample complexity:*

$$\Theta\left(n^{2/3}m^{1/3}\log^{1/3}(1/\delta)/\epsilon^{4/3}\right) + \\ +\Theta\left(((nm)^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta))/\epsilon^2\right) .$$

Moreover, this sample size upper bound is information-theoretically optimal, within a universal constant factor, for all n, m, ϵ, δ.

The main focus of this paper is on developing the techniques required to establish Theorems 1.2 and 1.3. Building on these techniques, we obtain optimal testers for two additional fundamental properties.

In the task of testing collections of distributions, we are given access to m distributions $p^{(1)}, \ldots, p^{(m)}$ supported on $[n]$ and we want to distinguish between the case that $p^{(1)} = p^{(2)} = \ldots = p^{(m)}$ and the case that $\min_q (1/m)\sum_{i=1}^m d_{TV}(p^{(i)}, q) \geq \epsilon$. Our algorithm is given samples of the form (i, j), where i is drawn uniformly at random from $[m]$ and $j \in [n]$ is drawn from $p^{(i)}$. While this problem has strong similarities to independence testing, it also has some significant differences. For this testing task, we show:

Theorem 1.4 (Testing Collections of Distributions). *There exists a computationally efficient (ϵ, δ)-tester for testing closeness of collections of m distributions on $[n]$ with sample complexity:*

$$\Theta\left(n^{2/3}m^{1/3}\log^{1/3}(1/\delta)/\epsilon^{4/3}\right) + \\ +\Theta\left(((nm)^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta))/\epsilon^2\right) .$$

Moreover, this sample size upper bound is information-theoretically optimal, within a universal constant factor, for all n, m, ϵ, δ.

Our final result is for the problem of testing closeness between two unknown discrete distributions when we have access to unequal sized samples from the two unknown distributions. This problem interpolates between the vanilla closeness testing task (with equal sized samples) and the task of identity testing (where one of the two distributions is known exactly). For this task, we show:

Theorem 1.5 (Closeness Testing with Unequal Sized Samples). *There exists a computationally efficient (ϵ, δ)-closeness tester for discrete distributions of support size n that draws $O(K+k)$ samples from one distribution and $O(k)$ samples from the other, as long as*

$$k \geq C\left(n\sqrt{\log(1/\delta)/\min(n, K)} + \log(1/\delta)\right)/\epsilon^2 ,$$

where $C > 0$ is a universal constant. Moreover, this sample size trade-off is information-theoretically optimal, within a universal constant factor, for all n, ϵ, δ.

1.3 Overview of Techniques

In this section, we provide a detailed overview of our upper and lower bound techniques. Full statements of the lower bounds and their proofs can be found in the full version of this paper. Our main technical and conceptual innovation lies in the development of our upper bounds. To keep this section concrete, we describe our techniques in the context of closeness and independence testing. Our algorithms for testing collections and closeness with unequal sized samples use very similar ideas to those of our independence tester.

Closeness Tester. To obtain a closeness tester that performs well in the high confidence regime, we need to design a test statistic that exhibits strong concentration bounds. A reasonable approach to enforce this requirement would be to ensure that the test statistic is Lipschitz in the samples, so that we can leverage an appropriate concentration inequality (e.g., McDiarmid's inequality) to obtain the necessary concentration. We note that the chi-squared closeness tester of [8] is Lipschitz, but not Lipschitz enough for the straightforward analysis to obtain an optimal bound. While we conjecture that the [8] closeness tester is indeed optimal, here we develop a new and easier to analyze closeness tester. Our new closeness tester (and its analysis) will also be crucially used for our independence tester.

We are now ready to describe the new statistic that our closeness tester relies on. Let X_i, Y_i be the number of samples assigned to bin (domain element) $i \in [n]$, from p and q respectively. A natural starting point is to consider the absolute value of the difference $|X_i - Y_i|$. Namely, we could consider the statistic $Z = \sum_{i=1}^n |X_i - Y_i|$ and output "YES" or "NO" based on its magnitude. Unfortunately, this random variable Z does not have mean zero in the completeness case (i.e., when $p = q$). Furthermore, one can construct instances where the expectation of this statistic is not even minimized when $p = q$. To fix this issue, we will need to subtract an appropriate proxy for what the value should be if $p = q$. To do this, we draw a second set of samples with X_i' and Y_i' samples in bin i from each of the distributions. We then use the test statistic

$$Z = \sum_{i=1}^n \left(|X_i - Y_i| + |X_i' - Y_i'| - |X_i - X_i'| - |Y_i - Y_i'|\right) .$$

If $p = q$, it is clear that X_i, X_i', Y_i, Y_i' are i.i.d., and so Z is mean zero. The challenging part of the proof involves showing that if p is ϵ-far from q, then $\mathbf{E}[Z]$ must be large. Since Z is Lipschitz, it satisfies strong concentration bounds, and so with sufficiently many samples we can distinguish the two cases with high probability. A careful analysis shows that this tester is indeed sample optimal for the entire parameter regime.

Independence Tester. Let p be a discrete distribution on $[n] \times [m]$. It is easy to see (and well-known) that the independence testing problem amounts to distinguishing the case where $p = q$ from the case that p is ϵ-far from q, where q is the product of p's marginals. Unfortunately, directly applying Theorem 1.2 to this domain of size nm gives a poor sample complexity in one of the three terms. In particular, the first term would be $n^{2/3}m^{2/3}$, not $n^{2/3}m^{1/3}$. Of course, this is an issue even for the constant confidence regime. We thus need a better bound when this term is dominant, which we

will obtain using tighter concentration bounds on our statistic Z from the previous subsection.

We start by observing that if Z is computed by drawing a total of k independent samples, the fact that Z is Lipschitz implies a variance bound of $O(k)$. By McDiarmid's inequality, it follows that Z is within $O(\sqrt{k \log(1/\delta)})$ of its mean value with probability $1 - \delta$. However, we note that the value of the output statistic for Z does not really depend on all of the samples. In particular, any bin (domain element) with exactly one sample drawn from it (from the combination of p and q) will not contribute to the statistic. Hence, if we let N be the number of non-isolated samples, then in some sense, the variance of Z will be bounded above by N. Formally speaking, some technical work is needed here, because there is a low probability of N being unusually small in which case it would bound the variance. To address this, we use a symmetrization argument to show that $|Z - \mathbf{E}[Z]| = O(\sqrt{(N + \log(1/\delta)) \log(1/\delta)})$ with probability at least $1 - \delta$ (see Lemma 4.5). If we can ensure that the number of non-isolated samples is not too large, this stronger concentration bound should allow us to use fewer samples.

In order to decrease the number of non-singleton samples in our distribution, it is natural to want our underlying distributions to have small ℓ_2 norm. An approach to achieve this is by using the flattening technique of [16]. The basic idea of flattening is to use some of our samples to identify the heavy bins in our distribution, and then to artificially subdivide these bins in order to decrease the total ℓ_2 norm of the distribution. This technique is especially useful for the product distribution q, as we can separately identify the heavy x-coordinates and heavy y-coordinates, rather than using what would need to be substantially more samples to identify all of the heavy pairs. *However, there are two major difficulties with using flattening in this setting. To circumvent these obstacles, new ideas are needed, as explained in the proceeding discussion.*

First, although flattening can be used to reduce the number of collisions coming from samples of q, it will not necessarily reduce the number of collisions from p-samples to acceptable levels. We get around this issue by noting that if most of the collisions contributing to N come largely from p-samples, then with high probability it will be case that $Z \gg N$, in which case the larger variance term will not hurt us much. A second, more difficult, problem to handle is this: although it is not hard to show that flattening works *on average*, it simply is not true that flattening yields a small number of collisions with sufficiently high probability. This is a major issue in our setting, since our goal is to obtain the optimal sample complexity with high confidence!

To circumvent the latter problem, we will need to substantially restructure our algorithm. Essentially, we will pick a set S of samples once at the beginning of our algorithm. We then randomly assign samples of S to be used either to flatten x and y coordinates, or to generate samples from p and q. If we got unlucky and our flattening was not sufficient (because the number of q-samples that collided was too large), we will try again using the same initial set S of samples, but re-randomizing the way these samples are used.

To show that this new algorithm works, we will need to establish two statements:

(1) For *any* set of initial samples S, the probability that we will need to try again is at most 50% (so, on average, we only need to try a constant number of times).
(2) The probability that a given try causes our algorithm to terminate with the wrong answer is at most δ.

Combining the second statement with the fact that on average we will only need $O(1)$ many tries before we get an answer, the total probability of failure will be bounded by $\delta \mathbf{E}[\#\text{ tries}] = O(\delta)$. This allows us to get a high-probability bound even though our analysis of flattening only works on average.

Sample Complexity Lower Bounds. We sketch our sample complexity lower bound for independence testing. Details can be found in the full version of this paper. The corresponding lower bound for closeness testing follows as a special case in a black-box manner.

Our lower bound proof follows the same outline as the lower bound proof in [16]. The gist of the argument in that work was that we reduced to the following problem: We have two explicit pseudo-distributions[1] D_{yes} (over independent pseudo-distributions) and D_{no} (over usually far from independent pseudo-distributions). We pick a random pseudo-distribution from one of these families, take $\text{Poi}(k)$ samples from it, hand them to the algorithm, and ask the algorithm to determine which ensemble we started with. It was shown in [16] that it is impossible to do this reliably by bounding the *mutual information* between the samples and the bit determining which ensemble was sampled from.

This approach, unfortunately, does not suffice for high probability bounds. [16] worked in the constant confidence regime, where the mutual information is close to 0. In contrast, in the high confidence regime, the mutual information will be close to 1. While, in principle, bounding the mutual information away from 1 might suffice to prove lower bounds in the high confidence regime, the mutual information bounds achievable with the [16] techniques are not sufficiently strong, in the sense that they can only bound the mutual information by a quantity bigger than 1, given enough samples.

To overcome this technical hurdle, we replace our bounds on mutual information with bounds on KL-divergence. Unlike the mutual information (which is bounded by 1 bit), the KL-divergence between our distributions can become arbitrarily large. It is also not hard to see that if two distributions can be distinguished with probability $1 - \delta$, the KL-divergence is $\Omega(\log(1/\delta))$. (See Fact 2.2.)

Given the above observation, our lower bound ensembles are identical to the ones used in [16]. Furthermore, the analytic techniques we use to bound the KL-divergence are very similar, using essentially the same expression as an upper bound on KL-divergence as was used as an upper bound on mutual information. Another technical issue is that we need to show that the reduction to our hard instance over pseudo-distributions still works for high probability testing, which is not difficult, but needs to be carefully checked.

1.4 Prior and Concurrent Work

Prior to this work, the question of developing sample-optimal testers in the high-confidence regime has been considered for

[1] A "pseudo-distribution" is like a distribution, except not necessarily normalized to sum to one. In other areas of mathematics, they are commonly referred to as finite measures.

uniformity testing (and, via Goldreich's reduction, identity testing). Specifically, [25] showed that Paninski's uniformity tester (based on the number of unique elements) has the sample-optimal sample complexity of $O(\sqrt{n\log(1/\delta)}/\epsilon^2)$ in the sublinear sample regime, i.e., when the sample size is $o(n)$. More recently, [14] gave a different tester that achieves the optimal sample complexity $O((\sqrt{n\log(1/\delta)}+\log(1/\delta))/\epsilon^2)$ in the entire regime of parameters.

As already mentioned, prior to our work, uniformity was the only property for which the high confidence regime has been analyzed. We now comment on some closely related literature. [8] gave a chi-squared tester and showed that it is sample-optimal in the constant confidence regime. We believe that the same tester is optimal in the high-confidence regime. However, a proof of this statement seems rather non-trivial. In particular, simple analyses based on McDiarmid's inequality [29] lead to sub-optimal sample complexity when the sample size is $\Omega(n)$. The new closeness tester introduced in this work is arguably simpler with a compact analysis, and it is crucial for our much more involved independence tester.

The work of [1] gave an independence tester that is sample optimal ' in the constant confidence regime for the special case that the two dimensions have the same support size (i.e., $n = m$). The performance of this tester is sub-optimal in the high-confidence regime, as it relies on a non-Lipschitz identity tester. [16] gave a sample-optimal independence tester for the general case (where $n \geq m$), which is the only known sample-optimal tester in the constant confidence regime for this problem. Unfortunately, this tester is also sub-optimal in the high-confidence regime for the following reason. [16] uses the flattening technique to reduce the problem under total variation (ℓ_1) distance to an ℓ_2-closeness testing problem. The issue is that the ℓ_2-testing task does not behave well in the high probability regime, so this approach does not suffice to give optimal testers in this setting. While our optimal independence tester in this paper also leverages the flattening technique, it requires several new conceptual and technical ideas.

Concurrent and independent work [27] provided testers for closeness and independence testing in the high-confidence regime. Their algorithms distinguish between the Type 1 and Type 2 error probabilities α and β respectively. Our results in this paper correspond to the setting that $\alpha = \beta = \delta$. Their testers have polynomial dependence on $1/\beta$ and therefore do not perform well in our setting. For constant β, their testers perform better than naive amplification but still sub-optimally in the parameter α. For example, their Theorem 8.1 gives a closeness tester with sample complexity of $m = O(n^{2/3}\log^{2/3}(1/\alpha)/\beta^{4/3} + n^{1/2}\log(1/\alpha)/\beta^2)$. Even for $\beta = \Theta(1)$, this is essentially quadratically worse in $\log(1/\alpha)$ than applying Theorem 1.2 with $\delta = \alpha$.

1.5 Organization

After setting up the required preliminaries in Section 2, we give our testing algorithms for closeness and independence in Sections 3 and 4. Our testers for other properties and sample complexity lower bounds are deferred to the full version of this paper [13].

2 PRELIMINARIES

We write $[n]$ to denote the set $\{1,\ldots,n\}$. We consider discrete distributions over $[n]$ with corresponding probability mass functions $p : [n] \to [0,1]$ satisfying $\sum_{i=1}^n p_i = 1$. We use the notation p_i to denote the probability of element i in distribution p. The ℓ_1 (resp. ℓ_2) norm of a distribution is identified with the ℓ_1 (resp. ℓ_2) norm of the corresponding vector, i.e., $\|p\|_1 = \sum_{i=1}^n p_i$ and $\|p\|_2 = \sqrt{\sum_{i=1}^n p_i^2}$. Similarly, the ℓ_1 (resp. ℓ_2) distance between distributions p and q is the ℓ_1 (resp. ℓ_2) norm of the vector of their difference. The total variation distance between distributions p, q on $[n]$ is $d_{\mathrm{TV}}(p,q) \stackrel{\text{def}}{=} \frac{1}{2}\cdot\|p-q\|_1$. The KL divergence between two discrete distributions p and q on $[n]$ is $D(p||q) = \sum_i p_i \log(p_i/q_i)$.

A Poisson distribution with parameter λ is denoted $\mathrm{Poi}(\lambda)$. The binomial and multinomial distributions are denoted $\mathrm{Binom}(n,p)$ and $\mathrm{Multinom}(n,\{p_i\}_{i=1}^k)$, respectively.

The main concentration inequality used in our upper bounds is McDiarmid's inequality.

Fact 2.1 (McDiarmid's Inequality[29]). *Let f be a multivariate function with m independent random inputs whose codomain is $\mathbb{R}$ and such that, for each $i \in [m]$, changing the ith coordinate alone can change the output by at most c_i additively. Then*

$$\mathbf{Pr}[|f(X) - \mathbf{E}[f(x)]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_i c_i^2}}.$$

A commonly used method for bounding from above the total variation distance in terms of KL divergence is Pinsker's inequality. However, Pinsker's inequality is mainly useful when the KL divergence is small. In the high probability regime, the KL divergence is larger than 1 and this gives no information about the total variation distance. Our sample complexity lower bounds instead use a different inequality, which is better suited for the high probability regime.

Fact 2.2 (see, e.g., Lemmas 2.1 and 2.6 of [35]). *For any pair of distributions p, q, we have that $d_{\mathrm{TV}}(p,q) \leq 1 - (1/2)e^{-D(p||q)}$. Equivalently, it holds $D(p||q) \geq \log(2/\delta)$, where $1-\delta$ is the total variation distance.*

3 SAMPLE-OPTIMAL CLOSENESS TESTER

In this section, we give our optimal closeness tester, described in pseudo-code below.

The main result of this section is the following theorem:

Theorem 3.1. *There exists a universal constant $C > 0$ such that the following holds: When*

$$k \geq C\left(n^{2/3}\log^{1/3}(1/\delta)/\epsilon^{4/3} + \left(n^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta)\right)/\epsilon^2\right), \tag{1}$$

Algorithm Test-Closeness is an (ϵ,δ)-closeness tester in total variation distance.

To prove Theorem 3.1, we will show that the expected value of our statistic $\widetilde{Z}$ in the completeness case is sufficiently separated from the expected value of $\widetilde{Z}$ in the soundness case, and also that the value of $\widetilde{Z}$ is highly concentrated around its expectation in both cases. We proceed to prove these two steps in the following subsections. We will assume that the parameter k in Step 1 of the algorithm satisfies (1).

Algorithm 1: TEST-CLOSENESS$(p, q, n, \epsilon, \delta)$

Input : sample access to distributions p, q over $[n]$, $\epsilon > 0$, and $\delta > 0$.

Output : "YES" if $p = q$, "NO" if $d_{TV}(p, q) \geq \epsilon$; both with probability at least $1 - \delta$.

Set $k = C\left(n^{2/3}\log^{1/3}(1/\delta)/\epsilon^{4/3} + \left(n^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta)\right)/\epsilon^2\right)$, where $C > 0$ is a sufficiently large universal constant.

Set $(\widetilde{m_p}, \widetilde{m_p}', \widetilde{m_q}, \widetilde{m_q}') = \text{Multinom}\,(4k, (1/4, 1/4, 1/4, 1/4))$.

Draw two multi-sets of independent samples from p of sizes $\widetilde{m_p}, \widetilde{m_p}'$ respectively, and two multi-sets of independent samples from q of sizes $\widetilde{m_q}, \widetilde{m_q}'$ respectively. Let $\widetilde{X} = (\widetilde{X}_i)_{i=1}^n$, $\widetilde{X}' = (\widetilde{X}_i')_{i=1}^n$, $\widetilde{Y} = (\widetilde{Y}_i)_{i=1}^n$, $\widetilde{Y}' = (\widetilde{Y}_i')_{i=1}^n$ be the corresponding histograms of the samples.

Compute the value of the random variable $\widetilde{Z} = \sum_{i=1}^n \widetilde{Z}_i$, where, for $i \in [n]$, we define

$$\widetilde{Z}_i = |\widetilde{X}_i - \widetilde{Y}_i| + |\widetilde{X}_i' - \widetilde{Y}_i'| - |\widetilde{X}_i - \widetilde{X}_i'| - |\widetilde{Y}_i - \widetilde{Y}_i'| \,.$$

Set the threshold $T = C'\sqrt{k\log(1/\delta)}$, where C' is a universal constant (derived from the analysis of the algorithm).

if $\widetilde{Z} \leq T$ **then**

  return "YES"

else

  return "NO"

end

3.1 Bounding the Expectation Gap

In this section, we will prove an $\Omega(\sqrt{k\log(1/\delta)})$ expectation gap between the completeness and soundness cases. We proceed by analyzing the expectation of a slightly modified random variable Z obtained by taking the number of samples drawn from p and q be Poisson distributed. We then relate the expectation of Z to the expectation of our actual statistic $\widetilde{Z}$.

Definition of modified random variable Z. Independently set $m_p = \text{Poi}(k)$, $m_p' = \text{Poi}(k)$, $m_q = \text{Poi}(k)$, $m_q' = \text{Poi}(k)$. Draw two multi-sets of independent samples from p of sizes m_p, m_p' respectively, and two multi-sets of independent samples from q of sizes m_q, m_q' respectively. Let $X = (X_i)_{i=1}^n$, $X' = (X_i')_{i=1}^n$, $Y = (Y_i)_{i=1}^n$, $Y' = (Y_i')_{i=1}^n$ be the corresponding histograms of the samples. We will analyze the random variable

$$Z = \sum_{i=1}^n Z_i, \text{ where } Z_i = |X_i - Y_i| + |X_i' - Y_i'| - |X_i - X_i'| - |Y_i - Y_i'| \,. \tag{2}$$

Let $m = m_p + m_p' + m_q + m_q'$ be the total number of samples drawn from p, q in the definition of Z. By construction, we have that $\widetilde{Z} = Z \mid (m = 4k)$. This will allow us to argue that $\mathbf{E}[Z]$ and $\mathbf{E}[\widetilde{Z}]$ are close to each other.

CLAIM 3.2. *We have that* $|\mathbf{E}[Z] - \mathbf{E}[\widetilde{Z}]| = O(\sqrt{k})$.

PROOF. Note that the statistic Z is 2-Lipschitz, i.e., adding a sample can change Z by at most 2. Therefore, $|\mathbf{E}[Z \mid m = a] - \mathbf{E}[Z \mid m = b]| \leq 2|a - b|$. This implies that

$$|\mathbf{E}[Z] - \mathbf{E}[\widetilde{Z}]| = O(\mathbf{E}[|m - 4k|]) = O(\sqrt{k}) \,,$$

as desired. □

It therefore suffices to show that there is sufficient separation between $\mathbf{E}[Z]$ in the completeness and soundness cases. Specifically, this subsection is devoted to the proof of the following lemma:

LEMMA 3.3 (EXPECTATION GAP). *Let Z be the statistic defined in (2). Then*

(i) *If $p = q$ (completeness), we have that $\mathbf{E}[Z] = 0$.*

(ii) *If $d_{TV}(p, q) \geq \epsilon$ (soundness), then $\mathbf{E}[Z] = \Omega(\sqrt{k\log(1/\delta)})$.*

Note that for each $i \in [n]$, $X_i, X_i' \sim \text{Poi}(kp_i)$, $Y_i, Y_i' \sim \text{Poi}(kq_i)$. Moreover, the random variables $\{X_i, X_i', Y_i, Y_i'\}_{i=1}^n$ are mutually independent.

The proof of Part (i) in Lemma 3.3 is straightforward and holds for all $k \geq 1$. Since $p = q$, it follows that, for any fixed $i \in [n]$, the random variables X_i, X_i', Y_i, Y_i' are identically distributed. Therefore, the random variables $|X_i - Y_i|$, $|X_i' - Y_i'|$, $|X_i - X_i'|$, and $|Y_i - Y_i'|$ are also identically distributed, which implies that $\mathbf{E}[|X_i - Y_i|] = \mathbf{E}[|X_i' - Y_i'|] = \mathbf{E}[|X_i - X_i'|] = \mathbf{E}[|Y_i - Y_i'|]$. Thus, $\mathbf{E}[Z_i] = 0$ for all $i \in [n]$, and therefore $\mathbf{E}[Z] = 0$.

The proof of Part (ii) in Lemma 3.3 is significantly more challenging. We note that the proof of Part (ii) crucially relies on the assumption that k is sufficiently large, satisfying (1).

We start with the following technical claim:

CLAIM 3.4. *For all $i \in [n]$, we have that*

$$\mathbf{E}[Z_i] = \Omega\left(\min\left\{|kp_i - kq_i|, |kp_i - kq_i|^2, \frac{|kp_i - kq_i|^2}{\sqrt{kp_i + kq_i}}\right\}\right) \,. \tag{3}$$

PROOF. Recall that for each $i \in [n]$, $X_i, X_i' \sim \text{Poi}(kp_i)$, $Y_i, Y_i' \sim \text{Poi}(kq_i)$ and that these random variables are mutually independent. This implies that $\mathbf{E}[|X_i - Y_i|] = \mathbf{E}[|X_i' - Y_i'|]$ and therefore

$$\mathbf{E}[Z_i] = 2\,\mathbf{E}[|X_i - Y_i|] - \mathbf{E}[|X_i - X_i'|] - \mathbf{E}[|Y_i - Y_i'|] \,.$$

Due to the absolute values in the above expression, we can assume without loss of generality that $a := kp_i \geq kq_i =: b$.

Let $c := a - b \geq 0$. Then we can write that $X_i, X_i' \sim \text{Poi}(b) + \text{Poi}(c)$ and $Y_i, Y_i' \sim \text{Poi}(b)$. Let B_1, B_2 and C_1, C_2 be mutually independent random variables with $B_1, B_2 \sim \text{Poi}(b)$ and $C_1, C_2 \sim \text{Poi}(c)$. Note that $B_\ell + C_{\ell'}$, for $\ell, \ell' \in \{1, 2\}$, have the same distribution as X_i and X_i'. By linearity of expectation, we can thus write

$$\begin{aligned}\mathbf{E}[Z_i] = (1/2)\,\mathbf{E}\Big[&|B_1 + C_1 - B_2| + |B_1 + C_2 - B_2| + \\ &+ |B_1 - C_1 - B_2| + |B_1 - C_2 - B_2| - |B_1 + C_1 - B_2 - C_2| - \\ &- |B_1 + C_2 - B_2 - C_1| - 2|B_1 - B_2|\Big] \,,\end{aligned} \tag{4}$$

where the first four terms above correspond to $2\,\mathbf{E}[|X_i - Y_i|]$, the fifth and sixth terms correspond to $-\mathbf{E}[|X_i - X_i'|]$, and the last term corresponds to $-\mathbf{E}[|Y_i - Y_i'|]$.

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) = (1/2)\,(|x + y| + |y - x| - 2|y|) \,.$$

By the definition of f and (4), we have that

$$\mathbf{E}[Z_i] = \mathbf{E}\left[f(C_1, B_1 - B_2) + f(C_2, B_1 - B_2) - f(C_1 - C_2, B_1 - B_2)\right] . \tag{5}$$

Now observe that $f(x, y) = \max\{0, |x| - |y|\}$ and that $f(x, y)$ is an increasing function of $|x|$.

For any $x_1, x_2 \geq 0$ and $y \in \mathbb{R}$, we have that $|x_1 - x_2| \leq \max\{x_1, x_2\}$, hence

$$\begin{aligned}(x_1 - x_2, y) = f(|x_1 - x_2|, y) &\leq f(\max\{x_1, x_2\}, y) \\ &= \max\{f(x_1, y), f(x_2, y)\} .\end{aligned}$$

This implies that

$$\begin{aligned}f(x_1, y) + f(x_2, y) - f(x_1 - x_2, y) \geq & f(x_1, y) + f(x_2, y) \\ & - \max\{f(x_1, y), f(x_2, y)\} \\ = & \min\{f(x_1, y), f(x_2, y)\} \\ = & f(\min\{x_1, x_2\}, y) .\end{aligned}$$

Using (5), the above inequality gives that

$$\begin{aligned}\mathbf{E}[Z_i] &\geq \mathbf{E}\left[f(\min\{C_1, C_2\}, B_1 - B_2)\right] \\ &= \mathbf{E}\left[\max\left\{0, \min\{C_1, C_2\} - |B_1 - B_2|\right\}\right] .\end{aligned} \tag{6}$$

Therefore, it suffices to establish a lower bound on the RHS of (6). We proceed to do so by considering two complementary cases, based on the value of the parameter $c \geq 0$.

Case I: $c < 1$.

In this case, we can write

$$\begin{aligned}\mathbf{E}[Z_i] &\geq \mathbf{Pr}\left[(\min\{C_1, C_2\} \geq 1) \wedge (B_1 = B_2)\right] \\ &= \mathbf{Pr}[C_1 \geq 1]^2 \mathbf{Pr}[B_1 = B_2] \\ &\geq \Omega\left(c^2 \min\left\{1, 1/\sqrt{b}\right\}\right) = \Omega\left(\min\left\{c^2, c^2/\sqrt{b}\right\}\right) ,\end{aligned}$$

where the first inequality follows from (6) (since $\min\{C_1, C_2\} - |B_1 - B_2| \} \geq 1$ under the corresponding event), the first equality uses the independence of B_1, B_2 and C_1, and the last inequality uses the fact that $\mathbf{Pr}[C_1 \geq 1] = 1 - e^{-c} \geq c/2$ (since $0 \leq c < 1$) and that $\mathbf{Pr}[B_1 = B_2] = \Omega(\min\{1, 1/\sqrt{b}\})$. To prove the latter lower bound, we will use the fact that B_1, B_2 are i.i.d. and that their common distribution B is supported on integers and has standard deviation $\sigma = \sqrt{b}$. By Chebyshev's inequality, we have that $\mathbf{Pr}\left[|B - b| = O(\sigma)\right] \geq 1/2$. Since B is has integer support, there exists a set of integers S with cardinality $|S| \leq 1 + O(\sigma)$ such that $\mathbf{Pr}[B \in S] \geq 1/2$. Now note that $\mathbf{Pr}[B_1 = B_2] = \sum_{i \geq 0} \mathbf{Pr}[B = i]^2 \geq \sum_{i \in S} \mathbf{Pr}[B = i]^2 \geq (1/|S|)\mathbf{Pr}[B \in S]^2 \geq 1/(4|S|)$, where the second inequality follows by the convexity of the quadratic function. Therefore, $\mathbf{Pr}[B_1 = B_2] = \Omega(1/(1 + O(\sigma))) = \Omega\left(\min\{1, 1/\sigma\}\right)$, as desired.

Case II: $c \geq 1$.

In this case, there exists a universal constant $\delta_0 > 0$ such that $\delta_0 = \mathbf{Pr}[\min\{C_1, C_2\} \geq c/2]$. We will show that $\mathbf{Pr}[|B_1 - B_2| \leq c/4] = \Omega(\min\{1, c/\sqrt{b}\})$. Using (6), the latter inequality implies that

$$\begin{aligned}\mathbf{E}[Z_i] &\geq (c/4)\,\mathbf{Pr}\left[\min\{C_1, C_2\} \geq c/2\right]\,\mathbf{Pr}[|B_1 - B_2| \leq c/4] \\ &= (c/4)\,\delta_0\,\Omega(\min\{1, c/\sqrt{b}\}) \\ &= \Omega\left(\min\{c, c^2/\sqrt{b}\}\right) .\end{aligned}$$

To establish the desired upper bound on $\mathbf{Pr}[|B_1 - B_2| \leq c/4]$, we apply the argument from Case I for the random variables $B_i' = \lfloor B_i/(c/4) \rfloor$, $i = 1, 2$. Note that the B_i' is an integer-valued random variable with standard deviation $\sigma' = 1 + O(\sqrt{b}/c)$, and therefore $\mathbf{Pr}[B_1' = B_2'] = \Omega(\min\{1, 1/\sigma'\}) = \Omega(\min\{1, c/\sqrt{b}\})$. Finally, we note that $\mathbf{Pr}[|B_1 - B_2| \leq c/4] \geq \mathbf{Pr}[B_1' = B_2']$. This completes Case II.

Recall that $c = |kp_i - kq_i|$ by definition. The proof of Claim 3.4 is now complete. □

Proof of Lemma 3.3 (ii). Suppose that $d_{TV}(p, q) \geq \epsilon$. For each bin $i \in [n]$, we assign i to set S_1, S_2, S_3 if

$$\min\left\{|kp_i - kq_i|, |kp_i - kq_i|^2, \frac{|kp_i - kq_i|^2}{\sqrt{kp_i + kq_i}}\right\}$$

is equal to $|kp_i - kq_i|$, $|kp_i - kq_i|^2$, or $\frac{|kp_i - kq_i|^2}{\sqrt{kp_i + kq_i}}$ respectively (breaking ties arbitrarily). This defines a partition of $[n]$ into three sets, S_1, S_2, S_3. Since $\sum_{i=1}^n |p_i - q_i| \geq \epsilon/2$, for at least one $j \in \{1, 2, 3\}$ we have that $\sum_{i \in S_j} |p_i - q_i| \geq \epsilon/6$. In each of these three cases, we will use Claim 3.4 to prove the desired expectation lower bound.

Case 1: $\sum_{i \in S_1} |p_i - q_i| \geq \epsilon/6$**.**

In this case, we have that $\mathbf{E}[Z] = \sum_{i=1}^n \mathbf{E}[Z_i] \geq \sum_{i \in S_1} \mathbf{E}[Z_i] = \Omega(k) \sum_{i \in S_1} |p_i - q_i| = \Omega(\epsilon k)$. Since k is assumed to satisfy (1) and in particular we have that $k \geq C \log(1/\delta)/\epsilon^2$, it follows that $\mathbf{E}[Z] = \Omega(\sqrt{k \log(1/\delta)})$, as desired.

Case 2: $\sum_{i \in S_2} |p_i - q_i| \geq \epsilon/6$**.**

In this case, we have that $\mathbf{E}[Z] = \sum_{i=1}^n \mathbf{E}[Z_i] \geq \sum_{i \in S_2} \mathbf{E}[Z_i] = \Omega(k^2) \sum_{i \in S_2} |p_i - q_i|^2 = \Omega(k^2\epsilon^2/n)$, where the last inequality follows from Cauchy-Schwarz and the fact that $|S_2| \leq n$. Since k is assumed to satisfy (1) and in particular $k \geq Cn^{2/3} \log^{1/3}(\delta)/\epsilon^{4/3}$, it follows that $\mathbf{E}[Z] = \Omega(\sqrt{k \log(1/\delta)})$, as desired.

Case 3: $\sum_{i \in S_3} |p_i - q_i| \geq \epsilon/6$**.** In this case, we can similarly write that

$$\begin{aligned}\mathbf{E}[Z] = \sum_{i=1}^n \mathbf{E}[Z_i] \geq \sum_{i \in S_3} \mathbf{E}[Z_i] &= \Omega(k^{3/2}) \sum_{i \in S_3} \frac{(p_i - q_i)^2}{(p_i + q_i)^{1/2}} \\ &= \Omega\left(k^{3/2}\epsilon^2/n^{1/2}\right) ,\end{aligned}$$

where the last bound follows from our assumption that $\sum_{i \in S_3} |p_i - q_i| \geq \epsilon/6$ and a careful application of the generalized Holder's inequality. Recall that for any triple of vectors $x, y, z \in \mathbb{R}^m$, we have that $\sum_i |x_i y_i z_i| \leq \|x\|_r \|y\|_s \|z\|_t$, where $1/r + 1/s + 1/t = 1$. Using this fact, we can write

$$\begin{aligned}\sum_{i \in S_3} |p_i - q_i| &= \sum_{i \in S_3} \frac{|p_i - q_i|}{(p_i + q_i)^{1/4}} (p_i + q_i)^{1/4} \\ &\leq \left(\sum_{i \in S_3} \frac{(p_i - q_i)^2}{(p_i + q_i)^{1/2}}\right)^{1/2} \left(\sum_{i \in S_3} (p_i + q_i)\right)^{1/4} \left(\sum_{i \in S_3} 1^4\right)^{1/4}\end{aligned}$$

where we used $x = \left(\frac{|p_i - q_i|}{(p_i + q_i)^{1/4}}\right)_{i \in S_3}$, $y = ((p_i + q_i)^{1/4})_{i \in S_3}$, $z = (1)_{i \in S_3}$, and $r = 2, s = t = 4$. Since $\sum_{i \in S_3} (p_i + q_i) \leq 2$ and $|S_3| \leq n$, we get that $\sum_{i \in S_3} \frac{(p_i - q_i)^2}{(p_i + q_i)^{1/2}} = \Omega(\epsilon^2/n^{1/2})$, as desired.

We have thus shown that $\mathbf{E}[Z] = \Omega(k^{3/2}\epsilon^2/n^{1/2})$. Since k is assumed to satisfy (1) and in particular $k \geq Cn^{1/2}\log^{1/2}(\delta)/\epsilon^2$, it follows that $\mathbf{E}[Z] = \Omega(\sqrt{k\log(1/\delta)})$, as desired.

This completes the proof of Lemma 3.3 (ii). □

3.2 Concentration of Test Statistic: Proof of Theorem 3.1

By Lemma 3.3, we have that $\mathbf{E}[Z] = 0$ in the completeness case and $\mathbf{E}[Z] = \Omega(\sqrt{k\log(1/\delta)})$ in the soundness case respectively. Combined with Claim 3.2, we have that in the completeness case $\mathbf{E}[\widetilde{Z}] = O(\sqrt{k})$ and in the soundness case $\mathbf{E}[\widetilde{Z}] = \Omega(\sqrt{k\log(1/\delta)})$.

The random variable $\widetilde{Z}$ depends on $4k$ inputs: the choice, for each of the $4k$ samples, of which distribution to be drawn from and which coordinate to land in. $\widetilde{Z}$ is 2-Lipschitz in these $4k$ inputs. An application of McDiarmid's inequality to $\widetilde{Z}$ gives that

$$\mathbf{Pr}\left[\left|\widetilde{Z} - \mathbf{E}[\widetilde{Z}]\right| \geq C'\sqrt{k\log(1/\delta)}\right] < 2e^{-2\frac{(C'\sqrt{(k\log(1/\delta))})^2}{4k\cdot 4}} = 2\delta^{(C')^2/8} = 2\delta^{C'''}$$

for some constant C'''. If we apply the variable substitution $\delta \leftarrow (\delta/2)^{1/C'''}$, the RHS above becomes δ and the number of samples only changes by a constant factor. Therefore, our tester is correct with probability at least $1-\delta$, as desired.

4 SAMPLE-OPTIMAL INDEPENDENCE TESTER

4.1 Intuition and Setup

The goal in independence testing is to distinguish between p and $q = p_x \times p_y$, i.e., the product of the marginal distributions of p on the two coordinates. Unfortunately, we cannot simply use our closeness tester to solve this problem, as the sample complexity would contain an $(nm)^{2/3}\log^{1/3}(1/\delta)/\epsilon^{4/3}$ term, which is sub-optimal even for constant δ. Instead, we must take advantage of the fact that q is a product distribution.

This issue is solved in the large δ case in [16] by flattening. The idea is that the error in their test statistic can be reduced if q is guaranteed to have small ℓ_2 norm. To achieve this, we use flattening to split up the heavy bins. This can be done especially effectively for product distributions, as we can use samples to identify the heavy bins in the marginals rather than having to individually identify all of the heavy bins in the product.

To make a technique like this work in our context, there are several obstacles that must be overcome. The first is that we need to know how flattening can be used to improve the concentration bounds on our test statistic Z which is defined later in this subsection. To see why this might be the case, we will observe that any bins with only a single sample do not contribute to Z, and thus do not contribute to its variance. In fact, with some extra work we can prove stronger concentration bounds on Z that depend on the number N of non-isolated samples. As distributions with small ℓ_2 norm will likely produce fewer non-isolated samples, this will hopefully improve our concentration bounds.

Unfortunately, while the basic flattening technique [16] works in the large δ regime, it does not work with high probability. To overcome this issue, we note that the goal of our flattening is actually not to produce a distribution with small ℓ_2 norm, but to ensure that the number of collisions among the samples used to compute Z is relatively small. For this we note that if we are given a fixed pool S of samples from which we draw samples both for the purposes of flattening and for computing Z, it can be shown that no matter what S is, there is always a good probability that the samples to compute Z have few collisions. The overall strategy for our tester will be to take this fixed set of samples and repeatedly try different subdivisions into flattening and testing samples until we find one that works.

The most basic unit of our tester will be an algorithm called BasicTest, which runs one iteration of this strategy and returns one of "YES", "NO", or "ABORT", with the last meaning that our attempt at flattening has failed and needs to be repeated.

Algorithm 2: FullTest($\overline{S}$): Given a distribution p over $[n]\times[m]$ (where $n \geq m$), test if p is a product distribution.

Input : Sample access to a 2-dimensional distribution p over $[n]\times[m]$

Output : "YES" if $p \in \mathcal{P}$, "NO" if $\inf_{q\in\mathcal{P}} d_{\mathrm{TV}}(p,q) \geq \epsilon$, where $\mathcal{P}$ is the set of product distributions, both with probability at least $1-\delta$.

1 $k \leftarrow C\left(n^{2/3}\log^{1/3}(1/\delta)/\epsilon^{4/3} + \left(n^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta)\right)/\epsilon^2\right)$, where $C > 0$ is a sufficiently large universal constant.
2 $S \leftarrow 100k$ samples from p.
3 $result \leftarrow$ ABORT
4 **while** $result = ABORT$ **do**
5 $\quad result \leftarrow$ BasicTest(S)
6 **end**
7 return $result$

The main result of this section is the following theorem:

Theorem 4.1. *There exists a universal constant $C > 0$ such that the following holds: When*

$$k \geq C\left(\frac{n^{2/3}m^{1/3}\log^{1/3}(1/\delta)}{\epsilon^{4/3}} + \frac{(nm)^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta)}{\epsilon^2}\right), \tag{7}$$

Algorithm FullTest is an (ϵ,δ)-independence tester in total variation distance.

Setup. Our independence testing procedure BasicTest has the following basic structure:

(1) Choose a large multiset set of samples $\overline{S}$.
(2) Choose from $\overline{S}$ a flattening $F = (F_x, F_y)$, and possibly ABORT.
(3) Choose from $\overline{S}$ a set S of "flattened" samples, and possibly ABORT.
(4) Use S to compute a test statistic Z.
(5) Accept or reject based on the test statistic.

At various points in the process, the algorithm may choose to ABORT (for example, if the number of non-singletons in S is 100× more than expected). We will show that if the algorithm is run on a random set $\overline{S}$ of samples, the probability of outputting a wrong

answer is $O(\delta)$, but that for *any* set $\overline{S}$ of samples the chance of aborting is at most $1/2$. Therefore, when we abort, we can start over from Step 2, and repeat until we output "YES" or "NO", without increasing the sample complexity and with only $O(\delta)$ failure probability.

Flattening. Flattening involves choosing a set F of samples from the distribution p with marginals p_x and p_y. We then flatten the rows and columns of p independently, giving us a new distribution p^f with marginals p_1^f and p_2^f. The following definition appears as Definition 2.4 in [16] and describes a subdivision of the domain of a distribution p that aims at reducing its ℓ_2 norm. For this transformation to be useful to us, we need to always make sure that the domain size does not increase by more than a constant factor as a result.

Definition 4.2 ([16])**.** Given a distribution p on $[n]$ and a multiset S of elements from $[n]$, we define the *split* distribution p_S over $[n + |S|]$ as follows: For $1 \le i \le n$, let f_i be the number of times element i appears in S, and $a_i = 1 + f_i$. Our new distribution p_S is supported on the set $B = \{(i, j) : i \in [n], 1 \le j \le a_i\}$. In order to get a sample (i, j) from p_S, we first draw i according to p and then j uniformly at random from $[a_i]$.

Note the following fact about split distributions:

Fact 4.3. *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then: (i) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively. (ii) It holds that $\|p_S - q_S\|_1 = \|p - q\|_1$.*

When we are dealing with multidimensional distributions, it will be useful to have a definition of flattening only on a specific marginal. The definition below is given for 2-dimensional distributions, but it can be easily generalized.

Definition 4.4. Given a distribution p on $[n]\times[m]$ with marginals p_x and p_y. Also let S be a multiset of elements from $[n]$ (respectively $[m]$), we define the *row-split* (respectively *column-split*) distribution p_S over $[n + |S|] \times [m]$ (respectively $[n] \times [m + |S|]$) as follows: in order to get a sample $((i, k), j)$ (respectively $(i, (j, k))$) from the *row-split* (respectively *column-split*) distribution p_S, we first draw (i, j) according to p and then independently draw k uniformly at random from $[a_i]$ (respectively $[a_j]$).

Test Statistic. Define the product distribution $q^f := p_1^f \times p_2^f$. Note that $d_{\mathrm{TV}}(p^f, q^f) = d_{\mathrm{TV}}(p, q)$ where $q = p_x \times p_y$. Therefore, the goal of determining whether p is a product distribution or far from it is equivalent to distinguishing between $p^f = q^f$ and p^f far from q^f. In addition to sampling from p^f, we can sample q^f by taking two samples from p^f: we combine the first coordinate of the first sample with the second coordinate of the second sample.

The sample set S consists of four pieces:

- S_{p0}, S_{p1}: two sets of $\mathrm{Poi}(k)$ samples from p^f.
- S_{q0}, S_{q1}: two sets of $\mathrm{Poi}(k)$ samples from q^f.

We let $X_u^{(p0)}$ denote the number of times element u appears in S_{p0}, and similarly for the other three sets. For each u in the range of q^f we get the test statistic:

$$Z_u := |X_u^{(p0)} - X_u^{(q0)}| + |X_u^{(p1)} - X_u^{(q1)}| - |X_u^{(p0)} - X_u^{(p1)}| - |X_u^{(q0)} - X_u^{(q1)}|$$

Our final test statistic is the sum of this:

$$Z := \sum_u Z_u.$$

Note that, if a given item u appears exactly once in the entire set S of samples, then $Z_u = 0$. We say that such a sample is a *singleton*, and define $N \le |S|$ to be the number of non-singleton samples.

4.2 Concentration of Z

The goal of this section is to prove that the test statistic Z concentrates. We will show this happens for any setting of the flattening F, and ignoring the possibility of ABORT (that is, if we ran even aborted procedures to completion). In particular, our goal is the following lemma:

Lemma 4.5. *For a fixed flattening F and any $\delta > 0$, there exists a constant $C > 0$ such that*

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| > C \cdot \sqrt{(N + \log(1/\delta))\log(1/\delta)}] \le \delta.$$

Intuitively, the idea is that since singletons do not change the statistic, the variance—and concentration—of Z should depend on the number of non-singletons N rather than the total number of samples k. Note that the concentration is relative to N, which is also a random variable.

We show this using *symmetrization*. For the sake of analysis we introduce an independent copy of the statistic Z', generated from another set S' of samples. Let $T = S \cup S'$ be the set of all samples used by Z and Z', and let M be the number of non-singletons in T.

Note that we could generate these same variables in a different way: rather than first generating S_{p0} and S'_{p0} with $\mathrm{Poi}(k)$ samples each and setting $T_{p0} = S_{p0} \cup S'_{p0}$, we can instead first sample T_{p0} with $\mathrm{Poi}(2k)$ samples, then randomly assign each sample in T_{p0} to one of S_{p0} and S'_{p0} (and similarly for $p1, q0, q1$). These are equivalent generative processes. This second process leads to the following lemma:

Lemma 4.6. *For every possible T, and any $\delta > 0$,*

$$\mathbf{Pr}[|Z - Z'| > \sqrt{8M\log(2/\delta)} \mid T] \le \delta.$$

Proof. We apply McDiarmid's inequality, and use the alternative generative process. Conditioned on $T = (T_{p0}, T_{p1}, T_{q0}, T_{q1})$, the only randomness lies in whether each sample v is placed in S or S'. Let c_v be the maximum amount that $|Z - Z'|$ can change by when $v \in T$ is switched between S and S'. Switching v can only change Z by at most 2, and similarly for Z', so $c_v \le 4$. Moreover, if v is a singleton in T, then switching v has zero effect on Z or Z', so $c_v = 0$. Hence

$$\sum_{v \in T} c_v^2 \le 16M.$$

Since Z and Z' are identically distributed, $\mathbf{E}[Z - Z' \mid T] = 0$. Therefore McDiarmid's inequality states that, for any t,

$$\mathbf{Pr}[|Z - Z'| \ge t \mid T] \le 2e^{-\frac{2t^2}{16M}}.$$

Setting t appropriately gives the result. □

Since our desired lemma is in terms of N, not M, we relate the two:

Lemma 4.7. *There exists a constant C such that, for every possible T, and any $\delta > 0$,*

$$\mathbf{Pr}[M > C(N + \log(1/\delta)) \mid T] \leq \delta.$$

Proof. We again use the alternative generative process. There are M non-singletons in T, which means we can pair them up into $M/2$ disjoint pairs of colliding elements. Each such pair has a $1/4$ chance of having both elements land in S, independent of every other pair. Let n be the number of such pairs that land entirely in S. By a Chernoff bound:

$$\mathbf{Pr}[n \leq M/16 \mid T] \leq e^{-M/C}$$

for some constant $C \geq 8$. Now, if T is such that $M \leq C\log(1/\delta)$, the lemma statement is trivially true. Otherwise, since $N \geq 2n$,

$$\mathbf{Pr}[M \geq 8N \mid T] \leq \delta$$

as desired. □

We also need to prove a constant-probability version of the result:

Lemma 4.8. *It holds that*

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| \geq C\sqrt{N+1}] \leq 1/2.$$

Proof. We will show this with Markov's inequality, by showing

$$\mathbf{E}[(Z - \mathbf{E}[Z])^2/(N+1)] = O(1) \tag{8}$$

using symmetrization. Since Z' is independent of Z, and by convexity,

$$\begin{aligned}\mathbf{E}[(Z - \mathbf{E}[Z])^2/(N+1)] &\leq \mathbf{E}[(Z - Z')^2/(N+1)] \\ &= \mathbf{E}_T[\mathbf{E}[(Z - Z')^2/(N+1)|T]]. \end{aligned} \tag{9}$$

For any fixed T, by Lemma 4.6 and Lemma 4.7 applied with $\delta/2$ and a union bound we have with probability $1 - \delta$ that both:

$$\begin{aligned}(Z - Z')^2 &\leq 8M\log(4/\delta) \\ N &\geq M/C - \log(2/\delta)\end{aligned}$$

The latter equation implies $N + 1 \geq M/(C\log(2/\delta))$, and hence

$$(Z - Z')^2/(N+1) \leq 8C\log(4/\delta)\log(2/\delta)$$

with probability $1 - \delta$. This strong concentration implies a bound in expectation:

$$\mathbf{E}[(Z - Z')^2/(N+1)|T] \leq O(C) = O(1).$$

Plugging back into (9) gives (8), which implies the result. □

We now have the tools for the main result of the section.

Proof of Lemma 4.5. Consider any two thresholds τ and τ', where τ is a random variable depending on the sampling used for Z and τ' depends on that for Z'. Because Z' is independent of Z, we have:

$$\begin{aligned}\mathbf{Pr}[|Z - \mathbf{E}[Z]| > \tau \cap |Z' - \mathbf{E}[Z]| < \tau'] = \mathbf{Pr}[|Z - \mathbf{E}[Z]| \\ > \tau]\mathbf{Pr}[|Z' - \mathbf{E}[Z]| < \tau'].\end{aligned}$$

On the other hand,

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| > \tau \cap |Z' - \mathbf{E}[Z]| < \tau'] \leq \mathbf{Pr}[|Z - Z'| > \tau - \tau'].$$

Hence

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| > \tau] \leq \mathbf{Pr}[|Z - Z'| > \tau - \tau']/\mathbf{Pr}[|Z' - \mathbf{E}[Z]| < \tau']. \tag{10}$$

We now define these two thresholds τ and τ'.

Defining τ'. By Lemma 4.8 applied to Z', with 50% probability we have

$$|Z' - \mathbf{E}[Z]| \leq O(\sqrt{N'+1}). \tag{11}$$

Define τ' to be this RHS.

By Lemma 4.7, with $1 - \delta$ probability we have

$$M = O(N + \log(1/\delta)). \tag{12}$$

(Note that we are no longer conditioning on T.) Since $N' \leq M$, this implies that there exists a constant $C > 0$such that

$$\tau' \leq C\sqrt{N + \log(1/\delta)} \tag{13}$$

with probability $1 - \delta$.

Defining τ. On the other hand, combining (12) with Lemma 4.6, with $1 - 2\delta$ probability we have

$$|Z - Z'| \leq O(\sqrt{(N + \log(1/\delta))\log(2/\delta)}).$$

We would like to define τ to be this RHS plus τ', but this would be invalid: τ must be independent of Z'. Hence we instead define τ to be this RHS plus $C\sqrt{N + \log(1/\delta)}$; by (13), this is larger than the RHS plus τ' with $1 - \delta$ probability. Hence:

$$\mathbf{Pr}[|Z - Z'| > \tau - \tau'] \leq 3\delta \tag{14}$$

for this τ, which is $O(\sqrt{(N + \log(1/\delta))\log(2/\delta)})$.

Combining the results. Plugging (14) and (11) into (10), we have for this τ that

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| > \tau] \leq 3\delta/(1/2) = 6\delta.$$

Using $\delta' = \delta/6$ gives the desired result. □

4.3 Algorithm

We begin with a helper algorithm BasicTest (Algorithm 3).

Our analysis will depend on two key facts:

(1) For *any* set of samples $\overline{S}$, the probability that BasicTest returns ABORT is at most $1/2$.
(2) If BasicTest is run on a set of i.i.d. samples from p, the probability that it returns an incorrect answer ("NO" if p is actually independent, or "YES" if p is ϵ-far from independent) is at most δ.

The latter of these points will hold because our algorithm will ABORT unless N_q is small. This, combined with Lemma 4.5 and Claim 4.14, will imply that the output is correct (along with a separate argument (see Lemma 4.12) for when $N \gg N_q$).

To show the first of these points, one can first use Markov to bound the probability of aborting due to F_x or F_y or ℓ or ℓ' being too large. The more interesting case is to show that N_q is bounded with appropriate probability. This will follow from the following lemma:

Algorithm 3: BasicTest($\overline{S}$): Given a distribution p over $[n] \times [m]$ (where $n \geq m$) test if p is a product distribution.

Input : A Multiset $\overline{S}$ of $100k$ samples from $[n] \times [m]$ with $k = C\left(\frac{n^{2/3}m^{1/3}\log^{1/3}(1/\delta)}{\epsilon^{4/3}} + \frac{\sqrt{nm\log(1/\delta)}}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}\right)$, where C is a sufficiently large universal constant.

Output : Information relating to whether these samples came from an independent distribution.

/* Choose flattening F */

$F_x, F_y \leftarrow \emptyset$

for $s \in \overline{S}$ **do**

  $F_x = F_x \cup \{s\}$ with prob $\min\{n/100k, 1/100\}$

  $F_y = F_y \cup \{s\}$ with prob $m/100k$; // note that $k > m$ always.

end

if $|F_x| > 10n$ *or* $|F_y| > 10m$ **then**

  return ABORT

end

/* Draw samples S_p^f, S_q^f */

Let $\overline{S}' = \{(x_i, y_i)\}$ be a uniformly random permutation of $\overline{S} \setminus (F_x \cup F_y)$

Draw $\ell, \ell' \sim \mathrm{Poi}(2k)$.

if $2\ell + \ell' > |\overline{S}'|$ **then**

  return ABORT

end

Let $S_q = \{(x_{2j-1}, y_{2j})\}_{j=1}^{\ell}, S_p = \{(x_j, y_j)\}_{j=2\ell+1}^{2\ell+\ell'}$

Create S_p^f, S_q^f by assigning to corresponding sub-bins uniformly at random

$N_p \leftarrow \sharp$samples in S_p^f that collide with another sample in S_p^f

$N_q \leftarrow \sharp$samples in S_q^f that collide with another in $S_p^f \cup S_q^f$.

if $N_q > c\max(k/m, k^2/mn)$ **then**

  return ABORT

end

if $N_p > 20N_q + C'\log(1/\delta)$ **then** // C' a sufficiently large constant

  return "NO"

end

/* Compute test statistic Z */

Flag each sample of S_p^f, S_q^f independently with prob. $1/2$.

Let $X_i^{(p0)}, X_i^{(q0)}$ be the number of times element i appears *flagged* in each set S_p^f, S_q^f respectively and $X_i^{(p1)}, X_i^{(q1)}$ be the corresponding counts on *unflagged* samples.

Compute the statistic $Z = \sum_i Z_i$, where $Z_i = |X_i^{(p0)} - X_i^{(q0)}| + |X_i^{(p1)} - X_i^{(q1)}| - |X_i^{(p0)} - X_i^{(p1)}| - |X_i^{(q0)} - X_i^{(q1)}|$.

if $Z < C' \cdot \sqrt{\min(k, (k^2/(mn) + k/m))\log(1/\delta)}$ **then**

  return "YES"

else

  return "NO"

end

Lemma 4.9. *For any set of samples $\overline{S}$,*

$$\mathbf{E}[N_q \mid \overline{S}] = O\left(\max\left(\frac{k^2}{nm}, k/m\right)\right),$$

where N_q is considered to be 0 in the case that the algorithm aborts before computing it.

Proof. Throughout this proof we will condition on $\overline{S}$. We note that $\ell \geq k/2$ except with probability exponentially small in k, in which case $N_q = O(k)$. Thus, the contribution from the case where $\ell < k/2$ is $O(1)$ and we can henceforth assume that $\ell \geq k/2$ (note that given the size of the parameters $k/m > 1$).

In order to bound N_q we bound it as a sum of simpler random variables whose expectations we can bound individually. For $1 \leq i \leq 100k$, we let N_i be 0 unless the i^{th} element of $\overline{S}$ is in S_p, and in that case, it is the number of elements of S_q^f that the corresponding element of S_p^f collides with (with the exception that we define N_i to be 0 if $\ell < k/2$). For $1 \leq i \neq j \leq 100k$ let $N_{i,j}$ be 0 unless one of the elements of S_q is obtained by taking the x-coordinate from the i^{th} element of $\overline{S}$ and y-coordinate from the j^{th} element of $\overline{S}$, and if so is equal to the number of other elements in S_q^f that the corresponding element of S_q^f collides with (with the exception that we define $N_{i,j}$ to be 0 if $\ell < k/2$). It is easy to see that

$$N_q \leq \sum_i N_i + \sum_{i,j} N_{i,j}. \tag{15}$$

Our final result will follow from two bounds: Firstly, for all i, we claim that

$$\mathbf{E}[N_i] = O(\max\left(\frac{k^2}{nm}, k/m\right)/k). \tag{16}$$

We also claim that for all i, j that

$$\mathbf{E}[N_{i,j}] = O(\max\left(\frac{k^2}{nm}, k/m\right)/k^2). \tag{17}$$

We begin with our proof of Equation (16) as it is slightly easier. Assume that the i^{th} element of $\overline{S}$ is (X, Y). Let C_X denote the number of other elements of $\overline{S}$ with the same x-coordinate and C_Y the number with the same y-coordinate. Upon flattening, let F_X and F_Y denote the number elements of F_x equal to X and the number of elements of F_y equal to Y, respectively. Note that F_X is distributed as a binomial distribution $\mathrm{Binom}(C_X, \min(n/100k, 1/100))$ and thus $\mathbf{E}[1/(F_X + 1)] = O(1/(C_X \min(n/k, 1)))$. Similarly, $\mathbf{E}[1/(F_Y + 1)] = O(k/(C_Y m))$.

Once we have conditioned on the flattening sets F_x and F_y, we consider $C_{X,Y}$, the number of elements of S_q equal to (X, Y), where C_{XY} is set to 0 if $\ell < k/2$ (recall that this case can safely be ignored in our final analysis). We claim that $\mathbf{E}[C_{X,Y}|F_x, F_y] \ll C_X C_Y/k$. This is because the expectation of $C_{X,Y}$ is a sum of all pairs of one of the C_X elements of S with the correct x-coordinate and one of the C_Y elements of S with the correct y-coordinate of the probability that this pair of elements is used to create an element of S_q. We claim that this probability is $O(1/k)$. In fact, this probability is at most $1/\ell$, where $\ell \geq k/2$ due to our conditioning. That is because even conditioning on ℓ and which 2ℓ elements of S are used to construct the elements of S_q, there is only an $O(1/\ell) = O(1/k)$ probability that

the two designated elements of S are adjacent to each other after the random permutation is applied.

However, once S_q is fixed, each of these $C_{X,Y}$ elements that might collide with our i^{th} element of S only do if they are mapped to the same sub-bin. This happens only with probability $1/((1+F_X)(1+F_Y))$. Therefore, we have that:

$$\mathbf{E}[N_i | C_{X,Y}, F_X, F_Y] = \frac{C_{X,Y}}{(1+F_X)(1+F_Y)}.$$

Therefore, using the fact that F_X, F_Y are independent random variables, we have that

$$\begin{aligned}\mathbf{E}[N_i] &\le \sup_{F_X,F_Y} (\mathbf{E}[C_{X,Y}|F_X,F_Y])\mathbf{E}[1/(1+F_X)]\mathbf{E}[1/(1+F_Y)] \\ &= O(C_X C_Y/k)O(\max(k/n,1)/C_X)O(k/(C_Y m)) \\ &= O(\max(k/(mn), 1/m)),\end{aligned}$$

as desired.

The proof of Equation (17) is similar. Assume that the i^{th} element of $\overline{S}$ has x-coordinate X and that the j^{th} element has y-coordinate Y. Let C_X and C_Y be the number of other elements of $\overline{S}$ with x-coordinate equal to X and y-coordinate equal to Y, respectively. Again let F_X and F_Y denote the number elements of F_x equal to X and the number of elements of F_y equal to Y, respectively. Once again $\mathbf{E}[1/(F_X+1)] = O(1/(C_X \min(n/k, 1)))$ and $\mathbf{E}[1/(F_Y+1)] = O(k/(C_Y m))$.

We now let $C_{X,Y}$ be 0 unless $\ell \ge k/2$ *and* the i^{th} and j^{th} elements pair to make an element of S_q, and in this case define it to be the number of other elements of S_q equal to (X,Y). We claim now that $\mathbf{E}[C_{X,Y}|F_x,F_y] \ll C_X C_Y/k^2$ (note that this differs from the above because of the k^2 in the denominator rather than k). This is because $C_{X,Y}$ is the sum over the $C_X C_Y$ pairs of other elements with the correct x and y values of the probability that this pair of elements of S *and* the pair of the i^{th} and j^{th} elements both end up in S_q. Even conditioning on F_x, F_y and ℓ, the probability that the random permutation of elements put the two elements of both of these pairs next to each other is $O(1/\ell^2) = O(1/k^2)$. Thus, $\mathbf{E}[C_{X,Y}|F_x,F_y] \ll C_X C_Y/k^2$.

From here the argument is the same as above. Each of these $C_{X,Y}$ elements of S_q has only a $1/((F_X+1)(F_Y+1))$ of colliding with our designated one after assigning them to random sub-bins. Thus, we have that Therefore, we have that:

$$\mathbf{E}[N_{i,j} | C_{X,Y}, F_X, F_Y] = \frac{C_{X,Y}}{(1+F_X)(1+F_Y)}.$$

And thus,

$$\begin{aligned}\mathbf{E}[N_{i,j}] &\le \sup_{F_X,F_Y} (\mathbf{E}[C_{X,Y}|F_X,F_Y])\mathbf{E}[1/(1+F_X)]\mathbf{E}[1/(1+F_Y)] \\ &= O(C_X C_Y/k^2)O(\max(k/n,1)/C_X)O(k/(C_Y m)) \\ &= O(\max(1/(mn), 1/(km))),\end{aligned}$$

as desired.

Our lemma now follows from combining Equations (15), (16) and (17). □

We are now prepared to prove the second of our main points about BasicTest.

Lemma 4.10. *For any sample multiset $\overline{S}$, the probability that BasicTest returns ABORT is at most $1/2$.*

Proof. First, consider the case that BasicTest returns ABORT in line 6, because either $|F_x| > 10n$ or $|F_y| > 10n$. Note that $F_x \sim \text{Binom}(100k, \min\{n/100k, 1/100\})$ and $F_y \sim \text{Binom}(100k, m/100k)$. Therefore, we have that: $\mathbf{E}[|F_x|] \le n$ and $\mathbf{E}[|F_y|] = m$. By applying Markov's inequality for each random variable and a union bound, we get that

$$\mathbf{Pr}[(|F_x| > 10n) \vee (|F_y| > 10n)] \le 1/5 .$$

The second possibility to return ABORT is in line 11 when $2\ell + \ell' > |\overline{S}'| \ge 100k - |F_x| - |F_y|$. Thus, we need to bound: $\mathbf{Pr}[2\ell + \ell' + |F_x| + |F_y| > 100k]$. Note that by linearity of expectation:

$$\mathbf{E}[2\ell+\ell'+|F_x|+|F_y|] = \mathbf{E}[2\ell]+\mathbf{E}[\ell']+\mathbf{E}[|F_x|]+\mathbf{E}[|F_y|] \le 4k+2k+k+k .$$

By applying Markov's inequality again, we get that:

$$\mathbf{Pr}[2\ell + \ell' + |F_x| + |F_y| > 100k] \le 8/100 .$$

It remains to bound the chance of ABORT on line 18. By Lemma 4.9 and Markov's inequality,

$$\mathbf{Pr}[N_q > c \cdot \max\{k/m, k^2/nm\}|\overline{S}] < 1/5 ,$$

for some constant c.

Using a union bound for all the above three cases, we get that the probability that BasicTest returns ABORT is at most $1/5 + 8/100 + 1/5 < 1/2$. □

For the rest of the analysis, we consider running BasicTest on a set $\overline{S}$ of random samples from some distribution p on $[n] \times [m]$. We note that we can simulate the algorithm in the following way: First, for each i from 1 to $100k$, if our algorithm wants to add an element to F_x or F_y, we generate a random element from p and add it to the appropriate set(s). If either $|F_x| > 10n$ or $|F_y| > 10m$ we abort, so we will condition on F_x and F_y for which this does not happen. Next, we generate an infinite sequence of elements (x_i, y_i) from p, and let S_q be the set of (x_{2j-1}, y_{2j}) for $j \in [1, \ell]$ and S_p the set of (x_j, y_j) for $j \in [2\ell+1, 2\ell+\ell']$. Note that conditioned on not returning ABORT, this gives sets F_x, F_y, S_p, S_q identically distributed as BasicTest. However, unconditionally, it gives an S_q and S_p sets of $\text{Poi}(2k)$ samples from $q := p_x \times p_y$ and p, respectively. Furthermore, we can compute Z, N_q and N regardless. Note that this statistic Z will be an instance of the statistic computed for our closeness tester applied to the distribution p^f and q^f. In particular, Lemmas 4.5, 4.6, 4.8 and 4.11 will still apply to it.

For the next several lemmas, we consider F_x and F_y as being fixed and Z, N_q and N being computed in this way regardless of potential aborts. In the next few lemmas, we wish to show that with high probability N will be $O(N_q)$ if p is a product distribution. This will allow us to use our bounds on N_q as bounds on N (or more precisely, allow us to reject if N is not bounded in terms of N_q).

Lemma 4.11. *For a fixed set of samples S_p^f, S_q^f, consider the distribution of Z over the partition into $p0/p1$ and $q0/q1$. We have:*

$$\mathbf{Pr}[Z < N_p/6 - 2N_q - 100] < 1/2.$$

Proof. Let $X_i^{(p)}$ denote the number of times element i appears in S_p^f, so that $N_p = \sum_{i:X_i^{(p)}>1} X_i^{(p)}$. Define the statistic $\widetilde{Z} = \sum_i \widetilde{Z}_i$,

where $\widetilde{Z}_i = X_i^{(p0)} + X_i^{(p1)} - |X_i^{(p0)} - X_i^{(p1)}| = 2\min(X_i^{(p0)}, X_i^{(p1)})$, to be the value Z would take if S_q^f were empty. Since Z is 2-Lipschitz and invariant to singletons, we have

$$|Z - \widetilde{Z}| \le 2N_q. \tag{18}$$

Hence, our goal is to show that $\widetilde{Z}$ is usually at least $N_p/4$. We have that

$$\mathbf{E}[\widetilde{Z}_i] \ge \lfloor X_i^{(p)}/2 \rfloor ,$$

because we can partition the elements i into $\lfloor X_i^{(p)}/2 \rfloor$ pairs, each of which has a $1/2$ chance of being divided between $p0$ and $p1$, and hence contributing 1 to each of $X_i^{(p0)}$ and $X_i^{(p1)}$, or 2 to $\widetilde{Z}_i$. We also have that

$$\mathbf{Var}(\widetilde{Z}_i) \le 4X_i^{(p)} ,$$

because $\widetilde{Z}_i$ is a 2-Lipschitz function of $X_i^{(p)}$ independent random choices, and of course $\mathbf{Var}(\widetilde{Z}_i) = 0$ if $X_i^{(p)} = 0$. Therefore,

$$\mathbf{E}[\widetilde{Z}] \ge N_p/3, \quad \mathbf{Var}[\widetilde{Z}] \le 4N_p .$$

By Chebyshev's inequality, this means

$$\mathbf{Pr}[\widetilde{Z} < N_p/3 - 4\sqrt{N_p}] \le 1/4, \quad \text{or}$$

$$\mathbf{Pr}[\widetilde{Z} < N_p/6 \text{ and } N_p > 600] \le 1/4.$$

Combined with (18), we have

$$\mathbf{Pr}[Z < N_p/6 - 2N_q \text{ and } N_p > 600] \le 1/4 .$$

But, of course, $\mathbf{Pr}[Z < 0] = 0$, so for all N_p we have that

$$\mathbf{Pr}[Z < N_p/6 - 2N_q - 100] \le 1/4 < 1/2 .$$

□

We can now bound the probability that we reject incorrectly on line 22.

Lemma 4.12. *If p is a product distribution, then the probability that BasicTest returns "NO" on line 22 is $O(\delta)$.*

This is essentially because if N is a sufficient multiple of N_q then by Lemma 4.11 we have that Z is likely to be at least a large multiple of N. However Lemma 3.3 says that $\mathbf{E}[Z] = 0$ and Lemma 4.5 says that $|Z - \mathbf{E}[Z]| \ll \sqrt{N\log(1/\delta)}$ with high probability.

Finally, we can bound the probability of BasicTest giving an incorrect output on lines 30 or 28.

Lemma 4.13. *If p is a product distribution, then the probability that BasicTest returns "NO" on line 30 is $O(\delta)$. Similarly, if $d_{TV}(p,q) > \epsilon$ then the probability that BasicTest returns "YES" is $O(\delta)$.*

This holds because if we reach this stage of the algorithm $N = N_p + N_q$. We know by previous checks that N_q is not too large and N_p is not much bigger than N_q. This gives us strong concentration bounds on Z and a careful analysis of the separation in expectations between the soundness and completeness cases will yield our result.

Proof. In order for the algorithm to return "YES" or "NO" on line 30, it has to avoid "aborting" or returning "NO" on line 22. Therefore, it must be the case that $N_p \le 20N_q + C'\log(1/\delta)$ and $N_q = O(\frac{k^2}{nm} + k/m)$, which implies $N = O(\frac{k^2}{nm} + k/m)$. Note also that $\frac{k^2}{nm} \ge \log(1/\delta)$,as well as $k \ge \log(1/\delta)$ by definition of k. Note also the trivial bound that $N = O(k)$.

Therefore, by Lemma 4.5, we have that:

$$\mathbf{Pr}[|Z - \mathbf{E}[Z]| > C\sqrt{\min(k, (k^2/mn + k/m))\log(1/\delta)}] \le \delta/2 ,$$

for some constant $C > 0$.

If p is a product distribution (i.e., $p = q$), then by Lemma 3.3, we have that $\mathbf{E}[Z] = 0$. Thus, the algorithm will return "NO" with probability at most $\delta/2$.

For the soundness case, where $d_{TV}(p,q) > \epsilon$, it suffices to show the following lower bound on the expected value of Z:

Claim 4.14. *If $d_{TV}(p,q) > \epsilon$, then*

$$\mathbf{E}[Z] \ge 2C'\sqrt{\min(k, (k^2/mn + k/m))\log(1/\delta)} .$$

Proof. Suppose that we condition on the flattening samples. This will determine the flattened distributions p^f, q^f. From the proof of Lemma 3.3 it follows that:

$$\mathbf{E}[Z|F_x, F_y] = \Omega\left(\min\left\{\epsilon k, \frac{k^2\epsilon^2}{|D_{p^f}|}, \frac{k^{3/2}\epsilon^2}{\sqrt{|D_{p^f}|}}\right\}\right),$$

where $|D_{p^f}| = \Theta(nm)$ is the domain size of the flattened distribution. We now distinguish the following three cases:

- **Case 1:** $\mathbf{E}[Z|F_x, F_y] = \Omega(\epsilon k)$.
  Using the fact that $k = \Omega(\log(1/\delta)/\epsilon^2)$, it follows that
  $$\mathbf{E}[Z|F_x, F_y] = \Omega(\sqrt{k\log(1/\delta)}).$$
- **Case 2:** $\mathbf{E}[Z|F_x, F_y] = \Omega\left(\frac{k^2\epsilon^2}{nm}\right)$
  - Using the fact that $k = \Omega\left(\frac{\sqrt{nm\log(1/\delta)}}{\epsilon^2}\right)$, we get that:

$$\mathbf{E}[Z|F_x, F_y] = \Omega\left(\frac{k\epsilon^2}{nm}\cdot\frac{\sqrt{nm\log(1/\delta)}}{\epsilon^2}\right) = \Omega\left(\sqrt{\frac{k^2}{nm}\log(1/\delta)}\right).$$

  - Using the fact that $k = \Omega\left(\frac{n^{2/3}m^{1/3}\log^{1/3}(1/\delta)}{\epsilon^{4/3}}\right)$, we get that:

$$\mathbf{E}[Z|F_x, F_y] = \Omega\left(\sqrt{\frac{k}{m}}\cdot\frac{\epsilon^2k^{3/2}}{n\sqrt{m}}\right) = \Omega\left(\sqrt{\frac{k}{m}}\cdot\frac{\epsilon^2n\sqrt{m\log(1/\delta)}}{\epsilon^2n\sqrt{m}}\right) = \Omega\left(\sqrt{(k/m)\log(1/\delta)}\right).$$

- **Case 3:** $\mathbf{E}[Z|F_x, F_y] = \Omega\left(\frac{k^{3/2}\epsilon^2}{\sqrt{nm}}\right)$.
  We note that this is larger than the expression in Case 2 unless $k > nm$. Thus, it suffices to show that $\frac{k^{3/2}\epsilon^2}{\sqrt{nm}} = \Omega(\sqrt{k\log(1/\delta)})$. However, this follows from the fact that $k = \Omega\left(\frac{\sqrt{nm\log(1/\delta)}}{\epsilon^2}\right)$.

Combining these two bounds, we get the required statement for any possible choice of flattening samples. Thus, the unconditional version of the statement also holds. □

This completes the proof of the lemma. □

Recall that the full algorithm is the following:

(1) Let $\overline{S}$ be a random set of $100k$ samples.

(2) Run BasicTest on $\overline{S}$ until it does not return ABORT.
(3) Return "YES"/"NO" as appropriate.

Lemma 4.15. *If $p = q$ the probability that FullTest returns "NO" is $O(\delta)$, and if $d_{TV}(p, q) > \epsilon$ the probability that it returns "YES" is $O(\delta)$.*

Proof. We bound the probability as follows:

$$\begin{aligned}
&\mathbf{Pr}[\text{FullTest incorrect}] \\
&= \sum_{t=0}^{\infty} \mathbf{Pr}[\text{BasicTest Returns ABORT } t \text{ times} \\
&\qquad\qquad \text{and then wrong output}] \\
&= \sum_{t=0}^{\infty} \mathbf{E}_{\overline{S}}[\mathbf{Pr}[\text{BasicTest returns ABORT}|\overline{S}]^t \\
&\quad \cdot \mathbf{Pr}[\text{BasicTest returns wrong output}|\overline{S}]] \\
&\leq \sum_{t=0}^{\infty} 2^{-t} \mathbf{E}_{\overline{S}}[\mathbf{Pr}[\text{BasicTest returns wrong output}|\overline{S}]] \\
&= \sum_{t=0}^{\infty} 2^{-t} \mathbf{Pr}[\text{BasicTest returns wrong output}] \\
&= 2\mathbf{Pr}[\text{BasicTest returns wrong output}] = O(\delta).
\end{aligned}$$

□

Proof of Theorem 4.1. By Lemma 4.15, we get that there exists some constant $c > 0$, such that Algorithm 2 outputs "NO" with probability at most $\delta' = c \cdot \delta$ if p is a product distribution, and outputs "YES" with probability at most δ' if $d_{TV}(p, p_x \times p_y) \geq \epsilon$. Since $\delta = \delta'/c$, the sample complexity is:

$$\Theta\left(n^{2/3} m^{1/3} \log^{1/3}(1/\delta')/\epsilon^{4/3}\right) + \\ + \Theta\left((nm)^{1/2} \log^{1/2}(1/\delta') + \log(1/\delta'))/\epsilon^2\right) .$$

as desired. □

REFERENCES

[1] J. Acharya, C. Daskalakis, and G. Kamath. 2015. Optimal Testing for Properties of Distributions. In *Proceedings of NIPS'15*.
[2] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. 2000. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*. 259–269. citeseer.ist.psu.edu/batu00testing.html
[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. 2013. Testing Closeness of Discrete Distributions. *J. ACM* 60, 1 (2013), 4.
[4] T. Batu, R. Kumar, and R. Rubinfeld. 2004. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*. 381–390.
[5] Clément L. Canonne. 2020. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library. 1–100 pages. https://doi.org/10.4086/toc.gs.2020.009
[6] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. 2017. Testing Bayesian Networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*. 370–448.
[7] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. 2017. Testing Conditional Independence of Discrete Distributions. *CoRR* abs/1711.11560 (2017). arXiv:1711.11560 http://arxiv.org/abs/1711.11560 In STOC'18.
[8] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. 2014. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *SODA*. 1193–1203.
[9] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. 2013. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*. 1833–1852.
[10] C. Daskalakis, N. Dikkala, and G. Kamath. 2019. Testing Ising Models. *IEEE Trans. Inf. Theory* 65, 11 (2019), 6829–6852.
[11] C. Daskalakis and Q. Pan. 2017. Square Hellinger Subadditivity for Bayesian Networks and its Applications to Identity Testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*. 697–703.
[12] L. Devroye and G. Lugosi. 2001. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer.
[13] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price. 2020. Optimal Testing of Discrete Distributions with High Probability. *CoRR* abs/2009.06540 (2020). https://arxiv.org/abs/2009.06540
[14] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. 2018. Sample-Optimal Identity Testing with High Probability. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018 (LIPIcs)*, Vol. 107. 41:1–41:14.
[15] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. 2019. Collision-Based Testers are Optimal for Uniformity and Closeness. *Chic. J. Theor. Comput. Sci.* 2019 (2019).
[16] I. Diakonikolas and D. M. Kane. 2016. A New Approach for Testing Properties of Discrete Distributions. In *FOCS*. 685–694. Full version available at abs/1601.05557.
[17] I. Diakonikolas, D. M. Kane, and V. Nikishkin. 2015. Optimal Algorithms and Lower Bounds for Testing Closeness of Structured Distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*.
[18] I. Diakonikolas, D. M. Kane, and V. Nikishkin. 2015. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*.
[19] I. Diakonikolas, D. M. Kane, and V. Nikishkin. 2017. Near-Optimal Closeness Testing of Discrete Histogram Distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*. 8:1–8:15.
[20] I. Diakonikolas, D. M. Kane, and J. Peebles. 2019. Testing Identity of Multidimensional Histograms. In *Conference on Learning Theory, COLT 2019*. 1107–1131.
[21] O. Goldreich. 2017. Commentary on two works related to testing uniformity of distributions. Available at http://www.wisdom.weizmann.ac.il/ oded/MC/229.html.
[22] O. Goldreich. 2020. The Uniform Distribution Is Complete with Respect to Testing Identity to a Fixed Distribution. In *Computational Complexity and Property Testing - On the Interplay Between Randomness and Computation*, O. Goldreich (Ed.). Lecture Notes in Computer Science, Vol. 12050. Springer, 152–172.
[23] O. Goldreich, S. Goldwasser, and D. Ron. 1998. Property testing and its connection to learning and approximation. *J. ACM* 45 (1998), 653–750.
[24] O. Goldreich and D. Ron. 2000. *On testing expansion in bounded-degree graphs*. Technical Report TR00-020. Electronic Colloquium on Computational Complexity.
[25] D. Huang and S. Meyn. 2013. Generalized Error Exponents for Small Sample Universal Hypothesis Testing. *IEEE Trans. Inf. Theor.* 59, 12 (Dec. 2013), 8157–8181.
[26] Y. Ingster and I. A. Suslina. 2003. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lecture Notes in Statistics, Vol. 169. Springer.
[27] I. Kim, S. Balakrishnan, and L. Wasserman. 2020. Minimax optimality of permutation tests. *CoRR* abs/2003.13208 (2020). Available at https://arxiv.org/abs/2003.13208.
[28] E. L. Lehmann and J. P. Romano. 2005. *Testing statistical hypotheses*. Springer.
[29] C. McDiarmid. 1989. *On the method of bounded differences*. Cambridge University Press, 148–188. https://doi.org/10.1017/CBO9781107359949.008
[30] M. Neykov, S. Balakrishnan, and L. Wasserman. 2020. Minimax Optimal Conditional Independence Testing. *CoRR* (2020). arXiv:math.ST/2001.03039
[31] J. Neyman and E. S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, 694-706 (1933), 289–337. https://doi.org/10.1098/rsta.1933.0009 arXiv:http://rsta.royalsocietypublishing.org/content/231/694-706/289.full.pdf+html
[32] L. Paninski. 2008. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory* 54 (2008), 4750–4755.
[33] R. Rubinfeld. 2012. Taming big probability distributions. *XRDS* 19, 1 (2012), 24–28.
[34] R. Rubinfeld and M. Sudan. 1996. Robust characterizations of polynomials with applications to program testing. *SIAM J. on Comput.* 25 (1996), 252–271.
[35] A. B. Tsybakov. 2009. *Introduction to Nonparametric Estimation*. Springer, New York, NY.
[36] G. Valiant and P. Valiant. 2014. An Automatic Inequality Prover and Instance Optimal Identity Testing. In *FOCS*.
[37] A. W. van der Vaart and J. A. Wellner. 1996. *Weak convergence and empirical processes*. Springer-Verlag, New York. With applications to statistics.