Testing Conditional Independence of Discrete Distributions

Clément L. Canonne
Stanford University
USA
ccanonne@cs.stanford.edu

Ilias Diakonikolas
University of Southern California
USA
diakonik@usc.edu

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

Alistair Stewart
University of Southern California
USA
stewart.al@gmail.com

ABSTRACT

We study the problem of testing *conditional independence* for discrete distributions. Specifically, given samples from a discrete random variable (X, Y, Z) on domain $[\Lambda_1] \times [\Lambda_2] \times [n]$, we want to distinguish, with probability at least $2/3$, between the case that X and Y are conditionally independent given Z from the case that (X, Y, Z) is ε-far, in ℓ_1-distance, from every distribution that has this property. Conditional independence is a concept of central importance in probability and statistics with a range of applications in various scientific domains. As such, the statistical task of testing conditional independence has been extensively studied in various forms within the statistics and econometrics communities for nearly a century. Perhaps surprisingly, this problem has not been previously considered in the framework of distribution property testing and in particular no tester with sublinear sample complexity is known, even for the important special case that the domains of X and Y are binary.

The main algorithmic result of this work is the first conditional independence tester with *sublinear* sample complexity for discrete distributions over $[\Lambda_1] \times [\Lambda_2] \times [n]$. To complement our upper bounds, we prove information-theoretic lower bounds establishing that the sample complexity of our algorithm is optimal, up to constant factors, for a number of settings. Specifically, for the prototypical setting when $\Lambda_1, \Lambda_2 = O(1)$, we show that the sample complexity of testing conditional independence (upper bound and matching lower bound) is

$$\Theta\Big(\max\left(n^{1/2}/\varepsilon^2, \min\Big(n^{7/8}/\varepsilon, n^{6/7}/\varepsilon^{8/7}\Big)\right)\Big).$$

To obtain our tester, we employ a variety of tools, including (1) a suitable weighted adaptation of the "flattening" technique, and (2) the design and analysis of an optimal (unbiased) estimator for the following statistical problem of independent interest: Given a degree-d polynomial $Q\colon \mathbb{R}^n \to \mathbb{R}$ and sample access to a distribution p over $[n]$, estimate $Q(p_1, \ldots, p_n)$ up to small additive error. Obtaining tight variance analyses for specific estimators of this form has been a major technical hurdle in distribution testing. As an important contribution of this work, we develop a general theory providing tight variance bounds for *all* such estimators. Our lower bounds, established using the mutual information method, rely on novel constructions of hard instances that may be useful in other settings.

CCS CONCEPTS

• **Mathematics of computing → Discrete mathematics**; **Probability and statistics**; **Multivariate statistics**; *Probabilistic algorithms*; • **Theory of computation → Streaming, sublinear and near linear time algorithms**; **Lower bounds and information complexity**; *Mathematical optimization*;

KEYWORDS

Distribution testing, property testing, probability distributions, conditional independence, discrete distributions, sublinear algorithms, hypothesis testing

ACM Reference Format:
Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2018. Testing Conditional Independence of Discrete Distributions. In *Proceedings of 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC'18).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3188745.3188756

1 INTRODUCTION

1.1 Background

Suppose we are performing a medical experiment. Our goal is to compare a binary response (Y) for two treatments (X), using data obtained at n levels of a possibly confounding factor (Z). We have a collection of observations group in strata (fixed values of Z). The stratified data are summarized in a series of 2×2 contingency tables, one for each strata. One of the most important hypotheses in this context is conditional independence of X and Y given Z. How many observations (X, Y, Z) do we need so that we can confidently test this hypothesis?

The above scenario is a special case of the following statistical problem: Given samples from a joint discrete distribution (X, Y, Z), are the random variables X, Y independent conditioned on Z? This is the problem of *testing conditional independence* – a fundamental statistical task with a variety of applications in a variety of fields,

STOC'18, June 25–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5559-9/18/06…$15.00
https://doi.org/10.1145/3188745.3188756

including medicine, economics and finance, etc. (see, e.g., [35, 43, 49] and references therein). Formally, we have the following definition:

Definition 1.1 (Conditional Independence). Let X, Y, Z be random variables over discrete domains $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively. We say that X and Y are *conditionally independent given* Z, denoted by $(X \perp Y) \mid Z$, if for all $(i, j, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ we have that: $\Pr[\, X = i, Y = j \mid Z = z \,] = \Pr[\, X = i \mid Z = z \,] \cdot \Pr[\, Y = j \mid Z = z \,]$.

Conditional independence is an important concept in probability theory and statistics, and is a widely used assumption in various scientific disciplines [17]. Specifically, it is a central notion in modeling causal relations [43] and of crucial importance in graphical modeling [39]. Conditional independence is, sometimes, a direct implication of economic theory. A prototypical such example is the Markov property of a time series process. The Markov property is a natural property in time series analysis and is broadly used in economics and finance [27]. Other examples include distributional Granger non-causality [31] — which is a particular case of conditional independence — and exogeneity [5].

Given the widespread applications of the conditional independence assumption, the statistical question of *testing* conditional independence has been studied extensively for almost a century. In 1924, R. A. Fisher [28] proposed the notion of partial correlation coefficient, which leads to Fisher's classical z-test for the case that the data comes from a *multivariate Gaussian distribution.* For discrete distributions, conditional independence testing is one of the most common inference questions that arise in the context of contingency tables [2]. In the context of graphical models, conditional independence testing is a cornerstone in the context of structure learning and testing of Bayesian networks (see, e.g., [9, 36, 37, 47] and references therein). Finally, conditional independence testing is a useful tool in recent applications of machine learning involving fairness [32].

One of the classical conditional independence tests in the discrete setting is the Cochran–Mantel–Haenszel test [13, 35], which requires certain strong assumptions about the marginal distributions. When such assumptions do not hold, a common tester used is a linear combination of χ-squared testers (see, e.g., [2]). However, even for the most basic case of distributions over $\{0, 1\}^2 \times [n]$, no finite sample analysis is known. A recent line of work in econometrics has been focusing on conditional independence testing in *continuous settings* [6, 18, 19, 29, 33, 34, 42, 44–46, 49, 51]. The theoretical results in these works are asymptotic in nature, while the finite sample performance of their proposed testers is evaluated via simulations.

In this paper, we will study the property of conditional independence in the framework of distribution testing. The field of *distribution property testing* [3] has seen substantial progress in the past decade, see [8, 30, 41] for two recent surveys and books. A large body of the literature has focused on characterizing the sample size needed to test properties of arbitrary distributions of a *given* support size. This regime is fairly well understood: for many properties of interest there exist sample-efficient testers [1, 7, 11, 12, 20–22, 24, 30, 38, 48]. Moreover, an emerging body of work has focused on leveraging *a priori* structure of the underlying distributions to obtain significantly improved sample complexities [4, 9, 14–16, 23–25].

1.2 Our Contributions

Rather surprisingly, the problem of testing conditional independence has not been previously considered in the context of distribution property testing. In this work, we study this problem for discrete distributions and provide the first conditional independence tester with sublinear sample complexity. To complement our upper bound, we also provide information-theoretic lower bounds establishing that the sample complexity of our algorithm is optimal for a number of important regimes. To design and analyze our conditional independence tester, we employ a variety of tools, including an optimal (unbiased) estimator for the following statistical task of independent interest: Given a degree-d polynomial $Q \colon \mathbb{R}^n \to \mathbb{R}$ and sample access to a distribution p over $[n]$, estimate $Q(p_1, \ldots, p_n)$ up to small additive error.

In this section, we provide an overview of our results. We start with some terminology. We denote by $\Delta(\Omega)$ the set of all distributions over domain Ω. For discrete sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, we will use $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$ to denote the property of conditional independence, i.e.,

$$\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}} := \{\, p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \;:\; \text{if } (X, Y, Z) \sim p,\ (X \perp Y) \mid Z \,\} \;.$$

We say that a distribution $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ is ε-far from $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$, if for every distribution $q \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$ we have that $\mathrm{d_{TV}}(p, q) > \varepsilon$. We study the following hypothesis testing problem:

> $\mathcal{T}(\Lambda_1, \Lambda_2, n, \varepsilon)$: Given sample access to a distribution p over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, with $|\mathcal{X}| = \Lambda_1$, $|\mathcal{Y}| = \Lambda_2$, $|\mathcal{Z}| = n$, and $\varepsilon > 0$, distinguish with probability at least $2/3$ between the following cases:
>
> - Completeness: $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$.
> - Soundness: $\mathrm{d_{TV}}(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}) \geq \varepsilon$.

Even though the focus of this paper is on testing under the total variation distance metric (or equivalently the ℓ_1-distance), we remark that our techniques yield algorithms under the mutual information metric as well, near-optimal for a wide range of parameters. The interested reader is referred to Section 6 for a short description of these implications.

The property of conditional independence captures a number of other important properties as a special case. For example, the $n = 1$ case reduces to the property of independence over $[\Lambda_1] \times [\Lambda_2]$, whose testing sample complexity was resolved only recently [22]. Arguably the prototypical regime of conditional independence corresponds to the other extreme. That is, the setting that the domains $\mathcal{X}, \mathcal{Y}$ are binary (or, more generally, of small constant size), while the domain $\mathcal{Z}$ is large. This regime exactly captures the well-studied and practically relevant setting of $2 \times 2 \times n$ contingency tables (mentioned in the motivating example of the previous section). For the setting where $\mathcal{X}, \mathcal{Y}$ are small, our tester and our sample complexity lower bound match, up to constant factors. Specifically, we prove:

Theorem 1.2. *There exists a computationally efficient tester for $\mathcal{T}(2, 2, n, \varepsilon)$ with sample complexity*

$$O\Big(\max\Big(\sqrt{n}/\varepsilon^2, \min\Big(n^{7/8}/\varepsilon, n^{6/7}/\varepsilon^{8/7}\Big)\Big)\Big) \;.$$

Moreover, this sample upper bound is tight, up to constant factors. That is, any tester for $\mathcal{T}(2, 2, n, \varepsilon)$ requires at least

$$\Omega\left(\max\left(\sqrt{n}/\varepsilon^2, \min\left(n^{7/8}/\varepsilon, n^{6/7}/\varepsilon^{8/7}\right)\right)\right)$$

samples.

To the best of our knowledge, prior to our work, no $o(n)$ sample algorithm was known for this problem. Our algorithm is quite simple: For every fixed value of $z \in [n]$, we consider the conditional distribution p_z. Note that p_z is a distribution over $\mathcal{X} \times \mathcal{Y}$. We construct an unbiased estimator Φ of the squared ℓ_2-distance of any distribution on $\mathcal{X} \times \mathcal{Y}$ from the product of its marginals. Our conditional independence tester uses this estimator in a black-box manner for each of the p_z's. In more detail, our tester computes a weighted linear combination of $\Phi(p_z)$, $z \in [n]$, and rejects if and only if this exceeds an appropriate threshold.

To obtain the required unbiased estimator of the squared ℓ_2-distance, we observe that this task is a special case of the following more general problem of broader interest: For a distribution $p = (p_1, \ldots, p_n)$ and an polynomial $Q : \mathbb{R}^n \to \mathbb{R}$, obtain an unbiased estimator for the quantity $Q(p_1, \ldots, p_n)$. We prove the following general result:

Theorem 1.3. *For any degree-d polynomial $Q : \mathbb{R}^n \to \mathbb{R}$ and distribution p over $[n]$, there exists a unique and explicit* unbiased *estimator U_N for $Q(p)$ given $N \geq d$ samples. Moreover, this estimator is linear in Q and its variance is at most*

$$\sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ 1\leq\|\mathbf{s}\|\leq d}} \left(\prod_{i=1}^{n} p_i^{s_i}\right)\left(\frac{\partial^{\|\mathbf{s}\|}Q(p)}{\partial X_1^{s_1}\ldots\partial X_n^{s_n}}\right)^2\left(\frac{(N-\|\mathbf{s}\|)!}{N!\prod_{i=1}^{n}s_i!}\right),$$

which itself can be further bounded as a function of Q^+, the degree-d polynomial obtained by taking the absolute values of all the coefficients of Q, and its partial derivatives.

We note that Theorem 1.3 can be appropriately extended to the setting where we are interested in estimating $Q(p, q)$, where p, q are discrete distributions over $[n]$ and Q is a real degree-d polynomial on $2n$ variables.

In addition to being a crucial ingredient for our general conditional independence tester, we believe that Theorem 1.3 is of broader interest. In a number of distribution testing problems, we need unbiased estimators for some specific polynomial Q of a distribution p (or a pair of distributions p, q). For example, the ℓ_2-tester of [12] (which has been used as a primitive to obtain a wide range of sample-optimal testers [22]) is an unbiased estimator for the squared ℓ_2-distance between two distributions p, q over $[n]$. While designing such an unbiased estimator may be relatively simple, its analysis is typically highly non-trivial. Specifically, obtaining tight bounds for the variance of such estimators has been a major technical hurdle in distribution testing. As an important contribution of this work, we develop a general theory providing tight variance bounds for *all* such estimators.

The conditional independence tester Theorem 1.2 straightforwardly extends to larger domains $\mathcal{X}, \mathcal{Y}$, alas its sample complexity becomes at least linear in the size of these sets. To obtain a sublinear tester for this general case, we require a number of additional conceptual and technical ideas. Our main theorem for conditional independence testing for domain $[\Lambda_1] \times [\Lambda_2] \times [n]$ is the following:

Theorem 1.4. *There exists a computationally efficient tester for $\mathcal{T}(\Lambda_1, \Lambda_2, n, \varepsilon)$ with sample complexity*

$$O\left(\max\left(\min\left(\frac{n^{7/8}\Lambda_1^{1/4}\Lambda_2^{1/4}}{\varepsilon}, \frac{n^{6/7}\Lambda_1^{2/7}\Lambda_2^{2/7}}{\varepsilon^{8/7}}\right), \frac{n^{3/4}\Lambda_1^{1/2}\Lambda_2^{1/2}}{\varepsilon}, \frac{n^{2/3}\Lambda_1^{2/3}\Lambda_2^{1/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}\Lambda_1^{1/2}\Lambda_2^{1/2}}{\varepsilon^2}\right)\right), \quad (1)$$

where we assume without loss of generality $\Lambda_1 \geq \Lambda_2$.

The expression of the sample complexity in Theorem 1.4 may seem somewhat unwieldy. In an attempt to interpret this bound, we consider several important special cases of interest:

- For $\Lambda_1 = \Lambda_2 = O(1)$, (1) reduces to the binary case for X, Y, recovering the tight bound of Theorem 1.2.
- For $n = 1$ (and $\Lambda_1 \geq \Lambda_2$), (1) recovers the optimal sample complexity of *independence testing*, i.e.,
$$\Theta\left(\max\left(\Lambda_1^{2/3}\Lambda_2^{1/3}/\varepsilon^{4/3}, \sqrt{\Lambda_1\Lambda_2}/\varepsilon^2\right)\right)$$
(see [22]).
- For $\Lambda_1 = \Lambda_2 = n$ (and $\varepsilon = \Omega(1)$), the sample complexity of (1) becomes $O(n^{7/4})$. In Theorem 1.5 below, we show that this bound is optimal as well.

We conclude with the aforementioned tight sample lower bound for constant values of ε, in the setting where all three coordinates are of approximately the same cardinality:

Theorem 1.5. *Any tester for $\mathcal{T}(n, n, n, 1/20)$ requires $\Omega(n^{7/4})$ samples.*

1.3 Some Notation

For $n \in \mathbb{N}$, we write $[n]$ for the set $\{1, \ldots, n\}$, and log for the binary logarithm. A probability distribution over discrete domain Ω is a function $p \colon \Omega \to [0, 1]$ such that $\|p\|_1 := \sum_{\omega\in\Omega} p(\omega) = 1$. We denote by $\Delta(\Omega)$ the set of all probability distributions over domain Ω. Recall that for two probability distributions $p, q \in \Delta(\Omega)$, their *total variation distance* is defined as $d_{\mathrm{TV}}(p, q) := \sup_{S\subseteq\Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{\omega\in\Omega}|p(\omega) - q(\omega)|$, i.e., $d_{\mathrm{TV}}(p, q) = \frac{1}{2}\|p - q\|_1$. Their ℓ_2-distance is the distance $\|p - q\|_2$ between their probability mass functions. Given a subset $\mathcal{P} \subseteq \Delta(\Omega)$ of distributions, the *distance from p to $\mathcal{P}$* is then defined as $d_{\mathrm{TV}}(p, \mathcal{P}) := \inf_{q\in\mathcal{P}} d_{\mathrm{TV}}(p, q)$. If $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$, we say that p is *ε-far* from $\mathcal{P}$; otherwise, it is *ε-close.*

1.4 Organization

Due to space limitations, most of the proofs and many results (including the lower bounds and testing algorithms for the general case) have been deferred to the full version of this paper [10]. The structure of this extended abstract is as follows: In Section 2, we give a detailed outline of our techniques; Section 3 provides some necessary preliminaries and notation, before we describe in Section 4 our algorithm for the case of constant $|\mathcal{X}|, |\mathcal{Y}|$ with a detailed sketch of its analysis. Section 5 then contains the details of our polynomial estimator bounds, and Section 6 outlines the generalization of our results to testing udner conditional mutual information.

2 OUR TECHNIQUES

2.1 Conditional Independence Tester for Binary $\mathcal{X}, \mathcal{Y}$

In the case where $\mathcal{X}$ and $\mathcal{Y}$ are binary, for each bin $z \in \mathcal{Z}$ we will attempt to estimate the squared ℓ_2-distance of the corresponding conditional distribution and the product of its conditional marginals. In particular, if $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ the square of $p_{00}p_{11} - p_{01}p_{10}$, where p_{ij} is the probability that $X = i$ and $Y = j$, for $Z = z$, is proportional to this difference. Since this square is a degree-4 polynomial in the samples, there is an unbiased estimator of this quantity that can be computed for any value $z \in \mathcal{Z}$ from which we have at least 4 samples. Furthermore, for values of $z \in \mathcal{Z}$ for which we have more than 4 samples, the additional samples can be used to reduce the error of this estimator. The final algorithm computes a weighted linear combination of these estimators (weighted so that the more accurate estimators from heavier bins are given more weight) and comparing it to an appropriate threshold.

The correctness of this estimator requires a rather subtle analysis. Recall that there are three different regimes of ε versus n in the optimal sample complexity and the tester achieves this bound without a case analysis. As usual, we require a bound on the variance of our estimator and a lower bound on the expectation in the soundness case.

On the one hand, a naive bound on the variance for our estimator for an individual bin turns out to be insufficient for our analysis. In particular, let p be a discrete probability distribution and $Q(p)$ a polynomial in the individual probabilities of p. Given $m \geq \deg(Q)$ independent samples from p, it is easy to see that there is a unique symmetric, unbiased estimator for $Q(p)$, which we call $U_m Q$. Our analysis will depend on obtaining tight bounds for the variance of $U_m Q$. It is not hard to show that this variance scales as $O(1/m)$, but this turns out to be insufficient for our purposes. In order to refine this estimate, we show that $\mathrm{Var}(U_m Q) = R(p)/m + O(1/m^2)$, for some polynomial R for which we devise a general formula. From this point on, we can show that for our Q (or in general any Q which is the square of a lower degree polynomial) $\mathrm{Var}(U_m Q) = O(Q(p)/m + 1/m^2)$. This provides us with a much sharper estimate on the variance of our estimator, except in cases where the mean is large enough that the extra precision is not necessary.

Another technical piece of our analysis is relating the mean of our estimator to the total variation distance of our distribution from being conditionally independent. In particular, our estimator is roughly the sum (over the $\mathcal{Z}$-bins with enough samples) of the squared ℓ_2 distance that the conditional distribution is from being independent. When much of the distance from conditionally independence comes from relatively heavy bins, this relation is a more or less standard ℓ_1/ℓ_2 inequality. However, when the discrepancy is concentrated on very light bins, the effectiveness of our tester is bounded by the number of these bins which obtain at least four samples, and a somewhat different analysis is required. In fact, out of the different cases in the performance of our algorithm, one of the boundaries is determined by a transition between the hard cases involving discrepancies supported on light bins to ones where the discrepancy is supported on heavy bins.

If the variables X and Y are no longer binary, our estimates for the discrepancy of an individual bin must be updated. In particular, we similarly use an unbiased estimator of the ℓ_2 distance between the conditional distribution and the product of its conditional marginals. We note however that variance of this estimator is large if the marginal distributions have large ℓ_2 norms. Therefore, in bins for which we have a large number of samples, we can employ an idea from [22] and use some of our samples to artificially break up the heavier bins, thus flattening these distributions. We elaborate on this case, and the required ingredients it entails, in the next subsection.

2.2 General Conditional Independence Tester

Assuming that we take at least four samples from any bin $z \in \mathcal{Z}$, we can compute an unbiased estimator for the squared ℓ_2 distance between p_z, the conditional distribution, and q_z the product of its conditional marginals. It is easy to see that this expectation is at least $\varepsilon_z^2/(|\mathcal{X}|\,|\mathcal{Y}|)$, where ε_z is the ℓ_1 distance between the conditional distribution and the closest distribution with independent X and Y coordinates. At a high level, our algorithm takes a linear combination of these bin-wise estimators (over all bins from which we got at least 4 samples), and compares it to an appropriate threshold. There is a number of key ideas that are needed so that this approach gives us the right sample complexity.

Firstly, we use the idea of *flattening*, first introduced in [22]. The idea here is that the variance of the ℓ_2 estimator is larger if the ℓ_2 norms of p and q are large. However, we can reduce this variance by artificially breaking up the heavy bins. In particular, if we have m samples from a discrete distribution of support size n, we can artificially add m bins and reduce the ℓ_2 norm of the distribution (in expectation) to at most $O(1/\sqrt{m})$. We note that it is usually not a good idea to employ this operation for $m \gg n$, as it will substantially increase the number of bins. Nor do we want to use all of our samples for flattening (since we need to use some for the actual tester). Trading off these considerations, using $\min(m/2, n)$ of our samples to flatten is a reasonable choice. We also remark that instead of thinking of p and q as distributions over $|\mathcal{X}|\,|\mathcal{Y}|$ bins, we exploit the fact that q is a two-dimensional product distribution over $|\mathcal{X}| \times |\mathcal{Y}|$. By flattening these marginal distributions independently, we can obtain substantially better variance upper bounds.

Secondly, we need to use appropriate weights for our bin-wise estimator. To begin with, one might imagine that the weight we should use for the estimator of a bin $z \in \mathcal{Z}$ should be proportional to the probability mass of that bin. This is a natural choice because heavier bins will contribute more to the final ℓ_1 error, and thus, we will want to consider their effects more strongly. The probability mass of a bin is approximately proportional to the number of samples obtained from that bin. Therefore, we might want to weight each bin by the number of samples drawn from it. However, there is another important effect of having more samples in a given bin. In particular, having more samples from a bin allows us to do more flattening of that bin, which decreases the variance of the corresponding bin-wise estimator. This means that we will want to assign more weight to these bins based on how much flattening is being done, as they will give us more accurate information about the behavior of that bin.

Finally, we need to analyze our algorithm. If the bin weights are chosen appropriately, we show that the final estimator A has

$\mathrm{Var}[A] = O(\min(n, m) + \mathbb{E}[A] + \mathbb{E}[A]^{3/2})$, where m is the number of samples we take and n is the number of bins. Furthermore, in the completeness case, we have that $\mathbb{E}[A] = 0$. In order to be able to distinguish between completeness and soundness, we need it to be the case that for all distributions ε-far from conditional independence it holds that $\mathbb{E}[A] \gg \sqrt{\min(n, m)}$. We know that if we are ε-far from conditional independence, we must have that $\sum_z \varepsilon_z w_z \gg \varepsilon$, where w_z is the probability that $Z = z$. In order to take advantage of this, we will need to separate the Z-bins into four cases based on the size of the w_z. Indeed, if we are far from conditional independence, then for at least one of these cases the sum of $\varepsilon_z w_z$ over bins of that type only will be $\gg \varepsilon$. Each of these four cases will require a slightly different analysis:

- Case 1: $w_z < 1/m$. In this case, the expected number of samples from bin z is small. In particular, the probability of even seeing 4 samples from the bin might well be small. Here, the expectation is dominated by the probability that we see enough samples from the bin.
- Case 2: $1/m < w_z < |\mathcal{X}|/m$: In this case, we are likely to get our 4 samples from the bin, but probably will get fewer than $|\mathcal{X}|$. This means that our flattening will not saturate either of the marginal distributions and we can reduce the squared ℓ_2 norm of q by a full factor of m_z (where m_z is the number of samples from this bin).
- Case 3: $|\mathcal{X}|/m < w_z < |\mathcal{Y}|/m$. In this case, we are likely to saturate our flattening over the X-marginal but not the Y-marginal. Thus, our flattening only decreases the ℓ_2 norm of the conditional distribution on that bin by a factor of $\sqrt{|\mathcal{X}|\, m_z}$.
- Case 4: $|\mathcal{Y}|/m < w_z$: Finally, in this case we saturate both the X- and Y-marginals, so our flattening decreases the ℓ_2 norm by a factor of $\sqrt{|\mathcal{X}|\,|\mathcal{Y}|}$.

Within each sub-case, the expectation of A is a polynomial in $m, |\mathcal{X}|, |\mathcal{Y}|$ multiplied by the sum over $z \in \mathcal{Z}$ of some polynomial in ε_z and w_z. We need to bound this from below given that $\sum_z \varepsilon_z w_z \gg \varepsilon$, and then set m large enough so that this lower bound is more than $\sqrt{\min(n, m)}$. We note that only in Case 1, is the case where $m < n$ relevant. Thus, our final bound will be a maximum over the 4 cases of the m required in the appropriate case.

2.3 Sample Complexity Lower Bound Construction for Binary $\mathcal{X}, \mathcal{Y}$

We begin by reviewing the lower bound methodology we follow: In this methodology, a lower bound is shown by adversarially constructing two distributions over pseudo-distributions (i.e., finite measures, not necessarily summing to one). Specifically, we construct a pair of ensembles D and D' of pairs of nearly-normalized pseudo-distributions such that distributions from D have the desired property and from D' are ε-far from it with high probability, and such that Poisson(s) samples from a distribution are insufficient to reliably determine from which ensemble the distribution was taken from, unless s is large enough.

To formally prove our lower bounds, we will use the mutual information method, as in [22]. In this section, we provide an intuitive description of our sample complexity lower bound for testing conditional independence, when $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $\mathcal{Z} = [n]$. (Our lower bound for the regime $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = [n]$ is proved using the same methodology, but relies on a different construction.) We construct ensembles D and D' — where draws from D are conditionally independent and draws from D' are ε-far from conditionally independent with high probability — and show that s samples from a distribution on (X, Y, Z) are insufficient to reliably distinguish whether the distribution came from D or D', when s is small. We define D and D' by treating each bin $z \in [n]$ of Z independently. In particular for each possible value $z \in [n]$ for Z, we proceed as follows: (1) With probability $\min(s/n, 1/2)$, we assign the point $Z = z$ probability mass $\max(1/s, 1/n)$ and let the conditional distribution on (X, Y) be uniform. Since the distribution is conditionally independent on these bins and identical in both ensembles, these "heavy" bins will create "noise" to confuse an estimator. (2) With probability $1 - \min(s/n, 1/2)$, we set the probability that $Z = z$ to be ε/n, and let the conditional distribution on (X, Y) be taken from either C or C', for some specific ensembles C and C'. In particular, we pick C and C' so that a draw from C is independent and a draw from C' is far from independent. These bins provide the useful information that allows us to distinguish between the two ensembles D and D'. *The crucial property is that we can achieve the above while guaranteeing that any third moment from C agrees with the corresponding third moment from C'.* This guarantee implies that if we draw 3 (or fewer) samples of (X, Y) from some bin $Z = z$, then the distribution on triples of (X, Y) will be identical if the conditional was taken from C or if it was taken from C'. That is, all information about whether our distribution came from D or D' will come from bins of type (2) for which we have at least 4 samples, of which there will be approximately $n(s\varepsilon/n)^4$. On the other hand, there will be about $\min(s, n)$ bins of type (1) with 4 samples in random configuration adding noise. Thus, we will not be able to distinguish reliably unless $n(s\varepsilon/n)^4 \gg \sqrt{\min(s, n)}$, as otherwise the "noise" due to the heavy bins will drown out the "signal" of the light ones.

To define C and C', we find appropriate vectors p, q over $\{0, 1\}^2$ so that $p + q$ and $p + 3q$ each are distributions with independent coordinates, but $p, p + 2q, p + 4q$ are not. We let C return $p + q$ and $p + 3q$ each with probability $1/2$, and let C' return $p, p + 2q$ or $p + 4q$ with probability $1/8, 3/4, 1/8$ respectively. If we wish to find the probability that 3 samples from a distribution r come in some particular pattern, we get $f(r)$ for some degree-3 polynomial f. If we want the difference in these probabilities for r a random draw from C and a random draw from C', we get $f(p+q)/2 + f(p+3q)/2 - f(p)/8 - f(p+2q)(3/4) - f(p+4q)/8$. We note that this is proportional to the fourth finite difference of a degree-3 polynomial, and is thus 0. Therefore, any combination of at most 3 samples are equally likely to show up for some Z-bin from D as from D'.

To rigorously analyze the above sketched construction, we consider drawing Poisson(s) samples from a random distribution from either D or D', and bound the shared information between the set of samples and the ensemble they came from. Since the samples from each bin are conditionally independent on the ensemble, this is at most n times the shared information coming from a single bin. By the above, the probabilities of seeing any triple of samples are the same for either D or D' and thus contribute nothing to the shared information. For sets of 4 or more samples, we note that the difference in probabilities comes only from the case where 4 samples are drawn from a bin of type (2), which happens with probability at most $O(s\varepsilon/n)^4$. However, this is counterbalanced by the fact that these sample patterns are seen with much higher frequency from bins of type (1) (as they have larger overall mass). Thus, the shared information for a combination including $m \geq 4$ samples will be $(O(s\varepsilon/n)^m)^2/\min(s/n, 1/2) \cdot \Omega(1)^m$. The contribution from $m > 4$ can be shown to be negligible, thus the total shared information summed over all bins is $O(\min(s,n) \cdot (s\varepsilon/n)^8)$. This must be $\Omega(1)$ in order to reliably distinguish, and this proves our lower bound.

3 PRELIMINARIES AND BASIC FACTS

For a distribution p we write $X \sim p$ to denote that the random variable X is distributed according to p. Finally, for $p \in \Delta(\Omega_1)$, $q \in \Delta(\Omega_2)$ we let $p \otimes q \in \Delta(\Omega_1 \times \Omega_2)$ be the product distribution with marginals p and q.

Property Testing. We work in the standard setting of distribution testing: a *testing algorithm for a property* $\mathcal{P} \subseteq \Delta(\Omega)$ is an algorithm which, granted access to independent samples from an unknown distribution $p \in \Delta(\Omega)$ as well as distance parameter $\varepsilon \in (0,1]$, outputs either accept or reject, with the following guarantees.

- if $p \in \mathcal{P}$, then it outputs accept with probability at least $2/3$;
- if $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$, then it outputs reject with probability at least $2/3$.

The two measures of interest here are the *sample complexity* of the algorithm (i.e., the number of samples from the distribution it takes in the worst case), and its running time.

Conditional Independence. We study the problem of testing conditional independence of discrete distributions. Let X, Y, Z be random variables over discrete domains $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively. Given samples from the joint distribution of (X, Y, Z), we want to determine whether X and Y are *conditionally independent given* Z, denoted by $(X \perp Y) \mid Z$, versus ε-far in total variation distance from every distribution of random variables (X', Y', Z') such that $(X' \perp Y') \mid Z'$.

Definition 3.1 (Conditional Independence). Let X, Y, Z be random variables over discrete domains $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively. We say that X and Y are *conditionally independent given* Z, denoted by $(X \perp Y) \mid Z$, if for all $(i,j,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ we have that: $\Pr[\, X = i, Y = j \mid Z = z \,] = \Pr[\, X = i \mid Z = z \,] \cdot \Pr[\, Y = j \mid Z = z \,]$.

For discrete sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, we will denote by $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$ the property of conditional independence, i.e.,

$$\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}} := \{\, p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \ :\ \text{if } (X,Y,Z) \sim p,\ (X \perp Y) \mid Z \,\} \ .$$

Recall that we say that a distribution $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ is ε-far from $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$, if for every distribution $q \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$ we have that $d_{\mathrm{TV}}(p,q) > \varepsilon$. Fix a distribution $q \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$ of minimum total variation distance to p. Then the marginals of q on each of the three coordinates may have different distributions.

We will also define testing conditional independence with respect to a different metric, namely the *conditional mutual information* [26, 50]. For three random variables X, Y, Z as above, the conditional mutual information of X and Y with respect to Z is defined as

$$I(X;Y|Z) := \mathbb{E}_Z[(I(X;Y) \mid Z)]$$

i.e., as the expected (with respect to Z) Kullback-Leibler divergence between the distributions of $(X,Y) \mid Z$ and the product of the distributions of $(X \mid Z)$ and $(Y \mid Z)$. In this variant of the problem (considered in Section 6), we will want to distinguish $I(X;Y|Z) = 0$ from $I(X;Y|Z) \geq \varepsilon$.

Notation. Let $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. For $z \in \mathcal{Z}$, we will denote by $p_z \in \Delta(\mathcal{X} \times \mathcal{Y})$ the distribution defined by

$$p_z(i,j) := \Pr_{(X,Y,Z)\sim p}[\, X = i, Y = j \mid Z = z \,]$$

and by $p_Z \in \Delta(\mathcal{Z})$ the distribution $p_Z(z) := \Pr_{(X,Y,Z)\sim p}[\, Z = z \,]$. By definition, for any $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$, we have that $p(i,j,z) = p_Z(z) \cdot p_z(i,j)$. For $z \in \mathcal{Z}$, we will denote by $p_{z,X} \in \Delta(\mathcal{X})$ the distribution $p_{z,X}(i) = \Pr_{(X,Y,Z)\sim p}[\, X = i \mid Z = z \,]$ and $p_{z,Y} \in \Delta(\mathcal{Y})$ the distribution $p_{z,Y}(j) = \Pr_{(X,Y,Z)\sim p}[\, Y = j \mid Z = z \,]$.

We can now define the product distribution of the conditional marginals:

Definition 3.2 (Product of Conditional Marginals). Fix any $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. For $z \in \mathcal{Z}$, we define the *product of conditional marginals of* p *given* $Z = z$ to be the product distribution $q_z \in \Delta(\mathcal{X} \times \mathcal{Y})$ defined by $q_z := p_{z,X} \otimes p_{z,Y}$, i.e., $q_z(i,j) = p_{z,X}(i) \cdot p_{z,Y}(j)$. We will also denote by q the mixture of product distributions $q := \sum_{z\in\mathcal{Z}} p_Z(z) q_z \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$, i.e., $q(i,j,z) := p_Z(z) \cdot q_z(i,j)$.

Basic Facts. =
We start with the following simple lemma:

Lemma 3.3. *Let* $p, p' \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. *Then we have that*

$$d_{\mathrm{TV}}(p,p') \leq \sum_{z\in\mathcal{Z}} p_Z(z) \cdot d_{\mathrm{TV}}(p_z, p'_z) + d_{\mathrm{TV}}(p_Z, p'_Z) \ , \qquad (2)$$

with equality if and only if $p_Z = p'_Z$.

Using Lemma 3.3, we deduce the following useful corollary:

Fact 1. *If* $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ *is* ε*-far from* $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$, *then, for every* $p' \in \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$, *either (i)* $d_{\mathrm{TV}}\left(p_Z, p'_Z\right) > \varepsilon/2$, *or (ii)* $\sum_{z\in\mathcal{Z}} p_Z(z) \cdot d_{\mathrm{TV}}\left(p_z, p'_z\right) > \varepsilon/2$.

The next lemma shows a useful structural property of conditional independence that will be crucial for our algorithm. It shows that if a distribution $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ is close to being conditionally independent, then it is also close to an appropriate mixture of its products of conditional marginals, specifically distribution q from Definition 3.2:

Lemma 3.4. *Suppose* $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ *is* ε*-close to* $\mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}$. *Then,* p *is* 4ε*-close to the distribution* $q = \sum_{z\in\mathcal{Z}} p_Z(z) q_z$.

The Case $\mathcal{Z} = [n]$, $\mathcal{X} = \mathcal{Y} = \{0,1\}$. We now focus on the case that $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ for $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $\mathcal{Z} = [n]$. With the same notation we have previously established, for any $z \in [n]$ we have that

$$\begin{aligned} 2d_{\mathrm{TV}}(p_z, q_z) &= 4\,|p_z(0,0) \cdot p_z(1,1) - p_z(0,1) \cdot p_z(1,0)| \\ &= 4\,|\mathrm{Cov}[(X \mid Z = z), (Y \mid Z = z)]| \ , \end{aligned}$$

or equivalently

$$d_{\mathrm{TV}}(p_z, q_z) = 2\,|\mathrm{Cov}[(X \mid Z = z), (Y \mid Z = z)]| = \|p_z - q_z\|_2 \ . \qquad (3)$$

This expression for the total variation distance will be useful in the analysis of the lower bound constructions.

Some technical result on Poisson random variables. We state below a bound on the moments of truncated Poisson random variables, which we will rely on in our analysis.

Claim 1. *There exists an absolute constant $C > 0$ such that, for $N \sim$ Poisson(λ),*

$$\mathrm{Var}[N\mathbb{1}_{\{N\geq 4\}}] \leq C\mathbb{E}\big[N\mathbb{1}_{\{N\geq 4\}}\big]\ .$$

Moreover, one can take $C = 4.22$.

4 CONDITIONAL INDEPENDENCE TESTER: THE CASE OF CONSTANT $|\mathcal{X}|, |\mathcal{Y}|$

Let $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times z)$. In this section, we describe and analyze our sample-optimal conditional independence tester for the case that $|\mathcal{X}|, |\mathcal{Y}| = O(1)$. Our tester uses as a black-box an unbiased estimator for the ℓ_2^2-distance between a 2-dimensional distribution and the product of its marginals. Specifically, we assume that we have access to an estimator Φ with the following performance: Given N samples $s = (s_1, \ldots, s_N)$ from a distribution $p \in \Delta(\mathcal{X}\times\mathcal{Y})$, Φ satisfies:

$$\mathbb{E}[\Phi(s)] = \|p - p_{\mathcal{X}} \otimes p_{\mathcal{Y}}\|_2^2 \tag{4}$$

$$\mathrm{Var}[\Phi(s)] \leq C\left(\frac{\mathbb{E}[\Phi(s)]}{N} + \frac{1}{N^2}\right)\ , \tag{5}$$

for some absolute constant $C > 0$. Such an estimator follows as a special case of our generic polynomial estimator in Theorem 1.3 whose proof is given in Section 5.

Notation. Given $p \in \Delta(\mathcal{X}\times\mathcal{Y})$, we denote its marginal distributions by $p_{\mathcal{X}}, p_{\mathcal{Y}}$. That is, we have that $p_{\mathcal{X}} \in \Delta(\mathcal{X})$ with $p_{\mathcal{X}}(x) := \Pr_{(X,Y)\sim p}[X = x]$, $x \in \mathcal{X}$, and similarly for $p_{\mathcal{Y}}$. Then, given $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, for any $z \in \mathcal{Z}$ we will denote by q_z the product distribution $p_{z,X} \otimes p_{z,Y}$.

Let M be a Poisson(m) random variable representing the number of samples drawn from $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. Given the multi-set S of M samples drawn from p, let $S_z := \{\,(x,y) : (x,y,z) \in S\,\}$ denote the multi-set of pairs $(x,y) \in \mathcal{X}\times\mathcal{Y}$ corresponding to samples $(x,y,z) \in S$, i.e., the multi-set of samples coming from the conditional distribution p_z. For convenience, we will use the notation $\sigma_z := |S_z|$. Let $A_z := \sigma_z \cdot \Phi(S_z) \cdot \mathbb{1}_{\{\sigma_z \geq 4\}}$, for all $z \in \mathcal{Z}$. Our final estimator is

$$A := \sum_{z\in\mathcal{Z}} A_z\ .$$

We set $\varepsilon' := \frac{\varepsilon}{\sqrt{|\mathcal{X}||\mathcal{Y}|}} = \Theta(\varepsilon)$, and choose

$$m \geq \beta\max\left(\sqrt{n}/\varepsilon'^2, \min\left(n^{7/8}/\varepsilon', n^{6/7}/\varepsilon'^{8/7}\right)\right)\ , \tag{6}$$

for a sufficiently large absolute constant $\beta > 0$. Our tester outputs "accept" if $A \geq \tau$ and "reject" otherwise, where τ is selected to be $\Theta\Big(\sqrt{\min(n,m)}\Big)$. A detailed pseudo-code for the algorithm is given in Algorithm 1.

Algorithm 1 TestCondIndependence

```
Require: Parameter $n := |\mathcal{Z}|$, $\Lambda_1 := |\mathcal{X}|$, $\Lambda_2 := |\mathcal{Y}|$, $\varepsilon \in (0,1]$, and sample access to $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$.
Set $m \leftarrow \beta\max\left(\sqrt{n}/\varepsilon'^2, \min\left(n^{7/8}/\varepsilon', n^{6/7}/\varepsilon'^{8/7}\right)\right)$, where $\varepsilon' := \varepsilon/\sqrt{\Lambda_1\Lambda_2}$    ▷ $\beta \geq 1$ is a sufficiently large constant
Set $\tau \leftarrow \zeta\sqrt{\min(n,m)}$    ▷ Threshold for accepting ($\zeta > 0$ is a small constant)
Draw $M \sim$ Poisson$(m)$ samples from $p$ and let $S$ be the multi-set of samples.
for all $z \in \mathcal{Z}$ do
    Let $S_z \subseteq \mathcal{X}\times\mathcal{Y}$ be the multi-set $S_z := \{\,(x,y) : (x,y,z) \in S\,\}$.
    if $|S_z| \geq 4$ then    ▷ Enough samples to call $\Phi$
        Compute $\Phi(S_z)$.
        Set $A_z \leftarrow |S_z| \cdot \Phi(S_z)$.
    else
        Set $A_z \leftarrow 0$.
    end if
end for
if $A := \sum_{z\in\mathcal{Z}} A_z \leq \tau$ then
    return accept
else
    return reject
end if
```

4.1 Proof of Correctness

In this section, we prove correctness of Algorithm 1. Specifically, we will show that: (1) If $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ (completeness), then Algorithm 1 outputs "accept" with probability at least 2/3, and (2) If $\mathrm{d_{TV}}(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}) > \varepsilon$, then Algorithm 1 outputs "reject" with probability at least 2/3. The proof proceeds by analyzing the expectation and variance of our statistic A and using Chebyshev's inequality. We note that β, ζ are absolute constants defined in the algorithm's pseudo-code.

4.1.1 Analyzing the Expectation of A. The main result of this subsection is the following proposition establishing the existence of a gap in the expected value of A in the completeness and soundness cases:

Proposition 4.1. *We have the following: (a) If $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, then $\mathbb{E}[A] = 0$. (b) If $\mathrm{d_{TV}}\Big(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\Big) > \varepsilon$, then*

$$\mathbb{E}[A] > \gamma\min\left(m\varepsilon'^2, m^4\varepsilon'^4/8n^3\right) \geq \frac{\beta\cdot\gamma}{8}\cdot\sqrt{\min(n,m)},$$

for some absolute constant $\gamma > 0$.

The rest of this subsection is devoted to the outline of the proof of Proposition 4.1. We start by providing a convenient lower bound on the expectation of A. We prove the following lemma:

Lemma 4.2. *For $z \in \mathcal{Z}$, let $\delta_z := \|p_z - q_z\|_2$ and $\alpha_z := m \cdot p_Z(z)$. Then, we have that:*

$$\mathbb{E}[A] \geq \gamma\cdot\sum_{z\in\mathcal{Z}} \delta_z^2\min(\alpha_z, \alpha_z^4)\ . \tag{7}$$

Given the expression of $\mathbb{E}[A]$ as linear combination of the δ_z^2's, the first statement of Proposition 4.1 is immediate. Indeed, if p is conditionally independent, then all δ_z's are zero. To establish the second statement, we will require a number of intermediate lemmata. We henceforth focus on the analysis of the soundness case, i.e., we will assume that $\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}\right) > \varepsilon$. We require the following useful claim:

Claim 2. *If $\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}\mid\mathcal{Z}}\right) > \varepsilon$, then $\sum_{z\in\mathcal{Z}} \delta_z\alpha_z > 2m\varepsilon'$.*

Lemma 4.2 suggests the existence of two distinct regimes: the value of the expectation of our statistic is dominated by (1) the "heavy" elements $z \in \mathcal{Z}$ for which $\alpha_z > 1$, or (2) the "light" elements $z \in \mathcal{Z}$ for which $\alpha_z \leq 1$. Formally, let $\mathcal{Z}_H := \{\, z \in \mathcal{Z} \;:\; \alpha_z > 1 \,\}$ and $\mathcal{Z}_L := \{\, z \in \mathcal{Z} \;:\; \alpha_z \leq 1 \,\}$, so that

$$\sum_{z\in\mathcal{Z}} \delta_z^2 \min(\alpha_z, \alpha_z^4) = \sum_{z\in\mathcal{Z}_H} \delta_z^2\alpha_z + \sum_{z\in\mathcal{Z}_L} \delta_z^2\alpha_z^4 \,. \tag{8}$$

By Claim 2, at least one of the following two cases holds: (1) $\sum_{z\in\mathcal{Z}_H} \delta_z\alpha_z >$ $m\varepsilon'$ or (2) $\sum_{z\in\mathcal{Z}_L} \delta_z\alpha_z > m\varepsilon'$. We analyze each case separately, establishing overall that

$$\mathbb{E}[A] > \gamma \min\left(m\varepsilon'^2, m^4\varepsilon'^4/8n^3\right) ,$$

which gives the first inequality of Proposition 4.1 (b). To complete the proof of the proposition, it suffices to show that

$$\min\left(m\varepsilon'^2, m^4\varepsilon'^4/n^3\right) \gg \sqrt{\min(n,m)} \,.$$

We show this by considering the two ranges for ε, $0 < \varepsilon' \leq 1/n^{1/8}$ and $1/n^{1/8} \leq \varepsilon' \leq 1$. For each range, we recall our setting of m and analyze the cases that could arise in the above expression (depending on the two min's). This completes the proof of Proposition 4.1. □

4.1.2 *Analyzing the Variance of A.*

We establish an upper bound on the variance of A as a function of its expectation:

Proposition 4.3. *For some absolute constant $C'' > 0$, we have*

$$\mathrm{Var}[A] \leq C''\left(\min(n,m) + \mathbb{E}[A]\right) \,. \tag{9}$$

This subsection is devoted to the proof sketch of Proposition 4.3. By the law of total variance, we have that:

$$\mathrm{Var}\, A = \mathbb{E}[\mathrm{Var}[A \mid \sigma]] + \mathrm{Var}\,\mathbb{E}[\, A \mid \sigma \,] \,.$$

We will proceed to bound each term from above, which will give the result. We start with the first term. Conditioned on $\sigma_z = |S_z|$, Eq. (5) gives that $\mathrm{Var}[A_z \mid \sigma_z] \leq C\sigma_z^2\left(\frac{\delta_z^2}{\sigma_z} + \frac{1}{\sigma_z^2}\right)\mathbb{1}_{\{\sigma_z\geq 4\}} = C\,(1 + \mathbb{E}[\, A_z \mid \sigma_z \,])\,\mathbb{1}_{\{\sigma_z\geq 4\}}$. Hence, for $\sigma := (\sigma_z)_{z\in\mathcal{Z}}$, we can write

$$\mathrm{Var}[A \mid \sigma] \leq C\left(\min(n,M) + \mathbb{E}[\, A \mid \sigma \,]\right) ,$$

where we used the inequality $\sum_{z\in\mathcal{Z}} \mathbb{1}_{\{\sigma_z\geq 4\}} \leq \sum_{z\in\mathcal{Z}} \mathbb{1}_{\{\sigma_z\geq 1\}} \leq \min(n,M)$. From this we readily get

$$\mathbb{E}[\mathrm{Var}[A \mid \sigma]] \leq C\left(\min(n,m) + \mathbb{E}[A]\right) ,$$

as desired. We now proceed to bound the second term. As shown in Lemma 4.2, $\mathbb{E}[\, A \mid \sigma \,] = \sum_{z\in\mathcal{Z}} \sigma_z\delta_z^2 \mathbb{1}_{\{\sigma_z\geq 4\}}$. By the independence of the σ_z's, we obtain that

$$\mathrm{Var}\left[\mathbb{E}[\, A \mid \sigma \,]\right] = \sum_{z\in\mathcal{Z}} \delta_z^4 \,\mathrm{Var}[\sigma_z \mathbb{1}_{\{\sigma_z\geq 4\}}] \,. \tag{10}$$

From (10) and Claim 1, recalling that $\delta_z \leq 2$, $z \in \mathcal{Z}$, we get that

$$\mathrm{Var}\left[\mathbb{E}[\, A \mid \sigma \,]\right] \leq 4C' \sum_{z\in\mathcal{Z}} \delta_z^2 \mathbb{E}\left[\sigma_z \mathbb{1}_{\{\sigma_z\geq 4\}}\right] = 4C'\mathbb{E}[A] \,.$$

This completes the proof of Proposition 4.3.

4.1.3 *Completing the Proof.*

Recall that the threshold of the algorithm is defined to be $\tau := \zeta\sqrt{\min(n,m)}$. In the completeness case, by Proposition 4.1 (a), we have that $\mathbb{E}[A] = 0$. Proposition 4.3 then gives that $\mathrm{Var}[A] \leq C'' \cdot \min(n,m)$. Therefore, by Chebyshev's inequality we obtain

$$\Pr[\, A \geq \tau \,] \leq \frac{\mathrm{Var}[A]}{\tau^2} \leq \frac{1}{\zeta^2} C'' \frac{\min(n,m)}{\min(n,m)} \leq \frac{1}{3} \,,$$

where the last inequality follows by choosing the constant ζ to be sufficiently small (compared to C'').

In the soundness case, by Chebyshev's inequality and recalling the lower bound on $\mathbb{E}[A]$ from Proposition 4.1 (b) (which implies $\tau \leq 8\frac{\zeta}{\beta\gamma}\mathbb{E}[A] \leq \mathbb{E}[A]/2$ as long as ζ is chosen sufficiently small) we get

$$\begin{aligned}\Pr[\, A < \tau \,] &\leq \Pr[\, |A - \mathbb{E}[A]| \geq \mathbb{E}[A]/2 \,] \leq 4\frac{\mathrm{Var}[A]}{\mathbb{E}[A]^2} \\ &\leq 4C''\left(\frac{\min(n,m)}{\mathbb{E}[A]^2} + \frac{1}{\mathbb{E}[A]}\right) \leq \frac{1}{3} \,,\end{aligned}$$

where the third inequality uses Proposition 4.3 and the fourth inequality uses Proposition 4.1 (b), assuming β is sufficiently large. This completes the proof of correctness. □

5 ESTIMATION OF A POLYNOMIAL IN p

In this section, we consider the following general problem: "given a degree-d n-variate polynomial $Q \in \mathbb{R}_d[X_1, \ldots, X_n]$ and access to i.i.d. samples from a distribution $p \in \Delta([n])$, how to estimate $Q(p) = Q(p_1, \ldots, p_n)$ to an additive error ε?"

In particular, we will analyze an *unbiased* estimator for $Q(p)$, and provide quantitative bounds on its variance. (Due to space constraints, we do not provide here the proof all results stated in this section, which can be found in the full version of this paper.)

Remark 1 (Reduction to homogeneous polynomials). It is sufficient to consider, without loss of generality, the case where $Q \in \mathbb{R}_d[X_1, \ldots, X_n]$ is a *homogeneous* polynomial, i.e., a sum of monomials of total degree exactly d. This is because otherwise one can multiply any monomial of total degree $d' < d$ by $\left(\sum_{i=1}^n X_i\right)^{d-d'}$: since $\sum_{i=1}^n p_i = 1$, this does not affect the value of $Q(p)$.

Based on the above remark, we hereafter assume Q is a homogeneous polynomial of degree d. Before stating the results, we will need to set some notation. Given a multi-set S of independent samples from a distribution $p \in \Delta([n])$, we let Φ_S denote the *fingerprint* of S, i.e., the vector $(\Phi_{S,1}, \ldots, \Phi_{S,n}) \in \mathbb{N}^n$ of counts: $\sum_{i=1}^n \Phi_{S,i} = |S|$, and $\Phi_{S,i}$ is the number of occurrences of i in S. Moreover, for a vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, we write X^α for the monomial $X^\alpha := \prod_{i=1}^n X_i^{\alpha_i}$, $\|\alpha\|$ for the ℓ_1 norm $\sum_{i=1}^n \alpha_i$, and $\binom{\|\alpha\|}{\alpha}$ for the multinomial coefficient $\frac{\|\alpha\|!}{\alpha_1!\cdots\alpha_n!}$. Finally, for any integer $d \geq 0$ we denote by $\mathcal{H}_d \subseteq \mathbb{R}_d[X_1, \ldots, X_n]$ the set of homogeneous degree-d n-variate polynomials.

Proposition 5.1 (Existence). *For every $N \geq d$, there exists an* unbiased linear *estimator for $Q(p)$, i.e., a linear function*

$$U_N : \mathbb{R}_d[X_1, \dots, X_n] \to \mathbb{R}_d[X_1, \dots, X_n]$$

such that $\mathbb{E}[U_N Q(\Phi_S)] = Q(p)$, where S is obtained by drawing N independent samples from p.

Proof. Fix $N \geq d$. Since we aim for a linear operator, it is enough to define it on all monomials X^α for $\|\alpha\| \leq d$. Let

$$U_N X^\alpha := \binom{N}{\alpha, N - \|\alpha\|}^{-1} \prod_{i=1}^{n} \binom{X_i}{\alpha_i}. \tag{11}$$

Note that $\deg U_N X^\alpha \leq \|\alpha\| \leq d$, so that indeed we have $U_N X^\alpha \in \mathbb{R}_d[X_1, \dots, X_n]$. Moreover, for S obtained from N independent samples from a distribution $p \in \Delta([n])$,

$$\mathbb{E}\left[U_N X^\alpha(\Phi_S)\right] = \binom{N}{\alpha, N - \|\alpha\|}^{-1} \mathbb{E}\left[\prod_{i=1}^{n} \binom{\Phi_{S,i}}{\alpha_i}\right].$$

However, since Φ_S is distributed according to a multinomial distribution with parameters N and $p_1, \dots p_n$,

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{n} \binom{\Phi_{S,i}}{\alpha_i}\right] &= \sum_{\|\beta\|=N} \binom{N}{\beta} p^\beta \prod_{i=1}^{n} \binom{\beta_i}{\alpha_i} \\
&= \sum_{\substack{\|\beta\|=N \\ \beta \geq \alpha}} \binom{N}{\beta} p^\beta \binom{N}{\beta}^{-1} \frac{N!}{\prod_{i=1}^{n} \alpha_i! (\beta_i - \alpha_i)!} \\
&= \frac{N!}{\prod_{i=1}^{n} \alpha_i!} \sum_{\substack{\|\beta\|=N \\ \beta \geq \alpha}} p^\beta \prod_{i=1}^{n} \frac{1}{(\beta_i - \alpha_i)!} \\
&= \binom{N}{\alpha, N - \|\alpha\|} \sum_{\substack{\|\beta\|=N \\ \beta \geq \alpha}} p^\beta (N - \|\alpha\|)! \prod_{i=1}^{n} \frac{1}{(\beta_i - \alpha_i)!} \\
&= \binom{N}{\alpha, N - \|\alpha\|} p^\alpha \sum_{\substack{\|\beta\|=N \\ \beta \geq \alpha}} p^{\beta - \alpha} \binom{N - \|\alpha\|}{\beta - \alpha} \\
&= \binom{N}{\alpha, N - \|\alpha\|} p^\alpha \sum_{\|\gamma\| = N - \|\alpha\|} p^\gamma \binom{N - \|\alpha\|}{\gamma} \\
&= \binom{N}{\alpha, N - \|\alpha\|} p^\alpha
\end{aligned}$$

the last equality recognizing the sum of the probability mass function of a multinomial distribution with parameters $N - \|\alpha\|$ and $p_1, \dots p_n$. This shows that $\mathbb{E}[U_N X^\alpha(\Phi_S)] = p^\alpha$; by linearity, we conclude that our estimator is indeed unbiased, i.e., $\mathbb{E}[U_N Q(\Phi_S)] = Q(p)$ for all $Q \in \mathbb{R}_d[X_1, \dots, X_n]$. □

Proposition 5.2 (Uniqueness). *This unbiased estimator is unique: that is, for every $N \geq d$, for any estimator $V_N : [n]^N \to \mathbb{R}$ satisfying*

$$\mathbb{E}[V_N(S)] = Q(p),$$

where S is a multiset of N independent samples drawn from p, one must have $V_N(S) = U_N Q(\Phi_S)$ for all S.

Proof. First, we can assume without loss of generality that V_N is a function of the fingerprint only, instead of the multiset of N samples itself. This is an immediate consequence of the fact that this fingerprint is a sufficient statistic (or, more elementary, that since all permutations of the samples are equally likely, one can consider instead $V'_N := \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_N} V_N \circ \sigma$). Therefore, we assume from now on that our estimator is of the form $V_N : \mathbb{N}^n \to \mathbb{R}$, and is restricted to inputs summing to N.

For every k, N, let U_N^k be the mapping from degree-d homogeneous polynomials to the set of their unbiased estimators on N samples.[1] We first show that it is sufficient to establish uniqueness only for the case $d = N$, i.e., to show that U_d^d maps polynomials to singletons. To argue this is enough, suppose $N > d$, and with have two different N-sample estimators V_N, W_N for a homogeneous degree-d polynomial Q. Considering $R := \left(\sum_{i=1}^{n} X_i\right)^{N-d} Q$ which is homogeneous of degree N and agrees with Q on every probability distribution p, we obtain two different N-sample estimators V_N, W_N for a homogeneous degree-N polynomial.

To prove the base case, we first describe a bijection φ (which will turn out to be U_d^d) between the set $\mathcal{E}_d$ of d-sample estimators and that of homogeneous polynomial $\mathcal{H}_d$. Specifically, given a polynomial $Q \in \mathcal{H}_d$ written as a weighted sum of degree-d monomials, $Q = \sum_{\|\alpha\|=d} c_\alpha X^\alpha$, we let $\varphi(Q)$ be the estimator whose value on a multiset S of d samples is

$$\varphi(Q)(\Phi_S) := \sum_{\|\alpha\|=d} \frac{c_\alpha}{\binom{d}{\alpha}} \mathbb{1}_{\{\Phi_S = \alpha\}} \tag{12}$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$. In particular, it immediately follows from the definition of the multinomial distribution that $\mathbb{E}[\varphi(Q)(\Phi_S)] = Q(p)$, when S is a multiset of d independent samples drawn from p:

$$\begin{aligned}
\mathbb{E}[\varphi(Q)(\Phi_S)] &= \sum_{\|\alpha\|=d} \frac{c_\alpha}{\binom{d}{\alpha}} \Pr[\,\Phi_S = \alpha\,] = \sum_{\|\alpha\|=d} \frac{c_\alpha}{\binom{d}{\alpha}} \binom{d}{\alpha} \prod_{i=1}^{n} p_i^{\alpha_i} \\
&= \sum_{\|\alpha\|=d} c_\alpha \prod_{i=1}^{n} p^\alpha
\end{aligned}$$

It is also clear that $\varphi \colon \mathcal{H}_d \to \mathcal{E}_d$ is a bijection.

Suppose now by contradiction that we have two different d-sample estimators $V_d, W_d \in \mathcal{E}_d$ for a single homogeneous polynomial $Q \in \mathcal{H}_d$. As then $\varphi^{-1}(V_d) \neq \varphi^{-1}(W_d)$, we may assume without loss of generality that $\varphi^{-1}(V_d) \neq Q$, which implies that V_d is an unbiased estimator for two distinct degree-d homogeneous polynomials, namely Q and $R := \varphi^{-1}(V_d)$.

In turn, we get that for every $p \in \Delta([n])$, $Q(p) = \mathbb{E}_S[V_d(\Phi_S)] = R(p)$; hence there difference $D := Q - R$ is a non-zero homogeneous degree-d polynomial which vanishes on every point $(x_1, \dots, x_n) \in \mathbb{N}^n$ with $\sum_{i=1}^{n} x_i = 1$. By homogeneity, for every non-zero $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}_+^n$,

$$D(\mathbf{x}) = \|\mathbf{x}\|_1^d D\!\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_1}\right) = \|\mathbf{x}\|_1^d \cdot 0 = 0$$

[1] The notation U_N^k, reminiscent of our linear estimator U_N, is not innocuous: indeed, after uniqueness is established we will see that U_N^k is the restriction of U_N (where $U_N Q$ is viewed as the singleton $\{U_N Q\}$) to degree-d homogeneous polynomials.

and therefore D vanishes on the whole non-negative quadrant $\mathbb{R}_+^n = \{\, \mathbf{x} \in \mathbb{R}^n \;:\; x_i \ge 0 \text{ for all } i \,\}$. Being identically zero on an open set, D must be the zero polynomial, leading to a contradiction. □

The above shows existence and uniqueness of an unbiased estimator, provided the number of samples N is at least the degree d of the polynomial (in p) we are trying to estimate. The proposition below shows this is necessary: if $N < d$, there is no unbiased estimator in general.

Proposition 5.3. *Let $Q \in \mathcal{H}_d$ be a homogeneous n-variate polynomial such that $\sum_{k=1}^n X$ does not divide Q. Then, there exists no unbiased estimator for $Q(p)$ from N samples unless $N \ge d$.*

Proof. Suppose by contradiction that, for such a $Q \in \mathcal{H}_d$, there exists an unbiased estimator for $Q(p)$ with $N < d$ samples. Then, since U_N^N (with the notation of the proof of Proposition 5.2) is invertible, this estimator is also an unbiased estimator for some homogeneous degree-N polynomial $R \in \mathcal{H}_N$. Therefore, it is also a unbiased estimator for the degree-d homogeneous polynomial $R' := R \cdot (\sum_{k=1}^n X_k)^{d-N} \in \mathcal{H}_d$. But by Proposition 5.2 one must then have $Q = R'$, which is impossible as $\sum_{k=1}^n X$ does not divide Q. □

Having established existence and uniqueness of our unbiased estimator, it remains to bound its variance.

Theorem 5.4. *Fix $N \ge d$, and let the mapping $U_N : \mathbb{R}_d[X_1, \dots, X_n] \to \mathbb{R}_d[X_1, \dots, X_n]$ be as above. Then, for every $Q \in \mathcal{H}_d$,*

$$\mathbb{E}\left[(U_N Q(\Phi_S))^2\right] = \sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|\le d}} \binom{\|\mathbf{s}\|}{\mathbf{s}} \frac{p^{\mathbf{s}}}{\|\mathbf{s}\|!^2} \left(\frac{d^{\|\mathbf{s}\|}Q(p)}{dX^{\mathbf{s}}}\right)^2 \binom{N-d}{d-\|\mathbf{s}\|}\binom{N}{\|\mathbf{s}\|, d-\|\mathbf{s}\|, N-d}^{-1} \tag{13}$$

where the expectation is over S obtained by drawing N independent samples from p.

By the above theorem, in order to analyze the variance of the estimator $\operatorname{Var} U_N Q(\Phi_S) = \mathbb{E}\left[(U_N Q(\Phi_S))^2\right] - \mathbb{E}[U_N Q(\Phi_S)]^2$, one needs to bound the different terms of

$$\begin{aligned}\mathbb{E}&\left[(U_N Q(\Phi_S))^2\right] \\ &= \sum_{h=0}^{d} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|=h}} \binom{h}{\mathbf{s}} p^{\mathbf{s}} \left(\frac{d^h Q(p)}{dX^{\mathbf{s}}}\right)^2 \binom{N-d}{d-h}\binom{N}{h, d-h, N-d}^{-1} \frac{1}{h!^2} \\ &= \sum_{h=0}^{d} T_h(Q,p,d,N)\end{aligned}$$

letting $T_h(Q,p,d,N)$ denote the inner sum for a given $0 \le h \le d$. In the rest of this section, we provide some useful bounds on some of these terms. First, we show that the first term will be mostly taken care of in the variance by the subtracted squared expectation, $\mathbb{E}[U_N Q(\Phi_S)]^2 = Q(p)^2$:

Claim 3. $T_0(Q,p,d,N) - Q(p)^2 = -Q(p)^2\left(\frac{d^2}{N} + O_d\left(\frac{1}{N^2}\right)\right) \le 0.$

(For our applications, the non-positivity will be enough as we only seek to upper bound the variance.)

In view of bounding the rest of the terms, let $Q^+ \in \mathcal{H}_d$ denote the polynomial obtained from Q by making all its coefficients non-negative: that is, if $Q = \sum_{\|\alpha\|=d} c_\alpha X^\alpha$, then $Q^+ := \sum_{\|\alpha\|=d} |c_\alpha| X^\alpha$. Then, we show the following:

Lemma 5.5. *Fix any $0 \le g \le d$. Then,*

$$\sum_{h=g}^{d} T_h(Q,p,d,N) = O\left(\frac{1}{N^g}\right) 2^d Q^+(p) \max_{\mathbf{s}:\|\mathbf{s}\|\ge g} \left|\frac{d^h Q(p)}{dX^{\mathbf{s}}}\right| .$$

5.1 Specific Case of Interest: ℓ_2 Distance Between p and $p_{\mathcal{X}} \otimes p_{\mathcal{Y}}$

We now instantiate the results of Section 5 to a case of interest, the polynomial Q corresponding to the ℓ_2 distance between a bivariate discrete distribution and the product of its marginals. In more detail, for any distribution $p \in \Delta(\mathcal{X} \times \mathcal{Y})$ where $|\mathcal{X}| = \Lambda_1$, $|\mathcal{Y}| = \Lambda_2$ (without loss of generality, we identify $\mathcal{X}$ and $\mathcal{Y}$ to $[\Lambda_1]$ and $[\Lambda_2]$ respectively), we let $p_\Pi := p_{\mathcal{X}} \otimes p_{\mathcal{Y}} \in \Delta(\mathcal{X} \times \mathcal{Y})$ be the product of its marginals. Moreover, define the degree-4 $(\Lambda_1\Lambda_2)$-variate polynomial $Q \in \mathbb{R}_4[X_{1,1}, X_{2,1}, \dots, X_{\Lambda_1,1}, X_{\Lambda_1,2}, \dots, X_{\Lambda_1,\Lambda_2}]$ as

$$\begin{aligned}Q&(X_{1,1}, \dots, X_{\Lambda_1,\Lambda_2}) \\ &:= \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2} \left(X_{i,j}\sum_{i'\neq i}\sum_{j'\neq j} X_{i',j'} - \sum_{i'\neq i} X_{i',j} \sum_{j'\neq j} X_{i,j'}\right)^2 .\end{aligned} \tag{14}$$

(An explicit expression for its unbiased estimator $U_N Q(\Phi_S)$ will be given in Eq. (15)). Specifically, we shall prove the following result:

Proposition 5.6. *Let Q be as in Eq. (14), and suppose that $b \ge \max(\|p\|_2^2, \|p_\Pi\|_2^2)$. Then, for $N \ge 4$,*

$$\operatorname{Var} U_N Q(\Phi_S) = O\left(\frac{Q(p)\sqrt{b}}{N} + \frac{b}{N^2}\right).$$

For consistency of notation with the previous section, we let $n := \Lambda_1\Lambda_2$ in what follows.

Claim 4. *For any p over $\mathcal{X} \times \mathcal{Y}$, we have $Q(p) = \|p - p_\Pi\|_2^2$.*

Firstly, we compute $U_N Q$ explicitly. By linearity of U_N, we can compute the unbiased estimator for each term separately, after writing $Q(X) = \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2} \Delta_{ij}(X)^2$ where

$$\Delta_{ij}(X) := X_{i,j}\sum_{i'\neq i}\sum_{j'\neq j} X_{i',j'} - \sum_{i'\neq i} X_{i',j} \sum_{j'\neq j} X_{i,j'} .$$

Now $U_N Q = U_N \Delta_{ij}^2$ and we want to compute $U_N \Delta_{ij}^2$. Note that the sums in $\Delta_{ij}(X)$ are over disjoint sets of X_ij's whose union is every X_ij. We can consider Δ_{ij} as a polynomial over the probabilities of a distribution with support of size 4, which consists of the events given by whether the marginal X is equal to i, and whether the marginal Y is equal to j. By uniqueness of the unbiased estimator, $U_N \Delta_{ij}^2$ is the same on this distribution of support 4 as on the original $\Lambda_1\Lambda_2$-size support distribution. Formally, we will write

$$\Delta_{ij}(X) := X_{i,j}X_{-i,-j} - X_{i,-j}X_{-i,j}$$

where $X_{-i,-j} := \sum_{i'\neq i}\sum_{j'\neq j} X_{i',j'}$, $X_{-i,j} := \sum_{i'\neq i} X_{i',j}$, and $X_{i,-j} := \sum_{j'\neq j} X_{i,j'}$. Squaring gives

$$\Delta_{ij}(X)^2 = X_{i,j}^2 X_{-i,-j}^2 + X_{i,-j}^2 X_{-i,j}^2 - X_{i,j}X_{-i,-j}X_{i,-j}X_{-i,j} ;$$

and it remains to apply U_N to each of these terms. We see that

$$\frac{N!}{(N-4)!}U_N X_{i,j}X_{-i,-j}X_{i,-j}X_{-i,j} = \Phi_{S,i,j}\Phi_{S,-i,-j}\Phi_{S,i,-j}\Phi_{S,-i,j}\,,$$
$$\frac{N!}{(N-4)!}U_N X_{i,j}^2 X_{-i,-j}^2 = \Phi_{S,i,j}(\Phi_{S,i,j}-1)\Phi_{S,-i,-j}(\Phi_{S,-i,-j}-1)\,,$$
$$\frac{N!}{(N-4)!}U_N X_{-i,j}^2 X_{i,-j}^2 = \Phi_{S,-i,j}(\Phi_{S,-i,j}-1)\Phi_{S,i,-j}(\Phi_{S,i,-j}-1)\,.$$

These counts are similarly summed so that, for example, $\Phi_{S,i,-j} = \sum_{j'\neq j}\Phi_{S,i,j'}$. Adding these together, we get that:

$$\begin{aligned}\frac{N!}{(N-4)!}U_N Q(\Phi_S)) &= \frac{N!}{(N-4)!}\sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2} U_N\Delta_{ij}(\Phi_S))^2\\ &= \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2}\big(\Phi_{S,i,j}(\Phi_{S,i,j}-1)\Phi_{S,-i,-j}(\Phi_{S,-i,-j}-1)\\ &\quad + \Phi_{S,-i,j}(\Phi_{S,-i,j}-1)\Phi_{S,i,-j}(\Phi_{S,i,-j}-1)\\ &\quad - 2\Phi_{S,i,j}\Phi_{S,-i,-j}\Phi_{S,i,-j}\Phi_{S,-i,j}\big)\\ &= \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2}\Big((\Phi_{S,i,j}\Phi_{S,-i,-j}-\Phi_{S,-i,j}\Phi_{S,-i,-j})^2\\ &\quad + \Phi_{S,i,j}\Phi_{S,-i,-j}(1-\Phi_{S,i,j}-\Phi_{S,-i,-j})\\ &\quad + \Phi_{S,-i,j}\Phi_{S,i,-j}(1-\Phi_{S,-i,j}-\Phi_{S,i,-j})\Big)\end{aligned} \tag{15}$$

where $\Phi_{S,-i,-j} := \sum_{i'\neq i}\sum_{j'\neq j}\Phi_{S,i',j'}$, $\Phi_{S,-i,j} = \sum_{i'\neq i}\Phi_{S,i',j}$ and $\Phi_{S,i,-j} = \sum_{j'\neq j}\Phi_{S,i,j'}$. This yields the explicit formula for our unbiased estimator of $Q(p)$.

We then turn to bounding its variance. From Theorem 5.4, we then have that, for $N \geq 4$,

$$\mathbb{E}\big[(U_N Q(\Phi_S))^2\big] = \sum_{h=0}^{4}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|=h}}\binom{h}{\mathbf{s}}\frac{p^{\mathbf{s}}}{h!^2}\left(\frac{d^hQ(p)}{dX^{\mathbf{s}}}\right)^2\binom{N-4}{4-h}\binom{N}{h,4-h,N-4}^{-1} \tag{16}$$

The rest of this section is devoted to bounding this quantity. For $h \in \{0,\dots,4\}$, we let $T_h(N)$ be the inner sum corresponding to h, so that $\mathbb{E}\big[(U_N Q(\Phi_S))^2\big] = \sum_{h=0}^{4}T_h(N)$.

For clarity, we (re-)introduce some notation: that is, we write $Q(X) = \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2}\Delta_{ij}(X)^2$ where $\Delta_{ij}(X) := X_{i,j}\sum_{i'\neq i}\sum_{j'\neq j}X_{i',j'} - \sum_{i'\neq i}X_{i',j}\sum_{j'\neq j}X_{i,j'}$ as before. Each Δ_{ij} is a degree-2 polynomial, with partial derivatives

$$\frac{\partial\Delta_{ij}}{\partial X_{k,\ell}} = \begin{cases} X_{i,j} & \text{if } k\neq i, \ell\neq j\\ \sum_{i'\neq i}\sum_{j'\neq j}X_{i',j'} & \text{if } k=i, \ell=j\\ -\sum_{i'\neq i}X_{i',j} & \text{if } k=i, \ell\neq j\\ -\sum_{j'\neq j}X_{i,j'} & \text{if } k\neq i, \ell=j\end{cases}$$

and

$$\frac{\partial^2\Delta_{ij}}{\partial X_{k,\ell}\partial X_{k',\ell'}} = (\delta_{ik}-\delta_{ik'})(\delta_{j\ell}-\delta_{j\ell'})\,.$$

- The first contribution, for $h=0$, is $O\big(Q(p)^2/N\big)$ by Claim 3 so we have T_0 under control: indeed,

$$Q(p) \leq 2\sqrt{b}$$

by the triangle inequality and the definition of b; so that $T_0(N) - Q(p)^2 = O\Big(Q(p)\sqrt{b}/N\Big)$.
- The second, $h=1$, contributes

$$\begin{aligned}T_1(N) &= \sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|=1}} p^{\mathbf{s}}\left(\frac{dQ(p)}{dX^{\mathbf{s}}}\right)^2\binom{N-4}{3}\binom{N}{1,3,N-4}^{-1}\\ &= 4\frac{\binom{N-4}{3}}{\binom{N}{4}}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|=1}} p^{\mathbf{s}}\left(\frac{dQ(p)}{dX^{\mathbf{s}}}\right)^2\end{aligned}$$

Since $\binom{N-4}{3}/\binom{N}{4} = O(1/N)$, it is enough to consider the other factor,

$$\sum_{\substack{\mathbf{s}\in\mathbb{N}^n\\ \|\mathbf{s}\|=1}} p^{\mathbf{s}}\left(\frac{dQ(p)}{dX^{\mathbf{s}}}\right)^2 = \sum_{k,\ell}p_{k,\ell}\left(\frac{dQ(p)}{dX_{k,\ell}}\right)^2\,.$$

We have, recalling the expression of the derivatives of Δ_{ij},

$$\begin{aligned}\frac{1}{2}\frac{dQ}{dX_{k,\ell}} &= \frac{1}{2}\sum_{i,j}2\Delta_{ij}\frac{d\Delta_{ij}}{dX_{k,\ell}}\\ &= \sum_{i\neq k}\sum_{j\neq\ell}X_{i,j}\Delta_{ij}(X) + \Delta_{k\ell}(X)\sum_{i\neq k}\sum_{j\neq\ell}X_{i,j}\\ &\quad - \sum_{j\neq\ell}\Delta_{kj}(X)\sum_{i\neq k}X_{i,j} - \sum_{i\neq k}\Delta_{i\ell}(X)\sum_{j\neq\ell}X_{i,j}\end{aligned}$$

Having this sum of four terms A_1, A_2, A_3, A_4 for $\frac{dQ}{dX_{k,\ell}}$, by Cauchy–Schwarz $\left(\frac{dQ}{dX_{k,\ell}}\right)^2 \leq 4(A_1^2+A_2^2+A_3^2+A_4^2)$, and so we can bound each of the square of these terms separately, ignoring cross factors.

  - For the first, we have (again by Cauchy–Schwarz)

$$\begin{aligned}\left(\sum_{i\neq k}\sum_{j\neq\ell}p_{i,j}\Delta_{ij}(p)\right)^2 &\leq \left(\sum_{i,j}p_{i,j}\Delta_{ij}(p)\right)^2\\ &\leq \left(\sum_{i,j}p_{i,j}^2\right)\left(\sum_{i,j}\Delta_{ij}(p)^2\right)\\ &\leq bQ(p) \leq \sqrt{b}Q(p)\end{aligned}$$

  so $\sum_{k,\ell}p_{k,\ell}\left(\sum_{i\neq k}\sum_{j\neq\ell}p_{i,j}\Delta_{ij}(p)\right)^2 \leq bQ(p)$.

  - For the second, since $\left(\Delta_{k\ell}(p)\sum_{i\neq k}\sum_{j\neq\ell}p_{i,j}\right)^2 \leq \Delta_{k\ell}(p)^2$, we have

$$\begin{aligned}\sum_{k,\ell}p_{k,\ell}\left(\Delta_{k\ell}(p)\sum_{i\neq k}\sum_{j\neq\ell}p_{i,j}\right)^2 &\leq \sum_{k,\ell}p_{k,\ell}\Delta_{k\ell}(p)^2\\ &\leq \sqrt{\sum_{k,\ell}p_{k,\ell}^2}\sqrt{\sum_{k,\ell}\Delta_{k\ell}(p)^4}\\ &\leq \sqrt{b}\sqrt{\left(\sum_{k,\ell}\Delta_{k\ell}(p)^2\right)^2}\end{aligned}$$

  which is equal to $\sqrt{b}Q(p)$.

- For the third and fourth term (similarly handled by symmetry),

$$\begin{aligned}
&\sum_{k,\ell} p_{k,\ell}\left(\sum_{j\neq\ell}\Delta_{kj}(p)\sum_{i\neq k}p_{i,j}\right)^2\\
&\qquad\leq \sum_{k,\ell} p_{k,\ell}\left(\sum_{j}\Delta_{kj}(p)\sum_{i\neq k}p_{i,j}\right)^2\\
&\qquad= \sum_{k}\left(\sum_{j}\Delta_{kj}(p)\sum_{i\neq k}p_{i,j}\right)^2\sum_{\ell}p_{k,\ell}\\
&\qquad\leq \sum_{k}\left(\sum_{j}\Delta_{kj}(p)^2\sum_{j}\left(\sum_{i\neq k}p_{i,j}\right)^2\right)\sum_{\ell}p_{k,\ell}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(Cauchy–Schwarz)}
\end{aligned}$$

and hence

$$\begin{aligned}
&\sum_{k,\ell} p_{k,\ell}\left(\sum_{j\neq\ell}\Delta_{kj}(p)\sum_{i\neq k}p_{i,j}\right)^2\\
&\quad\leq \sum_{k}\left(\sum_{j}\Delta_{kj}(p)^2\sum_{j}\left(\sum_{i}p_{i,j}\right)^2\right)\sum_{\ell}p_{k,\ell}\\
&\quad= \sum_{j}\left(\sum_{i}p_{i,j}\right)^2\cdot\sum_{k}\left(\sum_{j}\Delta_{kj}(p)^2\right)\sum_{\ell}p_{k,\ell}\\
&\quad\leq \sum_{j}\Big(\sum_{i}p_{i,j}\Big)^2\sqrt{\sum_{k}\Big(\sum_{j}\Delta_{kj}(p)^2\Big)^2\sum_{k}\Big(\sum_{\ell}p_{k,\ell}\Big)^2}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(Cauchy–Schwarz)}\\
&\quad\leq \sqrt{\sum_{j}\Big(\sum_{i}p_{i,j}\Big)^2\sum_{k}\Big(\sum_{\ell}p_{k,\ell}\Big)^2}\sqrt{\sum_{k}\Big(\sum_{j}\Delta_{kj}(p)^2\Big)^2}
\end{aligned}$$

where the last step relies on $\sum_j\big(\sum_i p_{i,j}\big)^2 \leq 1$ (since it is the squared ℓ_2 norm of a probability distribution, that of the first marginal of p) to write $\sum_j\big(\sum_i p_{i,j}\big)^2 \leq \sqrt{\sum_j\big(\sum_i p_{i,j}\big)^2}$. Continuing from there, and using monotonicity of ℓ_p norms to write $\sum_i v_i^2 \leq \big(\sum_i |v_i|\big)^2$,

$$\begin{aligned}
&\sum_{k,\ell} p_{k,\ell}\left(\sum_{j\neq\ell}\Delta_{kj}(p)\sum_{i\neq k}p_{i,j}\right)^2\\
&\quad\leq \sqrt{\sum_{j}\Big(\sum_{i}p_{i,j}\Big)^2\sum_{k}\Big(\sum_{\ell}p_{k,\ell}\Big)^2}\sum_{k}\sum_{j}\Delta_{kj}(p)^2\\
&\quad= \sqrt{\sum_{j}p_{\mathcal{Y}}(k)^2\sum_{k}p_{\mathcal{X}}(j)^2}Q(p) = \sqrt{\sum_{k,j}p_\Pi(k,j)^2}Q(p)\\
&\quad\leq \sqrt{b}Q(p)
\end{aligned}$$

and so $T_1(N) = O\Big(Q(p)\sqrt{b}/N\Big)$.

Gathering these four terms, and by the above discussion, we obtain

$$T_1(N) = 4\frac{\binom{N-4}{3}}{\binom{N}{4}}\sum_{k,\ell}p_{k,\ell}\left(\frac{dQ(p)}{dX_{k,\ell}}\right)^2 \leq 4\frac{\binom{N-4}{3}}{\binom{N}{4}}\cdot 8\cdot 4\sqrt{b}Q(p)$$

which is $O\left(\frac{\sqrt{b}Q(p)}{N}\right)$.

- Finally, for the rest of the contributions ($h \geq 2$), we invoke Lemma 5.5. Specifically, we first observe that, for any distribution $p \in \Delta(\mathcal{X}\times\mathcal{Y})$,

$$\begin{aligned}
Q^+(p) &= \sum_{i=1}^{\Lambda_1}\sum_{j=1}^{\Lambda_2}\left(p_{i,j}\sum_{i'\neq i}\sum_{j'\neq j}p_{i',j'} + \sum_{i'\neq i}p_{i',j}\sum_{j'\neq j}p_{i,j'}\right)^2\\
&\leq \sum_{i,j}\left(p_{i,j} + \sum_{i'=1}^{\Lambda_1}p_{i',j}\sum_{j'=1}^{\Lambda_2}p_{i,j'}\right)^2\\
&\leq 2\sum_{i,j}\left(p_{i,j}^2 + \left(\sum_{i'=1}^{\Lambda_1}p_{i',j}\right)^2\left(\sum_{j'=1}^{\Lambda_2}p_{i,j'}\right)^2\right)\\
&\leq 2\left(\|p\|_2^2 + \|p_\Pi\|_2^2\right) \leq 4b\,.
\end{aligned}$$

Next, we need to upper bound the high-order derivatives of Q. By Leibniz's rule, for $h \geq 2$ and $\|\mathbf{s}\| = h$,

$$\begin{aligned}
\frac{d^hQ}{dX^{\mathbf{s}}} &= \sum_{i,j}\frac{d^h\Delta_{ij}^2}{dX^{\mathbf{s}}} = \sum_{i,j}\sum_{\mathbf{s}'\leq\mathbf{s}}\prod_{\ell=1}^{n}\binom{s_\ell}{s'_\ell}\frac{d^{\|\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}'}}\frac{d^{\|\mathbf{s}\|-\|\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}-\mathbf{s}'}}\\
&\leq \sum_{\mathbf{s}'\leq\mathbf{s}}\prod_{i=\ell}^{n}\binom{s_\ell}{s'_\ell}\sqrt{\sum_{i,j}\left(\frac{d^{\|\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}'}}\right)^2\sum_{i,j}\left(\frac{d^{\|\mathbf{s}-\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}-\mathbf{s}'}}\right)^2}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(Cauchy–Schwarz)}\\
&\leq \max_{\mathbf{s}'\leq\mathbf{s}}\sum_{i,j}\left(\frac{d^{\|\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}'}}\right)^2\sum_{\mathbf{s}'\leq\mathbf{s}}\prod_{i=\ell}^{n}\binom{s_\ell}{s'_\ell}\\
&= 2^h\max_{\mathbf{s}'\leq\mathbf{s}}\sum_{i,j}\left(\frac{d^{\|\mathbf{s}'\|}\Delta_{ij}}{dX^{\mathbf{s}'}}\right)^2\,.
\end{aligned}$$

Since Δ_{ij} has degree 2, to bound this maximum we have to consider 3 cases: first, $\sum_{i,j}\left(\frac{d^0\Delta_{ij}(p)}{dX^0}\right)^2 = Q(p) \leq 4$. Second, recalling the partial derivatives of Δ_{ij} we computed earlier,

$$\begin{aligned}
\sum_{i,j}\left(\frac{d\Delta_{ij}(p)}{dX_{k,\ell}}\right)^2 &= \sum_{i\neq k}\sum_{j\neq\ell}p_{k,\ell}^2 + \left(\sum_{i'\neq k}\sum_{j'\neq\ell}p_{i',j'}\right)^2\\
&\qquad + \sum_{i'\neq k}p_{i',\ell}^2 + \sum_{j'\neq\ell}p_{k,j'}^2\\
&\leq 4\,.
\end{aligned}$$

Third,

$$\sum_{i,j}\left(\frac{d^2\Delta_{ij}(p)}{dX_{k,\ell}dX_{k',\ell'}}\right)^2 = \sum_{i,j}(\delta_{ik}-\delta_{ik'})^2(\delta_{j\ell}-\delta_{j\ell'})^2 \leq 4\,.$$

Combining all of the above cases results in $\left|\frac{d^h Q}{dX^{\mathbf{s}}}\right| \leq 2^4 \cdot 4$ for any $h \geq 2$ and $\|\mathbf{s}\| = h$, and from there

$$\sum_{h=2}^{4} T_h(N) = O\left(\frac{1}{N^2}\right) \cdot 2^4 \cdot 4 \cdot 4b = O\left(\frac{b}{N^2}\right).$$

Accounting for all the terms, we thus can bound the variance as

$$\operatorname{Var} U_N Q(\Phi_S) = (T_0(N) - Q(p)^2) + T_1(N) + \sum_{h=2}^{4} T_h(N)$$

which is $O\left(\frac{Q(p)\sqrt{b}}{N} + \frac{b}{N^2}\right)$; concluding the proof of Proposition 5.6.

Remark 2 (A Detour: Estimating a Polynomial under Poisson Sampling). We observe that analogues of our theorems hold under *Poisson* sampling (instead of multinomial sampling as treated in Section 5). We defer these results, which follow from a straightforward (yet slightly cumbersome) adaptation of the proofs of this section, to the full version of this paper.

6 TESTING WITH RESPECT TO MUTUAL INFORMATION

We conclude by considering a slightly different model from the one considered thus far. In particular, while the total variation metric is a reasonable one to measure what it means for X and Y to be far from conditionally independent, there is another metric that is natural in this context: *conditional mutual information.* Specifically, we modify the testing problem to distinguish between the cases where X and Y are conditionally independent on Z and the case where $I(X;Y|Z) \geq \varepsilon$. Our picture here is somewhat less complete, but we are still able to say something in the case where X, Y are binary.

Theorem 6.1. *If X and Y are binary random variables and Z has a support of size n, there exists a sample-efficient algorithm that distinguishes between $I(X;Y|Z) = 0$ and $I(X;Y|Z) \geq \varepsilon$ with sample complexity*

$$O\left(\max\left(\min\left(\frac{n^{6/7}}{\varepsilon^{8/7}}\log^{8/7}(1/\varepsilon), \frac{n^{7/8}}{\varepsilon}\log(1/\varepsilon)\right), \frac{\sqrt{n}}{\varepsilon^2}\log^2(1/\varepsilon)\right)\right).$$

Proof. This follows immediately upon noting that by Lemma 6.2 (stated and proven later), that if X and Y are ε-close in total variation distance from being conditionally independent on Z, then $I(X;Y|Z) \leq O(\varepsilon \log(1/\varepsilon))$; or, by the contrapositive, that $I(X;Y|Z) \geq \varepsilon$ implies that X and Y are $\Omega(\varepsilon/\log(1/\varepsilon))$-far in total variation distance from being conditionally independent on Z. Therefore, it suffices to run our existing conditional independence tester with parameter $\varepsilon' := \Omega(\varepsilon/\log(1/\varepsilon))$. The sample complexity of this tester is as specified. □

Remark 3 (On the optimality of this bound). It is not difficult to modify the analysis slightly in order to remove the logarithmic factors from the first two terms in the above expression. Intuitively, this is because these terms arise only when at least half of the mutual information comes from "light" bins, with mass at most $1/m$. In this case, these bins contribute at least $m^4 \sum_z \varepsilon_z^2 p_Z(z)^4 \gg m^4 \sum_z (p_Z(z)\varepsilon_z \log(1/\varepsilon_z))^4 \gg m^4\varepsilon^4/n^3$ to the expectation of Z, and the analysis proceeds from there as before.

It is also easy to show that in this regime our lower bounds still apply, as the hard instances also produced distributions with mutual information $\Omega(\varepsilon)$.[2] Therefore, we have matching upper and lower bounds as long as $\varepsilon \gg n^{-3/8}/\log^2 n$.

However, it seems likely that the correct behavior in the small ε regime is substantially different when testing with respect to mutual information. The difficult cases for total variation distance testing actually end up with mutual information merely $I(X;Y|Z) = O(\varepsilon^2)$. It is quite possible that a better algorithm or a better analysis of the existing algorithm could give substantially improved performance when $\varepsilon < n^{-3/8}$. In fact, it is conceivable that the sample complexity of $O(n^{7/8}/\varepsilon)$ could be maintained for a broad range of ε. The only lower bound that we know preventing this is a lower bound of $\Omega(\varepsilon \log(1/\varepsilon))$ by noting that there are distributions with $I(X;Y|Z) \geq \varepsilon$, but where (X, Y, Z) is $O(\varepsilon/\log(1/\varepsilon))$-far in variation distance from being conditionally independent.

Lemma 6.2. *Assume $(X, Y, Z) \sim p$, where $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ with $|\mathcal{X}| = \Lambda_1$, $|\mathcal{Y}| = \Lambda_2$, and $|\mathcal{Z}| = n$. Then, for every $\varepsilon \in (0, 1)$,*

- *If $\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\right) \leq \varepsilon$, then $I(X;Y|Z) \leq O(\varepsilon \log(\Lambda_1\Lambda_2/\varepsilon))$;*
- *If $\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\right) \geq \varepsilon$, then $I(X;Y|Z) \geq 2\varepsilon^2$.*

Proof. The second item is simply an application of Pinsker's inequality, recalling that

$$I(X;Y|Z) = \mathrm{d}_{\mathrm{KL}}((X,Y) \mid Z \mid\mid (X \mid Z) \otimes (Y \mid Z)).$$

i.e. the Kullback–Leibler divergence between the joint distribution of $(X, Y \mid Z)$ and the product of marginals $(X \mid Z)$ and $(Y \mid Z)$. As for the first, it follows from the relation between conditional mutual information and total variation distance obtained in [40]. □

ACKNOWLEDGMENTS

C. L. C. is a supported by a Motwani Postdoctoral Fellowship. Some of this work was performed while visiting USC, and a graduate student at Columbia University supported by NSF CCF-1115703 and NSF CCF-1319788. I. D. is supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. D. M. K. is supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.

REFERENCES

[1] J. Acharya, C. Daskalakis, and G. Kamath. 2015. Optimal Testing for Properties of Distributions. In *Proceedings of NIPS'15*.

[2] A. Agresti. 1992. A Survey of Exact Inference for Contingency Tables. *Statist. Sci.* 7, 1 (02 1992), 131–153.

[3] 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

[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] R. Blundell and J. L. Horowitz. 2007. A Non-Parametric Test of Exogeneity. *The Review of Economic Studies* 74, 4 (2007), 1035–1058. http://www.jstor.org/stable/4626172

[6] T. Bouezmarni and A. Taamouti. 2014. Nonparametric tests for conditional independence using conditional distributions. *Journal of Nonparametric Statistics* 26, 4 (2014), 697–719.

[2] I.e., the conditional mutual information of "no-distributions" is easily seen to actually be $\Omega(\varepsilon)$, while applying the relation between total variation distance and conditional mutual information as a black-box to the ε distance in total variation distance would incur a quadratic loss in ε.

[7] C. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. 2016. Testing Shape Restrictions of Discrete Distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*. 25:1–25:14.

[8] C. L. Canonne. 2015. A Survey on Distribution Testing: Your Data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)* 22 (2015), 63.

[9] 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.

[10] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2017. Testing Conditional Independence of Discrete Distributions. *CoRR* abs/1711.11560 (2017).

[11] C. L. Canonne, I. Diakonikolas, and A. Stewart. 2017. Fourier-Based Testing for Families of Distributions. *CoRR* abs/1706.05738 (2017).

[12] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. 2014. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *SODA*. 1193–1203.

[13] W. G. Cochran. 1954. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics* 10, 4 (1954), 417–451.

[14] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. 2013. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*. 1833–1852.

[15] C. Daskalakis, N. Dikkala, and G. Kamath. 2018. Testing Ising Models. In *SODA*. To appear.

[16] 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.

[17] A. P. Dawid. 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41, 1 (1979), 1–31. http://www.jstor.org/stable/2984718

[18] P. de Morais Andrade, J. M. Stern, and C. A. de Braganca Pereira. 2014. Bayesian Test of Significance for Conditional Independence: The Multinomial Model. *Entropy* 16, 3 (2014), 1376–1395.

[19] M. A. Delgado and W. G. Manteiga. 2001. Significance Testing in Nonparametric Regression Based on the Bootstrap. *The Annals of Statistics* 29, 5 (2001), 1469–1507.

[20] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. 2016. Collision-based Testers are Optimal for Uniformity and Closeness. *Electronic Colloquium on Computational Complexity (ECCC)* 23 (2016), 178.

[21] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. 2017. Sample-Optimal Identity Testing with High Probability. *CoRR* abs/1708.02728 (2017).

[22] 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.

[23] 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*.

[24] 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*.

[25] 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.

[26] R. L. Dobrušin. 1959. A general formulation of the fundamental theorem of Shannon in the theory of information. *Uspehi Mat. Nauk* 14, 6 (90) (1959), 3–104.

[27] D. Easley and M. O'Hara. 1987. Price, trade size, and information in securities markets. *Journal of Financial Economics* 19, 1 (1987), 69 – 90.

[28] R. A. Fisher. 1924. The distribution of the partial correlation coefficient. *Metron* 3 (1924), 329–332.

[29] G. Geenens and L. Simar. 2010. Nonparametric tests for conditional independence in two-way contingency tables. *Journal of Multivariate Analysis* 101, 4 (2010), 765–788.

[30] O. Goldreich. 2017. *Introduction to Property Testing*. Forthcoming. http://www.wisdom.weizmann.ac.il/~oded/pt-intro.html

[31] C.W.J. Granger. 1980. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* 2, Supplement C (1980), 329 – 352.

[32] M. Hardt, E. Price, and N. Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. 3315–3323.

[33] T.-M. Huang. 2010. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* 38, 4 (08 2010), 2047–2091.

[34] O. Linton and P. Gozalo. 1996. *Conditional Independence Restrictions: Testing and Estimation*. Cowles Foundation Discussion Papers 1140. Cowles Foundation for Research in Economics, Yale University.

[35] N. Mantel and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22, 4 (April 1959), 719–748. http://www.ncbi.nlm.nih.gov/pubmed/13655060 PMID: 13655060.

[36] K. Natori, M. Uto, and M. Ueno. 2017. Consistent Learning Bayesian Networks with Thousands of Variables. In *Proceedings of The 3rd International Workshop on Advanced Methodologies for Bayesian Networks (Proceedings of Machine Learning Research)*, Vol. 73. PMLR, 57–68. http://proceedings.mlr.press/v73/natori17a.html

[37] R. E. Neapolitan. 2003. *Learning Bayesian Networks*. Prentice-Hall, Inc.

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

[39] J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[40] M. S. Pinsker. 2005. On the estimation of information via variation. *Problemy Peredachi Informatsii* 41, 2 (2005), 3–8. https://doi.org/10.1007/s11122-005-0012-8

[41] R. Rubinfeld. 2012. Taming big probability distributions. *XRDS* 19, 1 (2012), 24–28.

[42] K. Song. 2009. Testing conditional independence via Rosenblatt transforms. *Ann. Statist.* 37, 6B (12 2009), 4011–4045.

[43] P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search* (2nd ed.). MIT press.

[44] L. Su and H. White. 2007. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics* 141, 2 (2007), 807 – 834.

[45] L. Su and H. White. 2008. A Nonparametric Hellinger Metric Test for Conditional Independence. *Econometric Theory* 24, 4 (2008), 829–864.

[46] L. Su and H. White. 2014. Testing conditional independence via empirical likelihood. *Journal of Econometrics* 182, 1 (2014), 27 – 44. Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions.

[47] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65, 1 (01 Oct 2006), 31–78.

[48] G. Valiant and P. Valiant. 2014. An Automatic Inequality Prover and Instance Optimal Identity Testing. In *FOCS*.

[49] X. Wang and Y. Hong. 2017. Characteristic Function Based Testing For Conditional Independence: A Nonparametric Regression Approach. *Econometric Theory* (2017), 1–35.

[50] A. D. Wyner. 1978. A definition of conditional mutual information for arbitrary ensembles. *Inform. and Control* 38, 1 (1978), 51–59. https://doi.org/10.1016/S0019-9958(78)90026-8

[51] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2011. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI'11)*. AUAI Press, 804–813.