Testing Conditional Independence of Discrete Distributions

Clément L. Canonne*
Stanford University
ccanonne@stanford.edu

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

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

Alistair Stewart
University of Southern California
alistais@usc.edu

December 1, 2017

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 $[\ell_1] \times [\ell_2] \times [n]$, 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 $[\ell_1] \times [\ell_2] \times [n]$. 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 $\ell_1, \ell_2 = O(1)$, we show that the sample complexity of testing conditional independence (upper bound and matching lower bound) is

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

To obtain our tester, we employ a variety of tools, including (1) a suitable weighted adaptation of the flattening technique [DK16], 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 (see, e.g., [CDVV14]). 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.

*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 Award CCF-1652862 (CAREER) and a Sloan Research Fellowship.

‡Supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.

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 grouped 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 applications in a variety of fields, including medicine, economics and finance, etc. (see, e.g., [MH59, SGS00, WH17] 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 [Daw79]. Specifically, it is a central notion in modeling causal relations [SGS00] and of crucial importance in graphical modeling [Pea88]. Conditional independence is, in several settings, 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 [EO87]. Other examples include distributional Granger non-causality [Gra80] — which is a particular case of conditional independence — and exogeneity [BH07].

Given the widespread applications of the conditional independence assumption, the statistical task of *testing* conditional independence has been studied extensively for nearly a century. In 1924, R. A. Fisher [Fis24] 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 [Agr92]. In the field of graphical models, conditional independence testing is a cornerstone in the context of structure learning and testing of Bayesian networks (see, e.g., [Nea03, TBA06, NUU17, CDKS17] and references therein). Finally, conditional independence testing is a useful tool in recent applications of machine learning involving fairness [HPS16].

One of the classical conditional independence tests in the discrete setting is the Cochran–Mantel–Haenszel test [Coc54, MH59], 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^2-squared testers (see, e.g., [Agr92]). However, even for the most basic case of distributions over $\{0, 1\}^2 \times [n]$, no finite sample analysis is known. (Interestingly enough, we note that our tester can be viewed as an appropriately weighted linear combination of χ^2-squared tests.) A recent line of work in econometrics has been focusing on conditional independence testing in *continuous settings* [LG96, DM01, SW07, SW08, Son09, GS10, Hua10, SW14, ZPJS11, BT14, dMASdBP14, WH17]. 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* [BFR+00] has seen substantial progress in the past decade,

see [Rub12, Can15, Gol17] 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 [Pan08, CDVV14, VV14, DKN15b, ADK15, CDGR16, DK16, DGPP16, CDS17, Gol17, DGPP17]. Moreover, an emerging body of work has focused on leveraging *a priori* structure of the underlying distributions to obtain significantly improved sample complexities [BKR04, DDS+13, DKN15b, DKN15a, CDKS17, DP17, DDK18, DKN17].

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 $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ to denote the property of conditional independence, i.e.,

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

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

> $\mathcal{T}(\ell_1, \ell_2, n, \varepsilon)$: Given sample access to a distribution p over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, with $|\mathcal{X}| = \ell_1$, $|\mathcal{Y}| = \ell_2$, $|\mathcal{Z}| = n$, and $\varepsilon > 0$, distinguish with probability at least $2/3$ between the following cases:
>
> - Completeness: $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$.
> - Soundness: $\mathrm{d}_{\mathrm{TV}}(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}) \geq \varepsilon$.

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 near-optimal algorithms under the mutual information metric as well. The interested reader is referred to Appendix A 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 $[\ell_1] \times [\ell_2]$, whose testing sample complexity was resolved only recently [DK16]. 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 1.1. *There exists a computationally efficient tester for $\mathcal{T}(2, 2, n, \varepsilon)$ with sample complexity*

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

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* $\Omega\Big(\max\Big(n^{1/2}/\varepsilon^2, \min\Big(n^{7/8}/\varepsilon, n^{6/7}/\varepsilon^{8/7}\Big)\Big)\Big)$ *samples.*

To the best of our knowledge, prior to our work, no $o(n)$ sample algorithm was known for this problem. Our algorithm in this regime is 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 a 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 1.2. *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 1.2 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 1.2 is of independent interest. Indeed, 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 [CDVV14] (which has been used as a primitive to obtain a wide range of sample-optimal testers [DK16]) is an unbiased estimator for the squared ℓ_2-distance between two distributions p, q over $[n]$. While the description of such unbiased estimators may be relatively simple, their analyses are 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 1.1 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 $[\ell_1] \times [\ell_2] \times [n]$ is the following:

Theorem 1.3. *There exists a computationally efficient tester for $\mathcal{T}(\ell_1, \ell_2, n, \varepsilon)$ with sample complexity*

$$O\left(\max\left(\min\left(\frac{n^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon}, \frac{n^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}}\right), \frac{n^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon}, \frac{n^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon^2}\right)\right) , \qquad (1)$$

where we assume without loss of generality that $\ell_1 \geq \ell_2$.

The expression of the sample complexity in Theorem 1.3 may seem somewhat unwieldy. In an attempt to interpret this bound, we consider several important special cases of interest:

- For $\ell_1 = \ell_2 = O(1)$, (1) reduces to the binary case for X, Y, recovering the tight bound of Theorem 1.1.
- For $n = 1$, our problem reduces to the task of *testing independence* of a distribution over $[\ell_1] \times [\ell_2]$, which has been extensively studied [BFF+01, LRR11, ADK15, DK16]. In this case, (1) recovers the optimal sample complexity of independence testing, i.e., $\Theta\left(\max\left(\ell_1^{2/3}\ell_2^{1/3}/\varepsilon^{4/3}, \sqrt{\ell_1\ell_2}/\varepsilon^2\right)\right)$ [DK16].
- For $\ell_1 = \ell_2 = n$ (and $\varepsilon = \Omega(1)$), the sample complexity of (1) becomes $O(n^{7/4})$. In Theorem 1.4 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 1.4. *Any tester for $\mathcal{T}(n, n, n, 1/20)$ requires $\Omega(n^{7/4})$ samples.*

1.3 Our Techniques

1.3.1 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 compares 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 rely 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 bound 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 polynomial 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 gives 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 [DK16] 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.

1.3.2 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*, introduced in [DK16]. 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. Let m be the number of samples we take and n be the domain of Z. If the bin weights are chosen appropriately, we show that the final estimator A has variance $O(\min(n, m) + \sqrt{\min(n, m)}\mathbb{E}[A] + \mathbb{E}[A]^{3/2})$, with high probability over the number of samples falling in each bin as well as the flattening we perform for each bin. (This high-probability statement, in turn,

is enough for us to apply Chebyshev's inequality to our final estimator in the end.) 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 fact, we will need to separate the Z-bins into four categories 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 will 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.

1.3.3 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. Specifically, we construct a pair of ensembles $\mathcal{D}$ and $\mathcal{D}'$ of pairs of nearly-normalized pseudo-distributions such that the following holds: (1) Pseudo-distributions drawn from $\mathcal{D}$ satisfy the desired property and pseudo-distribution drawn from $\mathcal{D}'$ are ε-far from satisfying the property with high probability, and (2) $\mathrm{Poisson}(s)$ samples 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 [DK16]. 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 $\mathcal{D}$ and $\mathcal{D}'$ — where draws from $\mathcal{D}$ are conditionally independent and draws from $\mathcal{D}'$ 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 $\mathcal{D}$ or $\mathcal{D}'$, when s is small. We define $\mathcal{D}$ and $\mathcal{D}'$ 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 $\mathcal{D}$ and $\mathcal{D}'$. *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 $\mathcal{D}$ or $\mathcal{D}'$ 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 $\mathcal{D}$ as from $\mathcal{D}'$.

To rigorously analyze the above sketched construction, we consider drawing $\mathrm{Poisson}(s)$ samples from a random distribution from either $\mathcal{D}$ or $\mathcal{D}'$, and bound the mutual 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 mutual information coming from a single bin. By the above, the probabilities of seeing any triple of samples are the same for either $\mathcal{D}$ or $\mathcal{D}'$ and thus contribute nothing to the mutual 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 mutual 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 mutual 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.

1.4 Organization

After setting up the required preliminaries in Section 2, we give our testing algorithm for the case of constant $|\mathcal{X}|, |\mathcal{Y}|$ in Section 3. In Section 4, we develop our theory for polynomial estimation. Section 5 leverages this theory, along with several other ideas, to obtain our general testing algorithm for conditional independence. Section 6 gives our information-theoretic lower bound for the setting of binary $|\mathcal{X}|, |\mathcal{Y}|$. In Section 7, we give an information-theoretic lower bound matching the sample complexity of our algorithm for the regime where $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}|$. In Appendix A, we discuss the implications of our results for conditional independence testing with regard to the conditional mutual information.

2 Preliminaries and Basic Facts

We begin with some standard notation and definitions that we shall use throughout the paper. For $m \in \mathbb{N}$, we write $[m]$ for the set $\{1, \ldots, m\}$, and $\log$ for the binary logarithm.

Distributions and Metrics A probability distribution over discrete domain Ω is a function $p\colon \Omega \to [0,1]$ such that $\|p\|_1 := \sum_{\omega\in\Omega} p(\omega) = 1$. Without the requirement that the total mass be one, p is said to be a *pseudo-distribution*. We denote by $\Delta(\Omega)$ the set of all probability distributions over domain Ω. For two probability distributions $p, q \in \Delta(\Omega)$, their *total variation distance* (or statistical distance) is defined as $\mathrm{d_{TV}}(p,q) := \sup_{S\subseteq\Omega}(p(S) - q(S)) = \frac{1}{2}\sum_{\omega\in\Omega} |p(\omega) - q(\omega)|$, i.e., $\mathrm{d_{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$, and 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 $\mathrm{d_{TV}}(p,\mathcal{P}) := \inf_{q\in\mathcal{P}} \mathrm{d_{TV}}(p,q)$. If $\mathrm{d_{TV}}(p,\mathcal{P}) > \varepsilon$, we say that p is *ε-far* from $\mathcal{P}$; otherwise, it is *ε-close*. 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 $\mathrm{d_{TV}}(p,\mathcal{P}) > \varepsilon$, 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 it draws from the underlying distribution) and its running time.

2.1 Conditional Independence

We record here a number of notations definitions regarding conditional independence. 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 $(X' \perp Y') \mid Z'$. For discrete sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, we will denote by $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ the property of conditional independence, i.e., $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}} := \{\, p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}) \,:\, (X,Y,Z) \sim p \text{ satisfies } (X \perp Y) \mid Z \,\}$. We say that a distribution $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ is ε-far from $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, if for every distribution $q \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ we have that $\mathrm{d_{TV}}(p,q) > \varepsilon$. Fix a distribution $q \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ 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 consider testing conditional independence with respect to a different metric, namely the *conditional mutual information* [Dob59, Wyn78]. 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) K-L 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 Appendix A), 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 2.1 (Product of Conditional Marginals)**.** Let $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. 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 $q := \sum_{z\in\mathcal{Z}} p_Z(z)q_z \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, i.e., $q(i,j,z) := p_Z(z)\cdot q_z(i,j)$.

2.2 Basic Facts

We start with the following simple lemma:

Lemma 2.1. *Let $p, p' \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. Then we have that*

$$\mathrm{d}_{\mathrm{TV}}(p,p') \le \sum_{z\in\mathcal{Z}} p_Z(z)\cdot \mathrm{d}_{\mathrm{TV}}(p_z,p'_z) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z)\ , \tag{2}$$

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

Using Lemma 2.1, we deduce the following useful corollary:

Fact 2.1. *If $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ is ε-far from $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, then, for every $p' \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, either (i) $\mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) > \varepsilon/2$, or (ii) $\sum_{z\in\mathcal{Z}} p_Z(z)\cdot\mathrm{d}_{\mathrm{TV}}(p_z,p'_z) > \varepsilon/2$.*

Proof. Let $p' \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$. Since p is ε-far from $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ we have that $\mathrm{d}_{\mathrm{TV}}(p,p') > \varepsilon$. By Lemma 2.1, we this obtain that $\sum_{z\in\mathcal{Z}} p_Z(z)\cdot\mathrm{d}_{\mathrm{TV}}(p_z,p'_z) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) > \varepsilon$, which proves the fact. □

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 2.1:

Lemma 2.2. *Suppose $p \in \Delta(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ is ε-close to $\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$. Then, p is 4ε-close to the distribution $q = \sum_{z\in\mathcal{Z}} p_Z(z)q_z$.*

2.3 Flattening Distributions

We now recall some notions and lemmata from previous work, regarding the technique of *flattening* of discrete distributions:

Definition 2.2 (Split distribution [DK16])**.** Given a distribution $p \in \Delta([n])$ and a multi-set S of elements of $[n]$, define the *split distribution* $p_S \in \Delta([n+|S|])$ as follows: For $1 \le i \le n$, let a_i denote 1 plus the number of elements of S that are equal to i. Thus, $\sum_{i=1}^n a_i = n + |S|$. We can therefore associate the elements of $[n+|S|]$ to elements of the set $B_S := \{\,(i,j) : i \in [n], 1 \le j \le a_i\,\}$. We now define a distribution p_S with support B_S, by letting a random sample from p_S be given by (i,j), where i is drawn randomly from p and j is drawn uniformly from $[a_i]$.

Fact 2.2 ([DK16, Fact 2.5]). *Let $p, q \in \Delta([n])$, and S a given multi-set of $[n]$. Then: (i) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively. (ii) It holds $\mathrm{d}_{\mathrm{TV}}(p_S, q_S) = \mathrm{d}_{\mathrm{TV}}(p, q)$.*

We will also require the analogue of [DK16, Lemma 2.6] (how flattening reduces the ℓ_2-norm of a distribution) for the non-Poissonized setting, i.e., when exactly m samples are drawn (instead of $\mathrm{Poisson}(m)$). The proof of this lemma is similar to that of [DK16, Lemma 2.6], and we include it in Appendix B for completeness:

Lemma 2.3. *Let $p \in \Delta([n])$. Then: (i) For any multi-sets $S \subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \leq \|p_S\|_2$, and (ii) If S is obtained by taking m independent samples from p, then $\mathbb{E}\left[\|p_S\|_2^2\right] \leq \frac{1}{m+1}$.*

Remark 2.3. Given S and $(a_i)_{i\in[n]}$ as in Definition 2.2, it is immediate that for any $p, q \in \Delta([n])$ it holds $\|p_S - q_S\|_2^2 = \sum_{i=1}^n \frac{(p_i - q_i)^2}{a_i}$ so that an ℓ_2^2 statistic for p_S, q_S can be seen as a particular rescaled ℓ_2 statistic for p, q.

2.4 Technical Facts on Poisson Random Variables

We state below some technical result on moments of truncated Poisson random variables, which we will use in various places of our analysis. The proof of these claims are deferred to Appendix B.

Claim 2.1. *There exists an absolute constant $C > 0$ such that, for $N \sim \mathrm{Poisson}(\lambda)$,*

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

Moreover, one can take $C = 4.22$.

Claim 2.2. *There exists an absolute constant $C > 0$ such that, for $X \sim \mathrm{Poisson}(\lambda)$ and $a, b \geq 0$,*

$$\mathrm{Var}[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}] \leq C\mathbb{E}\left[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\right] .$$

Claim 2.3. *There exists an absolute constant $C > 0$ such that, for $X \sim \mathrm{Poisson}(\lambda)$ and integers $a, b \geq 2$,*

$$\mathbb{E}\left[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\right] \geq C\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4) .$$

3 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 present and analyze our conditional independence tester for the case that $|\mathcal{X}|, |\mathcal{Y}| = O(1)$. Specifically, we will present a tester for this regime whose sample complexity is optimal, up to constant factors. 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:

$$\mathbb{E}[\Phi(s)] = \|p - p_{\mathcal{X}} \otimes p_{\mathcal{Y}}\|_2^2 \tag{3}$$

$$\mathrm{Var}[\Phi(s)] \leq C\left(\frac{\mathbb{E}[\Phi(s)]}{N} + \frac{1}{N^2}\right) , \tag{4}$$

for some absolute constant $C > 0$. Such an estimator follows as a special case of our generic polynomial estimators of Section 4.

Notation Let $p \in \Delta(\mathcal{X} \times \mathcal{Y})$. 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 $p_{\mathcal{Y}}$. Let $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. For $z \in \mathcal{Z}$, we will denote by q_z the product distribution $p_{z,X} \otimes p_{x,Y}$.

Let M be a $\mathrm{Poisson}(m)$ 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 statistic (that we will compare to a suitable threshold in the eventual test) is

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

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

$$m \geq \beta \max\left(\sqrt{n}/\varepsilon'^2, \min\left(n^{7/8}/\varepsilon', n^{6/7}/\varepsilon'^{8/7}\right)\right)\ , \tag{5}$$

for a sufficiently large absolute constant $\beta > 0$.

Interestingly enough, there are three distinct regions for this expression, based on the relation between n and ε, as illustrated in the following figure:

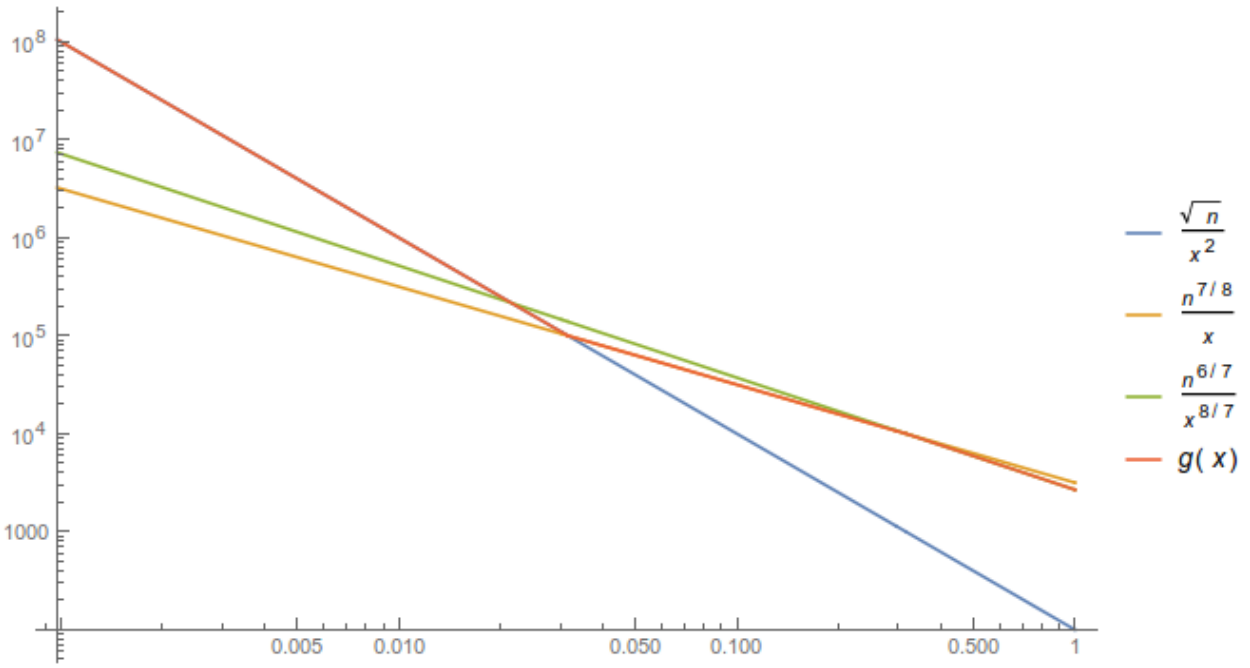

Figure 1: The three regimes of the sample complexity (for $n = 10000$ and $\varepsilon \in (0, 1]$) (log-log plot).

Our conditional independence tester outputs "accept" if $A \geq \tau$ and "reject" otherwise, where τ is selected to be $\Theta\left(\max\left(m\varepsilon'^2, \frac{m^4\varepsilon'^4}{n^3}\right)\right)$. A detailed pseudo-code for the algorithm is given in Algorithm 1.

Algorithm 1 TESTCONDINDEPENDENCE

Require: Parameter $n := |\mathcal{Z}|$, $\ell_1 := |\mathcal{X}|$, $\ell_2 := |\mathcal{Y}|$, $\varepsilon \in (0, 1]$, 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 $\varepsilon' := \varepsilon/\sqrt{\ell_1\ell_2}$ ▷ $\beta \geq 1$ is a sufficiently large absolute constant

Set $\tau \leftarrow \frac{\gamma}{2} \max\left(m\varepsilon'^2, \frac{m^4\varepsilon'^4}{n^3}\right)$, where $\gamma := 1 - \frac{5}{2e}$. ▷ Threshold for accepting

Draw $M \sim \mathrm{Poisson}(m)$ 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 Φ

    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 $A := \sum_{z\in\mathcal{Z}} A_z \geq \tau$ **then**

  return accept

else

  return reject

end if

3.1 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 $\mathrm{d}_{\mathrm{TV}}(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}) > \varepsilon$, 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, γ are absolute constants defined in the algorithm pseudo-code.

3.1.1 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 3.1. *We have the following: (a) If $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, then $\mathbb{E}[A] = 0$. (b) If $\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\right) > \varepsilon$, then $\mathbb{E}[A] > \gamma \min\left(m\varepsilon'^2, \frac{m^4\varepsilon'^4}{8n^3}\right) \geq \frac{\beta\cdot\gamma}{8} \cdot \sqrt{\min(n, m)}$.*

The rest of this subsection is devoted to the proof of Proposition 3.1. We start by providing a convenient lower bound on the expectation of A. We prove the following lemma:

Lemma 3.1. *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:*

$$\mathbb{E}[A] \geq \gamma \cdot \sum_{z\in\mathcal{Z}} \delta_z^2 \min(\alpha_z, \alpha_z^4)\ . \tag{6}$$

Proof. Conditioned on $\sigma_z = |S_z|$, Eq. (3) gives that $\mathbb{E}[\,A_z \mid \sigma_z\,] = \sigma_z\delta_z^2\mathbb{1}_{\{\sigma_z\geq 4\}}$. Therefore, for $\sigma := (\sigma_z)_{z\in\mathcal{Z}}$, we can write $\mathbb{E}[\,A \mid \sigma\,] = \sum_{z\in\mathcal{Z}} \sigma_z\delta_z^2\mathbb{1}_{\{\sigma_z\geq 4\}}$. Using the fact that the σ_z's are independent and

$\sigma_z \sim \mathrm{Poisson}(\alpha_z)$, we obtain the following closed-form expression for the expectation:

$$\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[\, A \mid \sigma\,]]_\sigma = \sum_{z\in\mathcal{Z}} \delta_z^2 \mathbb{E}\Big[\sigma_z \mathbb{1}_{\{\sigma_z \geq 4\}}\Big] = \sum_{z\in\mathcal{Z}} \delta_z^2 \cdot f(\alpha_z)\ , \tag{7}$$

where $f\colon \mathbb{R}_+ \to \mathbb{R}$ is the function $f(x) = e^{-x}\sum_{k=4}^{\infty} k\frac{x^k}{k!} = x - e^{-x}(x + x^2 + \frac{x^3}{2})$. Let $g\colon \mathbb{R}_+ \to \mathbb{R}$ be defined by $g(x) = \min(x, x^4)$. It is not hard to check that the function $f(x)/g(x)$ achieves its minimum at $x = 1$, where it takes the value $\gamma := 1 - \frac{5}{2e} > 0$. That is, $f(\alpha_z) \geq \gamma \cdot g(\alpha_z)$ and the lemma follows from (7). □

Given (7), the first statement of Proposition 3.1 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 $\mathrm{d_{TV}}\Big(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\Big) > \varepsilon$. We require the following useful claim:

Claim 3.1. *If* $\mathrm{d_{TV}}\Big(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\Big) > \varepsilon$, *then* $\sum_{z\in\mathcal{Z}} \delta_z\alpha_z > 2m\varepsilon'$.

Proof. We use the identity $\mathrm{d_{TV}}(p, q) = \sum_{z\in\mathcal{Z}} p_Z(z) \cdot \mathrm{d_{TV}}(p_z, q_z)$, which follows from Lemma 2.1 noting that $q_Z = p_Z$. By assumption, we have that $\mathrm{d_{TV}}(p, q) > \varepsilon$. We can therefore write

$$\sum_{z\in\mathcal{Z}} \delta_z\alpha_z = m \sum_{z\in\mathcal{Z}} \|p_z - q_z\|_2 \cdot p_Z(z) \geq \frac{m}{\sqrt{|\mathcal{X}|\,|\mathcal{Y}|}} \cdot \sum_{z\in\mathcal{Z}} \|p_z - q_z\|_1 p_Z(z) > \frac{2m}{\sqrt{|\mathcal{X}|\,|\mathcal{Y}|}}\varepsilon\ ,$$

where the last inequality is Cauchy–Schwarz. □

Lemma 3.1 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 3.1, at least one of the following 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.

(1) Suppose that $\sum_{z\in\mathcal{Z}_H} \delta_z\alpha_z > m\varepsilon'$. We claim that $\mathbb{E}[A] > \gamma\cdot m{\varepsilon'}^2$. Indeed, this follows from Lemma 3.1 and (8) using the following chain of (in)-equalities:

$$\sum_{z\in\mathcal{Z}_H} \delta_z^2\alpha_z \geq \frac{\Big(\sum_{z\in\mathcal{Z}_H} \delta_z\alpha_z\Big)^2}{\sum_{z\in\mathcal{Z}_H} \alpha_z} > m{\varepsilon'}^2\ ,$$

where the first inequality is Cauchy–Schwarz, and the second follows using that $\sum_{z\in\mathcal{Z}_H} \alpha_z \leq m$.

(2) Suppose that $\sum_{z\in\mathcal{Z}_L} \delta_z\alpha_z > m\varepsilon'$. We claim that $\mathbb{E}[A] > \gamma\cdot\frac{m^4{\varepsilon'}^4}{8n^3}$. Indeed, this follows from Lemma 3.1 and (8) using the following chain of (in)-equalities:

$$\sum_{z\in\mathcal{Z}_L} \delta_z^2\alpha_z^4 \geq \frac{1}{8n^3}\Big(\sum_{z\in\mathcal{Z}_L} \delta_z\alpha_z\Big)^4 > \frac{1}{8}\frac{m^4{\varepsilon'}^4}{n^3}\ .$$

The first inequality essentially follows by an application Jensen's inequality as follows: Let $\delta := \sum_{z \in \mathcal{Z}_L} \delta_z^{2/3}$. By Jensen's inequality we have:

$$\Big(\sum_{z \in \mathcal{Z}_L} (\delta_z^{2/3}/\delta) \cdot \delta_z^{1/3} \alpha_z \Big)^4 \leq \sum_{z \in \mathcal{Z}_L} (\delta_z^{2/3}/\delta) \cdot \delta_z^{4/3} \alpha_z^4 \, ,$$

or

$$\Big(\sum_{z \in \mathcal{Z}_L} \delta_z \alpha_z \Big)^4 \leq \delta^3 \sum_{z \in \mathcal{Z}_L} \delta_z^2 \alpha_z^4 \, .$$

Since $\delta_z \leq 2$ for all $z \in \mathcal{Z}$, it follows that $\delta \leq 2n$, which completes the proof of the claim.

We have thus far established that

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

giving the first inequality of Proposition 3.1 (b). To complete the proof of the proposition, it suffices to show that

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

We show this below by considering the following cases:

- If $n \geq \beta m$, we must be in the range $1/n^{1/8} \leq \varepsilon' \leq 1$ where $\max(\sqrt{n}/\varepsilon'^2, \min(n^{7/8}/\varepsilon', n^{6/7}/\varepsilon'^{8/7})) = n^{6/7}/\varepsilon'^{8/7}$. We get $\frac{m^4\varepsilon'^4}{n^3} \leq \beta^3 \frac{m^4\varepsilon'^4}{n^3} \leq m\varepsilon'^4 \leq m\varepsilon'^2$, and then since $\frac{m^4\varepsilon'^4}{n^3} \geq \beta^{7/2}\sqrt{m}$ by our choice of m in Eq. (5), we get that

  $$\min \left(m\varepsilon'^2, \frac{m^4 \varepsilon'^4}{n^3} \right) \geq \sqrt{\min(n, m)} \, ,$$

  as desired assuming that $\beta \geq 1$.
- If $\beta m \geq n$, we must be in the range $0 < \varepsilon' \leq 1/n^{1/8}$, and therefore $\min(n^{7/8}/\varepsilon', n^{6/7}/\varepsilon'^{8/7}) = n^{7/8}/\varepsilon'$. Since $m\varepsilon'^2 \geq \beta\sqrt{n}$ and $\frac{m^4\varepsilon'^4}{n^3} \geq \beta^4\sqrt{n}$ by our choice of m in Eq. (5), we get that

  $$\min \left(m\varepsilon'^2, \frac{m^4 \varepsilon'^4}{n^3} \right) \geq \sqrt{\min(n, m)} \, ,$$

  as desired assuming that $\beta \geq 1$.

This completes the proof of Proposition 3.1. □

3.1.2 Analyzing the Variance of A

The main result of this subsection is the following proposition establishing an upper bound on the variance of A as a function of its expectation:

Proposition 3.2. *We have that*

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

for some absolute constant C''.

The rest of this subsection is devoted to the proof of Proposition 3.2. By the law of total variance, we have that:

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

We will proceed to bound each term from above, which will give the proof. We start with the first term. Conditioned on $\sigma_z = |S_z|$, Eq. (4) gives that

$$\operatorname{Var}[A_z \mid \sigma_z] \le C\sigma_z^2 \left(\frac{\delta_z^2}{\sigma_z} + \frac{1}{\sigma_z^2}\right) \mathbb{1}_{\{\sigma_z \ge 4\}} = C\left(1 + \mathbb{E}[\, A_z \mid \sigma_z\,]\right) \mathbb{1}_{\{\sigma_z \ge 4\}} \ .$$

Therefore, for $\sigma := (\sigma_z)_{z \in \mathcal{Z}}$, we can write

$$\operatorname{Var}[A \mid \sigma] \le C\left(\min(n, M) + \mathbb{E}[\, A \mid \sigma\,]\right) \ , \tag{10}$$

where we used the inequality $\sum_{z \in \mathcal{Z}} \mathbb{1}_{\{\sigma_z \ge 4\}} \le \sum_{z \in \mathcal{Z}} \mathbb{1}_{\{\sigma_z \ge 1\}} \le \min(n, M)$. From Eq. (10), we immediately get

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

as desired.

We now proceed to bound the second term. As shown in Lemma 3.1, $\mathbb{E}[\, A \mid \sigma\,] = \sum_{z \in \mathcal{Z}} \sigma_z \delta_z^2 \mathbb{1}_{\{\sigma_z \ge 4\}}$. By the independence of the σ_z's, we obtain that

$$\operatorname{Var}\left[\mathbb{E}[\, A \mid \sigma\,]\right] = \sum_{z \in \mathcal{Z}} \delta_z^4 \operatorname{Var}[\sigma_z \mathbb{1}_{\{\sigma_z \ge 4\}}] \ . \tag{11}$$

From (11) and Claim 2.1, recalling that $\delta_z \le 2$, $z \in \mathcal{Z}$, we get that

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

This completes the proof of Proposition 3.2.

3.1.3 Completing the Proof

Recall that the threshold of the algorithm is defined to be $\tau := \frac{\gamma}{2} \max(m\varepsilon'^2, \frac{m^4\varepsilon'^4}{n^3})$.

In the completeness case, by Proposition 3.1 (a), we have that $\mathbb{E}[A] = 0$. Proposition 3.2 then gives that $\operatorname{Var}[A] \le C'' \cdot \min(n, m)$. Therefore, by Chebyshev's inequality we obtain

$$\Pr[\, A \ge \tau\,] \le \frac{\operatorname{Var}[A]}{\tau^2} \le \frac{4}{\gamma^2} C'' \frac{\min(n, m)}{\max(m\varepsilon'^2, \frac{m^4\varepsilon'^4}{n^3})} \le \frac{1}{3} \ ,$$

where the last inequality follows by choosing the constant β to be sufficiently large (compared to γ, C'').

In the soundness case, by Chebyshev's inequality we get:

$$\Pr[\, A < \tau\,] \le \Pr[\, |A - \mathbb{E}[A]| \ge \mathbb{E}[A]\,/2\,] \le 4\frac{\operatorname{Var}[A]}{\mathbb{E}[A]^2} \le 4C'' \left(\frac{\min(n, m)}{\mathbb{E}[A]^2} + \frac{1}{\mathbb{E}[A]}\right) \le \frac{1}{3} \ ,$$

where the third inequality uses Proposition 3.2 and the fourth inequality uses Proposition 3.1 (b), assuming β is sufficiently large. This completes the proof of correctness. □

4 Estimating a Polynomial of a Discrete Distribution

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 $p \in \Delta([n])$, we want to estimate the quantity $Q(p) = Q(p_1, \ldots, p_n)$ to within an additive error ε. In this section, we analyze an *unbiased* estimator for $Q(p)$ and provide quantitative bounds on its variance.

The structure of this section is as follows: In Section 4.1, we describe the unbiased estimator and establish its basic properties. In Section 4.2, we bound from above the variance of the estimator. Finally, Section 4.3 applies the aforementioned results to the setting that is relevant for our conditional independence tester.

4.1 Unbiased Estimator and its Properties

We start by noting that the general case can be reduced to the case that the polynomial Q is homogeneous.

Remark 4.1 (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 $(\sum_{i=1}^n X_i)^{d-d'}$: since $\sum_{i=1}^n p_i = 1$, this does not affect the value of $Q(p)$.

We henceforth assume Q is a homogeneous polynomial of degree d. Before stating our 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.

The estimators we consider are symmetric, that is only a function of the fingerprint Φ_S. We first focus on the special case $N = d$.

Lemma 4.1. *There exists an unbiased symmetric* linear *estimator for $Q(p)$, i.e., a linear function $U_d^d \colon \mathbb{R}_d[X_1, \ldots, X_n] \to \mathbb{R}_d[X_1, \ldots, X_n]$ such that*

$$\mathbb{E}\Big[U_d^d Q(\Phi_S)\Big] = Q(p) \ ,$$

where S is obtained by drawing d independent samples from p.

Proof. For any ordered d-tuple $T \in [n]^d$, by independence of the samples in S, we see that $\Pr[\, S = T \,] = \prod_{i=1}^n p_{T_i} = p^{\Phi_T}$. For any $\alpha \in \mathbb{N}^n$ with $\|\alpha\| = d$, the number of $T \in [n]^d$ with fingerprint α is $\binom{d}{\alpha}$. Thus, we have that $\Pr[\, \Phi_S = \alpha \,] = \binom{d}{\alpha} p^\alpha$. Noting that since $\|\alpha\| = \|\Phi_S\|$, $\prod_{i=1}^n \binom{\Phi_{S,i}}{\alpha_i} = \delta_{\Phi_S, \alpha}$, we can define

$$U_d^d X^\alpha(\Phi_S) := \binom{d}{\alpha}^{-1} \mathbb{1}_{\{\Phi_S = \alpha\}} = \binom{d}{\alpha}^{-1} \prod_{i=1}^n \binom{\Phi_{S,i}}{\alpha_i} \ . \tag{12}$$

Then we have $\mathbb{E}\Big[U_d^d X^\alpha(\Phi_S)\Big] = p^\alpha$. We extend this linearly to all $Q \in \mathcal{H}_d$. By linearity of expectation, we obtain an unbiased estimator for any such $Q(p)$. □

We can generalize this to $N \geq d$, by taking the average over all subsets of size d of S of the above estimator.

Proposition 4.1 (Existence). *For $N \geq d$ and $Q \in \mathcal{H}_d$ written in terms of monomials as $Q(X) = \sum_\alpha c_\alpha X^\alpha$, the symmetric linear estimator*

$$U_N Q(\Phi_S) := \sum_\alpha c_\alpha \binom{N}{\alpha, N - \|\alpha\|}^{-1} \prod_{i=1}^{n} \binom{X_i}{\alpha_i} \tag{13}$$

is an unbiased estimator for $Q(p)$.

Proof. For the case $N = d$, this follows from Lemma 4.1 and Eq. (12). For any set of d indices $I \subseteq [N], |I| = d$, the subset $S_I = \{ S_i : i \in I \}$ is a set of d independent samples from p, thus $U_d^d Q(\Phi_{S_I})$ is an unbiased estimator for $Q(p)$. To get a symmetric unbiased estimator (and to reduce the variance), we can take the average over all subsets S_I of S of size d. We claim that this estimator is U_N as defined above, i.e., that

$$U_N Q(\Phi_S) = \binom{N}{d}^{-1} \sum_{S' \subseteq S, |S'| = d} U_d^d Q(\Phi_{S'}) \,. \tag{14}$$

By linearity of expectation, the RHS is an unbiased estimator for $Q(p)$, and so (14) suffices to show the proposition. By linearity of U_N and U_d^d, we need to show (14) for each monomial X^α. Noting that the number of subsets S' of S of size d that have fingerprint α is $\prod_{i=1}^n \binom{\Phi_{S,i}}{\alpha_i}$, we have

$$\begin{aligned}
\binom{N}{d}^{-1} \sum_{S' \subseteq S, |S'| = d} U_d^d X^\alpha(\Phi_{S'}) &= \binom{N}{d}^{-1} \sum_{S' \subseteq S, |S'| = d} \binom{d}{\alpha}^{-1} \mathbb{1}_{\{\Phi_{S'} = \alpha\}} \\
&= \binom{N}{\alpha, N - \|\alpha\|}^{-1} \sum_{S' \subseteq S, |S'| = d} \mathbb{1}_{\{\Phi_{S'} = \alpha\}} \\
&= \binom{N}{\alpha, N - \|\alpha\|}^{-1} \prod_{i=1}^{n} \binom{\Phi_{S,i}}{\alpha_i} \\
&= U_N X^\alpha(\Phi_S) \,.
\end{aligned}$$

This completes the proof. □

Proposition 4.2 (Uniqueness). *The unbiased estimator $U_N Q(\Phi_S)$ of (13) is unique among symmetric estimators. That is, for every $N \geq d$, for any symmetric estimator $V_N \colon [n]^N \to \mathbb{R}$ satisfying $\mathbb{E}[V_N(\Phi_S)] = Q(p)$, where S is a multiset of N samples drawn from p, one must have $V_N(\Phi_S) = U_N Q(\Phi_S)$ for all S.*

Proof. 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 $R := (\sum_{i=1}^n X_i)^{N-d} Q$ 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.

When $N = d$, we have a map U_N^N from polynomials to estimators that gives an unbiased estimator for the polynomial. By (12), for $Q(X) = \sum_\alpha c_\alpha X^\alpha$, this is given by

$$U_d^d Q(\Phi_S) = \sum_\alpha c_\alpha \binom{d}{\alpha}^{-1} \mathbb{1}_{\{\Phi_S = \alpha\}} \,.$$

Any symmetric estimator on N samples can by written as a linear combination of $\mathbb{1}_{\{\Phi_S=\alpha\}}$. Hence, given an estimator V_N, we can find a unique polynomial Q_{V_n} with $U_d^d Q_{V_n}(\Phi_S) = V_n$ by choosing c_α to match the coefficients in this linear combination, i.e., U_d^d is a bijection between polynomials and symmetric estimators. Thus, if we have two different N-sample estimators V_N, W_N for a homogeneous degree-N polynomial Q, at least one of them is $U_d^d R$ for some homogeneous degree-N polynomial R.

Now we have an estimator V_N that is unbiased for two different homogeneous degree-N polynomials Q and R. So we get that for every $p \in \Delta([n])$, $Q(p) = \mathbb{E}_S[V_N(\Phi_S)] = R(p)$. Hence, their difference $D := Q - R$ is a non-zero homogeneous degree-N polynomial which vanishes on every point $(x_1, \ldots, x_n) \in \mathbb{N}^n$ with $\sum_{i=1}^n x_i = 1$. By homogeneity, for every non-zero $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}_+^n$,

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

and therefore D vanishes on the whole non-negative quadrant $\mathbb{R}_+^n = \{\, \mathbf{x} \in \mathbb{R}^n : \; x_i \geq 0 \text{ for all } i \,\}$. 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 4.3. *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 $N \geq d$.*

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 4.2) is invertible, this estimator is also an unbiased estimator for some homogeneous degree-N polynomial $R \in \mathcal{H}_N$. Therefore, it is also an 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 4.2, one must then have $Q = R'$, which is impossible since $\sum_{k=1}^n X$ does not divide Q. □

4.2 Bounding the Variance of the Unbiased Estimator

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

Theorem 4.2. *Fix $N \geq d$, and let $U_N \colon \mathbb{R}_d[X_1, \ldots, X_n] \to \mathbb{R}_d[X_1, \ldots, X_n]$ be as above. Then, for every $Q \in \mathcal{H}_d$,*

$$\mathbb{E}\Big[(U_N Q(\Phi_S))^2\Big] = \sum_{\substack{\mathbf{s} \in \mathbb{N}^n \\ \|\mathbf{s}\| \leq d}} p^{\mathbf{s}} \left(\frac{d^{\|\mathbf{s}\|} Q(p)}{dX^{\mathbf{s}}}\right)^2 \frac{(N-d)!^2}{N!(N-2d+\|\mathbf{s}\|)! \prod_{i=1}^n s_i!} \,, \tag{15}$$

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

Proof. In order to establish the identity, we first consider monomials: for $\alpha, \beta \in \mathbb{N}^n$, we will analyze $\mathbb{E}\Big[U_N X^\alpha(\Phi_S)) U_N X^\beta(\Phi_S))\Big]$, before extending it to $Q \in \mathcal{H}_d$, relying on the linearity of U_N. First, note that by definition of U_N (in Eq. (13)),

$$U_N X^\alpha U_N X^\beta = \frac{1}{\binom{N}{\alpha, N-\|\alpha\|}\binom{N}{\beta, N-\|\beta\|}} \prod_{i=1}^n \binom{X_i}{\alpha_i}\binom{X_i}{\beta_i} \,. \tag{16}$$

We will use the following fact:

Claim 4.1. *For $0 \le a, b \le n$, we have*

$$\binom{n}{a}\binom{n}{b} = \sum_{s=0}^{\min(a,b)} \binom{n}{a+b-s}\binom{a+b-s}{a-s, b-s, s}.$$

Proof. The left-hand-side $\binom{n}{a}\binom{n}{b}$ is the number of subsets A,B of $[n]$ with $|A| = a$,$|B| = b$. We can group the set of such pairs of subsets by the size of their intersection $s = |A \cap B|$. Summing the size of these classes gives that

$$\binom{n}{a}\binom{n}{b} = \sum_{s=0}^{\min(a,b)} \binom{n}{a-s, b-s, s, n-(a+b)+s},$$

which is easily seen to be equivalent to the claim by multiplying out the factorials. □

We can then rewrite, combining Eq. (16) and Claim 4.1, and setting $\pi_{\alpha,\beta} := \frac{1}{\binom{N}{\alpha, N-\|\alpha\|}\binom{N}{\beta, N-\|\beta\|}}$ for convenience, that

$$\begin{aligned} U_N X^\alpha U_N X^\beta &= \pi_{\alpha,\beta} \prod_{i=1}^{n} \sum_{s=0}^{\min(\alpha_i,\beta_i)} \binom{X_i}{\alpha_i+\beta_i-s}\binom{\alpha_i+\beta_i-s}{\alpha_i-s, \beta_i-s, s} \\ &= \pi_{\alpha,\beta} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le\min(\alpha,\beta)}} \prod_{i=1}^{n} \binom{X_i}{\alpha_i+\beta_i-s_i}\binom{\alpha_i+\beta_i-s_i}{\alpha_i-s_i, \beta_i-s_i, s_i}. \end{aligned}$$

Taking the expectation over an N-sample multiset S, we obtain

$$\mathbb{E}\Big[U_N X^\alpha(\Phi_S) U_N X^\beta(\Phi_S)\Big] = \pi_{\alpha,\beta} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le\min(\alpha,\beta)}} \prod_{i=1}^{n} \binom{\alpha_i+\beta_i-s_i}{\alpha_i-s_i, \beta_i-s_i, s_i} \mathbb{E}\left[\prod_{i=1}^{n} \binom{\Phi_{S,i}}{\alpha_i+\beta_i-s_i}\right].$$

Recalling the proof of Proposition 4.1, we have

$$\begin{aligned} \mathbb{E}\left[\prod_{i=1}^{n} \binom{\Phi_{S,i}}{\alpha_i+\beta_i-s_i}\right] &= \binom{N}{\alpha+\beta-\mathbf{s}, N-(\|\alpha\|+\|\beta\|-\|\mathbf{s}\|)} \mathbb{E}\Big[U_N X^{\alpha+\beta-\mathbf{s}}(\Phi_S)\Big] \\ &= \binom{N}{\alpha+\beta-\mathbf{s}, N-(\|\alpha\|+\|\beta\|-\|\mathbf{s}\|)} p^{\alpha+\beta-\mathbf{s}}, \end{aligned}$$

leading to

$$\begin{aligned} &\mathbb{E}[U_N X^\alpha(\Phi_S)\ U_N X^\beta(\Phi_S)\Big] \\ &\quad= \pi_{\alpha,\beta} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le\min(\alpha,\beta)}} p^{\alpha+\beta-\mathbf{s}} \binom{N}{\alpha+\beta-\mathbf{s}, N-(\|\alpha\|+\|\beta\|-\|\mathbf{s}\|)} \prod_{i=1}^{n} \binom{\alpha_i+\beta_i-s_i}{\alpha_i-s_i, \beta_i-s_i, s_i} \\ &\quad= \pi_{\alpha,\beta} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le\min(\alpha,\beta)}} p^{\alpha+\beta-\mathbf{s}} \binom{N}{\alpha-\mathbf{s}, \beta-\mathbf{s}, \mathbf{s}, N-(\|\alpha\|+\|\beta\|-\|\mathbf{s}\|)}. \end{aligned}$$

To get a better hold on this expression and extend the analysis to general homogeneous polynomials (instead of monomials), we first massage the expression above under the additional constraint that $\|\alpha\| = \|\beta\| = d$.

$$\begin{aligned}
&\mathbb{E}[U_N X^\alpha(\Phi_S)\ U_N X^\beta(\Phi_S)\Big] \\
&\qquad = \frac{(N-d)!^2}{N!}\prod_{i=1}^n \alpha_i! \prod_{i=1}^n \beta_i! \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le \min(\alpha,\beta)}} \frac{p^{\alpha+\beta-\mathbf{s}}}{(N-2d+\|\mathbf{s}\|)!\prod_{i=1}^n(\alpha_i-s_i)!(\beta_i-s_i)!s_i!} \\
&\qquad = \frac{(N-d)!^2}{N!} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le \min(\alpha,\beta)}} \frac{p^{\alpha+\beta-\mathbf{s}}}{(N-2d+\|\mathbf{s}\|)!\prod_{i=1}^n s_i!} \prod_{i=1}^n \frac{\alpha_i!}{(\alpha_i-s_i)!} \prod_{i=1}^n \frac{\beta_i!}{(\beta_i-s_i)!}\ .
\end{aligned}$$

Recalling that $\frac{d^{\|\mathbf{s}\|}X^\alpha}{dX^\mathbf{s}} = \prod_{i=1}^n \frac{\alpha_i!}{(\alpha_i-s_i)!} X_i^{\alpha_i-s_i}$ for $\mathbf{s}\le\alpha$, we have

$$p^\mathbf{s}\frac{d^{\|\mathbf{s}\|}p^\alpha}{dX^\mathbf{s}}\frac{d^{\|\mathbf{s}\|}p^\beta}{dX^\mathbf{s}} = p^{\alpha+\beta-\mathbf{s}}\prod_{i=1}^n\frac{\alpha_i!}{(\alpha_i-s_i)!}\prod_{i=1}^n\frac{\beta_i!}{(\beta_i-s_i)!}$$

for $\mathbf{s}\le\min(\alpha,\beta)$, from which

$$\mathbb{E}\Big[U_N X^\alpha(\Phi_S) U_N X^\beta(\Phi_S)\Big] = \frac{(N-d)!^2}{N!}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \mathbf{s}\le\min(\alpha,\beta)}} \frac{p^\mathbf{s}}{(N-2d+\|\mathbf{s}\|)!\prod_{i=1}^n s_i!}\frac{d^{\|\mathbf{s}\|}p^\alpha}{dX^\mathbf{s}}\frac{d^{\|\mathbf{s}\|}p^\beta}{dX^\mathbf{s}}\ .$$

By linearity of U and differentiation, this implies that, for any $Q, R\in\mathcal{H}_d$,

$$\mathbb{E}[U_N Q(\Phi_S) U_N R(\Phi_S)] = \frac{(N-d)!^2}{N!}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|\le d}} \frac{p^\mathbf{s}}{(N-2d+\|\mathbf{s}\|)!\prod_{i=1}^n s_i!}\frac{d^{\|\mathbf{s}\|}Q(p)}{dX^\mathbf{s}}\frac{d^{\|\mathbf{s}\|}R(p)}{dX^\mathbf{s}}\ .$$

Choosing $R = Q$ yields Eq. (15). □

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

$$\mathbb{E}\Big[(U_N Q(\Phi_S))^2\Big] = \sum_{h=0}^d \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} p^\mathbf{s}\left(\frac{d^h Q(p)}{dX^\mathbf{s}}\right)^2 \frac{(N-d)!^2}{N!(N-2d+h)!\prod_{i=1}^n s_i!} = \sum_{h=0}^d T_h(Q,p,d,N)\ ,$$

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

Corollary 4.1. *For $h\ge 0$,*

$$T_h(Q,p,d,N) \le \frac{(N-h)!}{N!}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} p^\mathbf{s}\left(\frac{d^h Q(p)}{dX^\mathbf{s}}\right)^2 \frac{1}{\prod_{i=1}^n s_i!}\ ,$$

and so

$$\operatorname{Var} U_N Q(\Phi_S) \le \sum_{h=1}^d \frac{(N-h)!}{N!}\sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} p^\mathbf{s}\left(\frac{d^h Q(p)}{dX^\mathbf{s}}\right)^2 \frac{1}{\prod_{i=1}^n s_i!}\ .$$

Proof. We have

$$\frac{(N-d)!^2}{N!(N-2d+h)!} = \prod_{i=1}^{d-h} \frac{N-2d+h+i}{N-d+i} \prod_{j=0}^{h-1} \frac{1}{N-j} \leq \frac{(N-h)!}{N!}$$

which gives the bound on $T_h(Q,p,d,N)$. For $h=0$, this gives that $T_0(Q,p,d,N) \leq Q(p)^2 = \mathbb{E}[U_N U_N Q(\Phi_S)]^2$ and so if we expand $\operatorname{Var} U_N Q(\Phi_S) = \mathbb{E}[U_N U_N Q(\Phi_S)^2] - \mathbb{E}[U_N U_N Q(\Phi_S)]^2$, the $T_0(Q,p,d,N)$ term is at least cancelled by the $-\mathbb{E}[U_N U_N Q(\Phi_S)]^2$. □

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 4.2. *Fix any $0 \leq g \leq d$. Then,*

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

Proof. For g as above, we have

$$\begin{aligned}\sum_{h=g}^{d} T_h(Q,p,d,N) &= \sum_{h=g}^{d} O\Big(\frac{1}{N^h}\Big) \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} \frac{1}{\prod_{i=1}^n s_i!} p^{\mathbf{s}} \left(\frac{d^h Q(p)}{dX^{\mathbf{s}}}\right)^2 \\ &= O\Big(\frac{1}{N^g}\Big) \sum_{h=g}^{d} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} \frac{1}{\prod_{i=1}^n s_i!} p^{\mathbf{s}} \left(\frac{d^h Q(p)}{dX^{\mathbf{s}}}\right)^2 .\end{aligned}$$

A useful observation is that since Q is homogeneous of degree d, then so is $X^{\mathbf{s}} \frac{d^h Q}{dX^{\mathbf{s}}}$ for every $\mathbf{s}$. Consider a term $c_\alpha X^\alpha$ in Q; a term X^α will appear in $X^{\mathbf{s}} \left|\frac{d^h Q}{dX^{\mathbf{s}}}\right|$ if and only if $\alpha \geq \mathbf{s}$, in which case this term will be

$$\left(\prod_{i=1}^n \frac{\alpha_i!}{(\alpha_i - s_i)!}\right) |c_\alpha| X^\alpha = |c_\alpha| X^\alpha \prod_{i=1}^n s_i! \prod_{i=1}^n \binom{\alpha_i}{s_i} .$$

Therefore, summing over all $\mathbf{s}$, we get

$$\sum_{h=g}^{d} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} \frac{1}{\prod_{i=1}^n s_i!} X^{\mathbf{s}} \left|\frac{d^h (c_\alpha X^\alpha)}{dX^{\mathbf{s}}}\right| = \sum_{h=g}^{d} \sum_{\substack{\mathbf{s}\leq\alpha \\ \|\mathbf{s}\|=h}} |c_\alpha| X^\alpha \prod_{i=1}^n \binom{\alpha_i}{s_i} \leq \sum_{\mathbf{s}\leq\alpha} |c_\alpha| X^\alpha \prod_{i=1}^n \binom{\alpha_i}{s_i} = 2^d |c_\alpha| X^\alpha ,$$

where the inequality is an abuse of notation, assuming X is a non-negative vector. For the last equality, we relied on the facts that $\|\alpha\| = d$ and

$$\sum_{\mathbf{s}\leq\alpha} \prod_{i=1}^n \binom{\alpha_i}{s_i} = \prod_{i=1}^n \sum_{s:s\leq\alpha_i} \binom{\alpha_i}{s} = \prod_{i=1}^n 2^{\alpha_i} = 2^{\|\alpha\|} .$$

By linearity and the definition of Q^+, this yields

$$\sum_{h=g}^{d} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} \frac{1}{\prod_{i=1}^n s_i!} p^{\mathbf{s}} \left| \frac{d^h Q(p)}{dX^{\mathbf{s}}} \right| \le 2^d Q^+(p) \ ,$$

and thus

$$\begin{aligned} \sum_{h=g}^{d} T_h(Q,p,d,N) &= O\left(\frac{1}{N^g}\right) \sum_{h=g}^{d} \sum_{\substack{\mathbf{s}\in\mathbb{N}^n \\ \|\mathbf{s}\|=h}} \frac{1}{\prod_{i=1}^n s_i!} p^{\mathbf{s}} \left(\frac{d^h Q(p)}{dX^{\mathbf{s}}} \right)^2 \\ &\le O\left(\frac{1}{N^g}\right) 2^d Q^+(p) \max_{\mathbf{s}:\|\mathbf{s}\|\ge g} \left| \frac{d^h Q(p)}{dX^{\mathbf{s}}} \right| \ . \end{aligned}$$

This completes the proof. □

Remark 4.3 (Poissonized case). Frequently, in distribution testing we analyze Poissonized statistics. That is, instead of S being a set of N samples, we consider a set S of $\mathrm{Poisson}(N)$ samples. In this case, $\Phi_{S,i}$ is independent for different i's and $\mathbb{E}\left[\prod_{i=1}^n \binom{S_i}{\alpha_i}\right] = p^\alpha \prod_{i=1}^n \frac{N}{\alpha_i!}$. Thus, we can define an unbiased estimator for $U'_N Q$ for a polynomial $Q(p)$ by taking linear combinations of $U'_N X^\alpha(\Phi_S) = N^{-\|\alpha\|} \prod_{i=1}^n \binom{S_i}{\alpha_i} \alpha_i!$. The theory in the Poissonized setting is a little different: this estimator is not unique and is unbiased for any $N > 0$, including non-integral N and $N < d$. However, the expression for $\mathbb{E}\left[U'_N X^\alpha(\Phi_S)^2\right]$ is very similar, and is obtained by an analogous proof. The difference is that we obtain a term N^{-h} instead of $\frac{(N-d)!^2}{N!(N-2d+h)!}$. The bound on the variance in Corollary 4.1 holds for the unbiased estimators in both the Poissonized and non-Poissonized cases.

4.3 Case of Interest: ℓ_2-Distance between p and $p_{\mathcal{X}} \otimes p_{\mathcal{Y}}$

We now instantiate the results of the previous subsections 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 $|\mathcal{X}| = \ell_1$, $|\mathcal{Y}| = \ell_2$, 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, let $Q \in \mathbb{R}_4[X_{1,1}, X_{2,1}, \ldots, X_{\ell_1,1}, X_{\ell_1,2}, \ldots, X_{\ell_1,\ell_2}]$ be the degree-4 $(\ell_1\ell_2)$-variate polynomial defined as

$$Q(X_{1,1}, \ldots, X_{\ell_1,\ell_2}) := \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \left(X_{i,j} \sum_{i'\neq i} \sum_{j'\neq j} X_{i',j'} - \sum_{i'\neq i} X_{i',j} \sum_{j'\neq j} X_{i,j'} \right)^2 . \tag{17}$$

An explicit expression for its unbiased estimator $U_N Q(\Phi_S)$ will be given in Eq. (18). Specifically, we shall prove the following result:

Proposition 4.4. *Let Q be as in Eq. (17), and suppose that $b \ge \max(\|p\|_2^2, \|p_\Pi\|_2^2)$. Then, for $N \ge 4$,*

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

For consistency of notation with the previous section, we let $n := \ell_1\ell_2$ in what follows.

Claim 4.2. *For any p over $\mathcal{X} \times \mathcal{Y}$, we have $Q(p) = \|p - p_\Pi\|_2^2$.*

Proof. Unraveling the definitions, we can write

$$\begin{aligned}
\|p - p_\Pi\|_2^2 &= \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} (p(i,j) - p_\Pi(i,j))^2 = \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} \left(p(i,j) - \sum_{j'=1}^{\ell_2} p(i,j') \sum_{i'=1}^{\ell_1} p(i',j)\right)^2 \\
&= \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} \left(p(i,j) - \left(p(i,j) + \sum_{j' \neq j} p(i,j')\right)\left(p(i,j) + \sum_{i' \neq i} p(i',j)\right)\right)^2 \\
&= \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} \left(p(i,j)\left(1 - p(i,j) - \sum_{i' \neq i} p(i',j) - \sum_{j' \neq j} p(i,j')\right) - \sum_{i' \neq i} p(i',j) \sum_{j' \neq j} p(i,j')\right)^2 \\
&= \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} \left(p(i,j) \sum_{i' \neq i}\sum_{j' \neq j} p(i',j') - \sum_{i' \neq i} p(i',j) \sum_{j' \neq j} p(i,j')\right)^2 = Q(p) \ ,
\end{aligned}$$

as claimed. □

Firstly, we compute $U_N Q$ explicitly. By linearity of U_N, we can compute the unbiased estimator for each term separately, after writing $Q(X) = \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} \Delta_{ij}(X)^2$, 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 $U_N Q = \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2} U_N \Delta_{ij}^2$ and we want to compute $U_N \Delta_{ij}^2$. Note that the sums in $\Delta_{ij}(X)$ are over disjoint sets of $X_{i,j}$'s whose union is every $X_{i,j}$. 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 $\ell_1\ell_2$-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

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

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

and

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

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:

$$\begin{aligned}
\frac{N!}{(N-4)!} U_N Q(\Phi_S) &= \frac{N!}{(N-4)!} \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} U_N \Delta_{ij}(\Phi_S)^2 \\
&= \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} (\Phi_{S,i,j}(\Phi_{S,i,j} - 1)\Phi_{S,-i,-j}(\Phi_{S,-i,-j} - 1) \\
&\quad + \Phi_{S,-i,j}(\Phi_{S,-i,j} - 1)\Phi_{S,i,-j}(\Phi_{S,i,-j} - 1) - 2\Phi_{S,i,j}\Phi_{S,-i,-j}\Phi_{S,i,-j}\Phi_{S,-i,j}) \\
&= \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \Big((\Phi_{S,i,j}\Phi_{S,-i,-j} - \Phi_{S,-i,j}\Phi_{S,-i,-j})^2 \\
&\quad + \Phi_{S,i,j}\Phi_{S,-i,-j}(1 - \Phi_{S,i,j} - \Phi_{S,-i,-j}) + \Phi_{S,-i,j}\Phi_{S,i,-j}(1 - \Phi_{S,-i,j} - \Phi_{S,i,-j})) \Big) ,
\end{aligned} \tag{18}$$

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 4.2, we then have that, for $N \geq 4$,

$$\mathbb{E}\Big[(U_N Q(\Phi_S))^2\Big] = \sum_{h=0}^{4} \sum_{\substack{\mathbf{s} \in \mathbb{N}^n \\ \|\mathbf{s}\| = h}} \binom{h}{\mathbf{s}} p^{\mathbf{s}} \left(\frac{d^h Q(p)}{dX^{\mathbf{s}}} \right)^2 \binom{N-4}{4-h} \binom{N}{h, 4-h, N-4}^{-1} \frac{1}{h!^2} . \tag{19}$$

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 $Q(X) = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \Delta_{ij}(X)^2$, 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 $O(Q(p)^2/N)$ by Corollary 4.1, so we have T_0 under control. Indeed,
$$Q(p) \leq 2\sqrt{b}$$
by the triangle inequality and the definition of b. So, $T_0(N) - Q(p)^2 = O\Big(Q(p)\sqrt{b}/N\Big)$.
- The second, $h = 1$, contributes

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

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

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

Recalling the expression of the derivatives of Δ_{ij}, we have that

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

Having this sum of four terms A_1, A_2, A_3, A_4 for $\frac{dQ}{dX_{k,\ell}}$, by Cauchy–Schwarz it holds that $\left(\frac{dQ}{dX_{k,\ell}}\right)^2 \leq 4(A_1^2 + A_2^2 + A_3^2 + A_4^2)$, 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)

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

  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

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

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

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

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

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

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

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

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

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

- Finally, for the rest of the contributions ($h \geq 2$), we invoke Lemma 4.2. Specifically, we first observe

that, for any distribution $p \in \Delta(\mathcal{X} \times \mathcal{Y})$,

$$Q^+(p) = \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_2}\left(p_{i,j}\sum_{i'\neq i}\sum_{j'\neq j}p_{i',j'} + \sum_{i'\neq i}p_{i',j}\sum_{j'\neq j}p_{i,j'}\right)^2 \leq \sum_{i,j}\left(p_{i,j} + \sum_{i'=1}^{\ell_1}p_{i',j}\sum_{j'=1}^{\ell_2}p_{i,j'}\right)^2$$
$$\leq 2\sum_{i,j}\left(p_{i,j}^2 + \left(\sum_{i'=1}^{\ell_1}p_{i',j}\right)^2\left(\sum_{j'=1}^{\ell_2}p_{i,j'}\right)^2\right) \leq 2\left(\|p\|_2^2 + \|p_\Pi\|_2^2\right) \leq 4b\,.$$

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

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

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

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

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 $\left|\frac{d^hQ}{dX^{\mathbf{s}}}\right| \leq 2^4\cdot 4$ 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 can thus bound the variance as

$$\operatorname{Var}U_NQ(\Phi_S) = (T_0(N) - Q(p)^2) + T_1(N) + \sum_{h=2}^{4}T_h(N) = O\left(\frac{Q(p)\sqrt{b}}{N} + \frac{b}{N^2}\right),$$

concluding the proof of Proposition 4.4.

Remark 4.4 (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 4). We defer these results, which follow from a straightforward (yet slightly cumbersome) adaptation of the proofs of this section, to an updated version of this paper.

5 The General Conditional Independence Tester

In this section, we present and analyze our general algorithm for testing conditional independence. The structure of this section is as follows: In Section 5.1, we begin by describing how we flatten the marginals of the distribution p_z, for each bin z for which we receive enough samples. After this flattening is performed, in Section 5.2 we explain how we use the remaining samples for each such bin z to compute a statistic A as an appropriate weighted sum of bin-wise statistics A_z. Before going further, we discuss in Section 5.3 the eventual result our analysis yields and comment on the sample complexity bound of our algorithm. In Section 5.4, we explain the three different sources of randomness involved in our estimator, in order to clarify what will follow – as we will crucially later condition on part of this randomness to obtain bounds on some of its conditional expectations and variances. Section 5.5 then details how the analysis of our statistic A is performed (Sections 5.5.1 and 5.5.2 respectively contain the analysis of the expectation and variance of A, conditioned on *some* of the randomness at play). Finally, Section 5.6 puts everything together and derives the correctness guarantee of our overall algorithm.

5.1 Flattening $\mathcal{X}, \mathcal{Y}$ for any Given Bin z

Given a multiset S of $N \geq 4$ independent samples from $p \in \Delta(\mathcal{X} \times \mathcal{Y})$, where $|\mathcal{X}| = \ell_1$, $|\mathcal{Y}| = \ell_2$, we perform the following. Losing at most three samples, we can assume $N = 4 + 4t$ for some integer t. Let $t_1 := \min(t, \ell_1)$ and $t_2 := \min(t, \ell_2)$. We divide S into two disjoint multi-sets $S_{\mathcal{F}}, S_{\mathcal{T}}$ of size $t_1 + t_2$ and $2t + 4$ respectively, where the subscripts $\mathcal{F}$ and $\mathcal{T}$ stand for *Flatten* and *Test*.

- We use $S_{\mathcal{F}}$ to flatten $\mathcal{X} \times \mathcal{Y}$, as per Definition 2.2. Namely, first we partition it into two multi-sets $S^1_{\mathcal{F}}$, $S^2_{\mathcal{F}}$ of size t_1, t_2. Looking at the projections $\pi_{\mathcal{X}} S^1_{\mathcal{F}}$, $\pi_{\mathcal{Y}} S^2_{\mathcal{F}}$ of $S^1_{\mathcal{F}}, S^2_{\mathcal{F}}$ onto $\mathcal{X}$ and $\mathcal{Y}$ respectively, we have two multi-sets of t_1 and t_2 elements. We then let $T \subseteq \mathcal{X} \times \mathcal{Y}$ obtained by, for each $(x, y) \in \mathcal{X} \times \mathcal{Y}$, adding in T $a_{x,y}$ copies of (x, y), where

$$1 + a_{x,y} := (1 + a_x)(1 + a'_y) \tag{20}$$

  with a_x (resp. a'_y) being the number of occurrences of x in $\pi_{\mathcal{X}} S^1_{\mathcal{F}}$ (resp. of y in $\pi_{\mathcal{Y}} S^2_{\mathcal{F}}$). Note that $|T| + \ell_1 \ell_2 = (|\mathcal{X}| + t_1)(|\mathcal{Y}| + t_2)$, and that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, by a similar proof as that of Lemma 2.3 (using the fact that a_x and a'_y are independent),

$$\mathbb{E}\left[\frac{1}{1 + a_{x,y}}\right] = \mathbb{E}\left[\frac{1}{1 + a_x}\right] \mathbb{E}\left[\frac{1}{1 + a_y}\right] \leq \frac{1}{(1 + t_1)(1 + t_2) p_{\mathcal{X}}(x) p_{\mathcal{Y}}(x)} ,$$

  and so, letting q_T denote the product of the marginals of p_T,

$$\mathbb{E}\left[\|q_T\|_2^2\right] \leq \frac{1}{(1 + t_1)(1 + t_2)} . \tag{21}$$

- Next, we use the $2t + 4 \geq 4$ samples from $S_{\mathcal{T}}$ to estimate the squared ℓ_2-distance between p_T and q_T, as per Section 4. Here, Remark 2.3 will come in handy, as it allows us to do it implicitly without having to actually map p to p_T. Indeed, recalling that the polynomial Q for which we wish to estimate $Q(p_T)$ is of the form

$$Q(X) = \sum_{(i,j) \in \mathcal{X} \times \mathcal{Y}} \Delta_{ij}(X)^2 ,$$

we will instead estimate $R_T(p)$, where R is defined as

$$R_T(X) := \sum_{(i,j)\in\mathcal{X}\times\mathcal{Y}} c_{i,j}\Delta_{ij}(X)^2$$

with $c_{i,j} := \frac{1}{1+a_{i,j}}$ for all $(i,j) \in \mathcal{X}\times\mathcal{Y}$. From Remark 2.3, it is immediate that $R_T(p) = Q(p_T) = \|p_T - q_T\|_2^2$, and further by inspection of the proof of Proposition 4.4 it is not hard to see that the variance of our estimator $U_N R_T$ on p is the same as that of $U_N Q$ on p_T.

Let $B := \|q_T\|_2^2$. Note that B is a random variable, determined by the choice of $S_{\mathcal{F}}$. The first observation is that, while the statement of Proposition 4.4 would be with regard to the maximum of $\|p_T\|_2^2$, $\|q_T\|_2^2$, we would like to relate it to B. To do so, observe that

$$\|p_T\|_2^2 \le (\|q_T\|_2 + \|p_T - q_T\|_2)^2 \le 2\left(\|q_T\|_2^2 + \|p_T - q_T\|_2^2\right) = 2\left(B + Q(p_T)\right)$$

so we can use $B' := 2B + 2Q(p_T)$ instead of our original bound B.

Therefore, our bound B can be used in the statement of Proposition 4.4, leading to a variance for our estimator of

$$\mathrm{Var}\left[U_N R_T\right] = O\left(\frac{Q(p_T)\sqrt{B'}}{N} + \frac{B'}{N^2}\right) = O\left(\frac{Q(p_T)\sqrt{B}}{N} + \frac{Q(p_T)^{3/2}}{N} + \frac{B}{N^2}\right). \tag{22}$$

Now, recall that by Lemma 2.3 (more precisely, Eq. (21)), we only have a handle on the *expectation* of B. We could try to first obtain instead a high-probability bound on its value by proving sufficiently strong concentration followed by a union bound over all estimators that we may run (i.e., all n bins in $\mathcal{Z}$). However, this would lead to a rather unwieldy argument. Instead, as outlined in Section 5.5, we will analyze our estimators by carefully conditioning on some of the randomness (the one underlying the flattening we perform for each bin), and only convert the bounds obtained into high-probability statements at the end, by a combination of Markov's and Chebyshev's inequalities.

5.2 From Flattening to an Algorithm

We now explain how the guarantees established above are sufficient to use in our algorithm. We will use the same notations as above, but now specifying the bin $z \in \mathcal{Z}$: that is, we will write $p_z, q_z, T_z, p_{z,T_z}, q_{z,T_z}$ instead of p, q, T, p_T, q_T to make the dependence on the bin we condition on explicit. In what follows, we write $\sigma = (\sigma_z)_{z\in\mathcal{Z}}$, $T = (T_z \mid \sigma_z)_{z\in\mathcal{Z}}$.

We let

$$A_z := \sigma_z \cdot \omega_z \cdot \Phi(S_z) \cdot \mathbb{1}_{\{\sigma_z \ge 4\}} ,$$

for all $z \in \mathcal{Z}$, where $\omega_z := \sqrt{\min\left(\sigma_z, \ell_1\right)\min\left(\sigma_z, \ell_2\right)}$. Our final statistic is

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

That is, compared to algorithm of Section 3, we now re-weight the statistics by $\sigma_z\omega_z$ instead of σ_z (since, intuitively, the flattening is done with "$t_{1,z}, t_{2,z}$" samples for which $\sqrt{t_{1,z}t_{2,z}} = \Theta(\omega_z)$ samples, we multiply the weight by the "flattening amount").

Recalling that $\ell_1 \geq \ell_2$ without loss of generality, we set

$$m \geq \zeta \max\Bigg(\min\left(\frac{n^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon}, \frac{n^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}}, \frac{n\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon} \right), \min\left(\frac{n^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon}, \frac{\ell_1^2\ell_2^2}{\varepsilon^4}, \frac{n\ell_1^{1/2}\ell_2^{3/2}}{\varepsilon} \right), \\ \min\left(\frac{n^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}}, \frac{\ell_1\ell_2}{\varepsilon^4}, \frac{\sqrt{n}\ell_1\sqrt{\ell_2}}{\varepsilon^2}, \frac{n\ell_1^{3/2}\ell_2^{1/2}}{\varepsilon} \right), \min\left(\frac{\sqrt{n\ell_1\ell_2}}{\varepsilon^2}, \frac{\ell_1\ell_2}{\varepsilon^4} \right) \Bigg) , \tag{23}$$

for some sufficiently big absolute constant $\zeta \geq 1$. The resulting pseudo-code is given in Algorithm 2.

Algorithm 2 TESTCONDINDEPENDENCEGENERAL

Require: Parameter $n := |\mathcal{Z}|$, $\ell_1 := |\mathcal{X}|$, $\ell_2 := |\mathcal{Y}|$, $\varepsilon \in (0,1]$, and sample access to $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$.

Set m as in Eq. (23) ▷ $\zeta \geq 1$ is an absolute constant
Set $\tau \leftarrow \zeta^{1/4}\sqrt{\min(n,m)}$. ▷ Threshold for accepting
Draw $M \sim \text{Poisson}(m)$ 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 Φ
    Set $N_z \leftarrow 4\lfloor (|S_z| - 4)/4 \rfloor$, and let S'_z be the multi-set of the first N_z elements of S_z. ▷ $N_z = 4 + 4t_z$ for some integer t_z.
    Set $t_{1,z} \leftarrow \min(t_z, \ell_1)$, $t_{2,z} \leftarrow \min(t_z, \ell_2)$, and divide S'_z into disjoint $S'_{\mathcal{F},z}$, $S'_{\mathcal{T},z}$ of size $t_{1,z} + t_{2,z}$ and $\sigma_z := 2t_z + 4$, respectively.
    $(a^{(z)}_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \leftarrow$ IMPLICITFLATTENING$(S'_{\mathcal{F},z})$ ▷ Flatten $\mathcal{X} \times \mathcal{Y}$ using $S'_{\mathcal{F},z}$ as explained in the first bullet of Section 5.1, by calling Algorithm 3
    $\Phi_z \leftarrow$ UNBIASEDESTIMATOR$((a^{(z)}_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}}, S'_{\mathcal{F},z})$ ▷ Compute $\Phi(S'_{\mathcal{F},z})$, the unbiased estimator of Q as defined in the second bullet of Section 5.1, by calling Algorithm 4
    Set $A_z \leftarrow \sigma_z\omega_z \cdot \Phi_z$, where $\omega_z \leftarrow \sqrt{\min(\sigma_z, \ell_1)\min(\sigma_z, \ell_2)}$.
  else
    Set $A_z \leftarrow 0$.
  end if
end for
if $A := \sum_{z\in\mathcal{Z}} A_z \geq \tau$ **then**
  return accept
else
  return reject
end if

5.3 Discussion of the Sample Complexity

The expression of our sample complexity in Eq. (23) may seem rather complicated. We argue here that it captures at least some of the regimes of our four parameters in a tight way:

- For $\ell_1 = \ell_2 = 2$, we fall back to the case $\mathcal{X} = \mathcal{Y} = \{0,1\}$, for which we had proven a tight bound in Section 3. Note that in this case the expression of m in Eq. (23) reduces to

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

  matching the bounds of Section 3.

Algorithm 3 IMPLICITFLATTENING

Require: Multi-set $S \subseteq \mathcal{X} \times \mathcal{Y}$.

▷ This simulates the construction of the "flattening set" as per Section 5.1; by Remark 2.3, it is actually sufficient to compute the corresponding normalization coefficients $a_{x,y}$, which we perform below.

▷ All b_x and c_y are initialized to 0

for all $(x, y) \in S$ **do**

  $b_x \leftarrow b_x + 1$

  $c_y \leftarrow c_y + 1$

end for

  ▷ Note that the step below can be done more efficiently by only looping through elements (x, y) for which either b_x or c_y is positive

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ **do**

  $a_{x,y} \leftarrow (1 + b_x)(1 + c_y) - 1$ ▷ Implement Eq. (20)

end for

return $(a_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$

Algorithm 4 UNBIASEDESTIMATOR

Require: Set of coefficients $(a_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$, multi-set of samples $S \subseteq \mathcal{X} \times \mathcal{Y}$.

▷ This computes the unbiased estimator $U_N R_T$ for $Q(p_T) = R_T(p)$ from the samples in S, as explained in Section 5.1: where

$$R_T(X) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{1}{1 + a_{x,y}} \Delta_{x,y}(X)^2$$

Let $N \leftarrow |S|$.

▷ Recall that $\Phi_{S,x,y}$ denotes the count of occurrences of (x, y) in the multi-set S

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ **do** ▷ Compute for $U_N \Delta_{x,y}(\Phi_S)^2$, from Eq. (18)

  $\Phi_{S,-x,-y} \leftarrow \sum_{x'\neq x} \sum_{y'\neq y} \Phi_{S,x',y'}$

  $\Phi_{S,-x,y} \leftarrow \sum_{x'\neq x} \Phi_{S,x',y}$

  $\Phi_{S,x,-y} \leftarrow \sum_{y'\neq y} \Phi_{S,x,y'}$

  $C_{i,j} \leftarrow (\Phi_{S,i,j}\Phi_{S,-i,-j} - \Phi_{S,-i,j}\Phi_{S,-i,-j})^2 + \Phi_{S,i,j}\Phi_{S,-i,-j}(1 - \Phi_{S,i,j} - \Phi_{S,-i,-j}) + \Phi_{S,-i,j}\Phi_{S,i,-j}(1 - \Phi_{S,-i,j} - \Phi_{S,i,-j})$

end for

return $\frac{(N-4)!}{N!} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{1}{1+a_{x,y}} C_{i,j}$

- For $n = 1$ (and $\ell_1 \geq \ell_2$ as before) we fall back to the independence testing problem [BFF+01, LRR11, ADK15, DK16], for which the tight sample complexity is known to be $\Theta\left(\max\left(\ell_1^{2/3}\ell_2^{1/3}/\varepsilon^{4/3}, \sqrt{\ell_1\ell_2}/\varepsilon^2\right)\right)$ [DK16]. It is easy to see that, with these parameters, Eq. (23) reduces to $O\left(\max\left(\ell_1^{2/3}\ell_2^{1/3}/\varepsilon^{4/3}, \sqrt{\ell_1\ell_2}/\varepsilon^2\right)\right)$ as well.
- For $\ell_1 = \ell_2 = n$ (and ε not too small), the choice of m reduces to $O(n^{7/4}/\varepsilon)$. This matches the $\Omega\left(n^{7/4}\right)$ lower bound of Section 7 for $\varepsilon = 1/20$.

Remark 5.1. We further note that the expression of Eq. (23), which emerges from the analysis, can be simplified by a careful accounting of the regimes of the parameters. Namely, one can show that it is equivalent

to

$$m \geq \beta \max\left(\min\left(\frac{n^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon}, \frac{n^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}}\right), \frac{n^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon}, \frac{n^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon^2}\right). \quad (24)$$

5.4 The Different Sources of Randomness

As the argument will heavily rely on conditioning on *some* of the randomness at play and analyzing the resulting conditional expectations and variances, it is important to clearly state upfront what the different sources of randomness are and how we refer to them.

In what follows, we will use the following notations: for each bin $z \in \mathcal{Z}$,

- σ_z is the number of samples from p we obtain with the Z coordinate falling in bin z;
- T_z is the randomness corresponding to the flattening of $\mathcal{X}, \mathcal{Y}$ for the corresponding bin z (as described in Section 5.1);
- R_z is the randomness of the estimator Φ_{S_z} on bin z.

Accordingly, we will write $\sigma = (\sigma_z)_{z\in\mathcal{Z}}$, $T = (T_z)_{z\in\mathcal{Z}}$, and $R = (R_z)_{z\in\mathcal{Z}}$ for the three sources of randomness (over all bins).

5.5 Analyzing A

The goal of this subsection is to show that, with high probability over σ, T, the following holds:

- If p is indeed conditionally independent, then $\mathbb{E}[\,A \mid \sigma, T\,] = 0$ and $\mathrm{Var}[A \mid \sigma, T] = O(\min(n, m))$.
- If p is far from conditionally independent, then $\mathbb{E}[\,A \mid \sigma, T\,] = \Omega(\sqrt{\min(n, m)})$ and $\mathrm{Var}[A \mid \sigma, T]$ is "not too big"compared to $\min(n, m)$ and $\mathbb{E}[\,A \mid \sigma, T\,]$.

This high-probability guarantee will allow us to use Chebyshev's inequality in Section 5.6 to conclude that, by comparing A to a suitably chosen threshold, we can distinguish between the two cases with high probability (both over σ, T and R).

The reason for which we only obtain the above guarantees "with high probability over σ, T" is, roughly speaking, that we need to handle the complicated dependencies between A and T, which prevent us from analyzing $\mathbb{E}[A]$ and $\mathrm{Var}[A]$ directly. To do so, we introduce an intermediate statistic, D (which itself only depends on σ and R, but not on the flattening randomness T), and relate it to $\mathbb{E}[\,A \mid \sigma, T\,]$. This enables us to analyze the expectation and variance of D instead of $\mathbb{E}[\,A \mid \sigma, T\,]$, before concluding by Markov's and (another application of) Chebyshev's inequality that these bounds carry over to $\mathbb{E}[\,A \mid \sigma, T\,]$ *with high probability over* σ, T.

5.5.1 The Expectation of A

We have that

$$\mathbb{E}[\,A_z \mid \sigma_z, T_z\,] = \sigma_z\omega_z\|p_{T_z} - q_{T_z}\|_2^2\mathbb{1}_{\{\sigma_z\geq 4\}} \quad (25)$$

but $Q(p_{T_z}) = \|p_{T_z} - q_{T_z}\|_2^2$ depends on T_z. To get around this, we will start by analyzing $D := \sum_{z\in\mathcal{Z}} D_z$, where

$$D_z := \sigma_z\omega_z\frac{\varepsilon_z^2}{\ell_1\ell_2}\mathbb{1}_{\{\sigma_z\geq 4\}},$$

and $\varepsilon_z := \mathrm{d_{TV}}(p_z, q_z)$. Note that now D only depends on R and σ (and no longer on T). For simplicity, we will often write $\varepsilon'_z = \frac{\varepsilon_z}{\sqrt{\ell_1 \ell_2}}$. Next we show that whatever flattenings $T = (T_z)_{z \in \mathcal{Z}}$ we use given $\sigma = (\sigma_z)_{z \in \mathcal{Z}}$, D is a lower bound for the conditional expectation of A:

Lemma 5.1. $\mathbb{E}[\, A \mid \sigma, T \,] \geq D$.

Proof. Using that $\|p_{T_z} - q_{T_z}\|_2 \geq \frac{2\varepsilon_z}{\sqrt{(\ell_1 + t_{1,z})(\ell_2 + t_{2,z})}} \geq \frac{\varepsilon_Z}{\sqrt{\ell_1 \ell_2}}$, we have that

$$D = \sum_{z \in \mathcal{Z}} \sigma_z \omega_z \frac{\varepsilon_z^2}{\sqrt{\ell_1 \ell_2}} \mathbb{1}_{\{\sigma_z \geq 4\}} \leq \sum_{z \in \mathcal{Z}} \sigma_z \omega_z \|p_{T_z} - q_{T_z}\|_2^2 \mathbb{1}_{\{\sigma_z \geq 4\}} = A \,.$$

□

We will require the following analogue of Lemma 3.1 for D:

Lemma 5.2. *For $z \in \mathcal{Z}$, let $\alpha_z := m \cdot p_Z(z)$. Then, we have that:*

$$\mathbb{E}[D] \geq \gamma \cdot \sum_{z \in \mathcal{Z}} \varepsilon_z^2 \min(\alpha_z \beta_z, \alpha_z^4) \tag{26}$$

for some absolute constant $\gamma > 0$, where $\beta_z := \sqrt{\min(\alpha_z, \ell_1) \min(\alpha_z, \ell_2)}$.

Proof. From the definition of D, we obtain that its expectation is:

$$\mathbb{E}[D] = \sum_z \mathbb{E}_\sigma \left[\sigma_z \omega_z \mathbb{1}_{\{\sigma_z \geq 4\}} \varepsilon_z'^2 \right] .$$

Now

$$\mathbb{E}[D] = \sum_z \varepsilon_z'^2 \mathbb{E}_\sigma \left[\sigma_z \omega_z \mathbb{1}_{\{\sigma_z \geq 4\}} \right] = \Omega(1) \sum_z \varepsilon_z'^2 \min(\alpha_z \beta_z, \alpha_z^4) \,,$$

using the fact (Claim 2.3) that, for a Poisson random variable X with parameter λ, $\mathbb{E}\left[X \sqrt{\min(X, a) \min(X, b)} \mathbb{1}_{\{X \geq 4\}} \right] \geq \gamma \min(\lambda \sqrt{\min(\lambda, a) \min(\lambda, b)}, \lambda^4)$, for some absolute constant $\gamma > 0$. □

We will leverage this lemma to show the following lower bound on the expectation of D:

Proposition 5.1. *If $\mathrm{d_{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\right) > \varepsilon$, then $\mathbb{E}[D] = \Omega\left(\zeta \sqrt{\min(n, m)}\right)$ (where ζ the constant in the definition of m).*

Proof. Since $\mathrm{d_{TV}}\left(p, \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\right) > \varepsilon$, we have that $\sum_{z \in \mathcal{Z}} \varepsilon'_z \alpha_z \geq \frac{1}{2\sqrt{\ell_1 \ell_2}} \sum_{z \in \mathcal{Z}} \varepsilon_z \alpha_z > \frac{m\varepsilon}{\sqrt{\ell_1 \ell_2}}$.

We once again divide $\mathcal{Z}$ into heavy and light bins, $\mathcal{Z}_H := \{\, z : \alpha_z^3 \geq \beta_z \,\}$ and $\mathcal{Z}_L := \mathcal{Z} \setminus \mathcal{Z}_H$. By the above, we must have $\sum_{z \in \mathcal{Z}_H} \varepsilon_z \alpha_z > m\varepsilon$ or $\sum_{z \in \mathcal{Z}_L} \varepsilon_z \alpha_z > m\varepsilon$. We proceed as in the proof of Proposition 3.1 to handle these two cases.

- In the first case, we want to lower bound $\sum_{z \in \mathcal{Z}_H} \varepsilon_z'^2 \alpha_z \beta_z$. We consider three sub-cases, partitioning $\mathcal{Z}_H$ in 3: (1) $\mathcal{Z}_{H,1} := \{\, z \in \mathcal{Z} : \ell_2 \leq \ell_1 < \alpha_z \,\}$, (2) $\mathcal{Z}_{H,2} := \{\, z \in \mathcal{Z} : \ell_2 \leq \alpha_z \leq \ell_1 \,\}$, and (3) $\mathcal{Z}_{H,3} := \{\, z \in \mathcal{Z} : \alpha_z < \ell_2 \leq \ell_1 \,\}$. By a similar argument, at least one of these sets is such that $\sum_{z \in \mathcal{Z}_{H,i}} \varepsilon_z \alpha_z > \frac{1}{3} m\varepsilon'$.

– In the first sub-case:

$$\sum_{z\in\mathcal{Z}_{H,1}} \varepsilon_z'^2\alpha_z\beta_z = \sqrt{\ell_1\ell_2}\sum_{z\in\mathcal{Z}_{H,1}} \varepsilon_z'^2\alpha_z \geq \sqrt{\ell_1\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,1}}\varepsilon_z'\alpha_z\right)^2}{\sum_{z\in\mathcal{Z}_{H,1}}\alpha_z} \geq \sqrt{\ell_1\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,1}}\varepsilon_Z'\alpha_z\right)^2}{m}$$

by Cauchy–Schwarz and recalling that $\sum_{z\in\mathcal{Z}_{H,1}}\alpha_z \leq \sum_{z\in\mathcal{Z}}\alpha_z = m$; and again by Jensen's inequality after taking expectations on both sides,

$$\sum_{z\in\mathcal{Z}_{H,1}} \varepsilon_z'^2\alpha_z\beta_z \geq \sqrt{\ell_1\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,1}}\varepsilon_z'\alpha_z\right)^2}{m} > \sqrt{\ell_1\ell_2}\frac{1}{36}m\varepsilon'^2 = \frac{1}{36}\frac{m\varepsilon^2}{\sqrt{\ell_1\ell_2}}\,. \tag{27}$$

– In the second sub-case:

$$\sum_{z\in\mathcal{Z}_{H,2}} \varepsilon_z'^2\alpha_z\beta_z = \sqrt{\ell_2}\sum_{z\in\mathcal{Z}_{H,2}} \varepsilon_z'^2\alpha_z^{3/2} \geq \sqrt{\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,2}}\varepsilon_z'\alpha_z\right)^2}{\sum_{z\in\mathcal{Z}_{H,2}}\sqrt{\alpha_z}} \geq \sqrt{\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,2}}\varepsilon_z'\alpha_z\right)^2}{\min(\sqrt{mn}, m/\sqrt{\ell_1})}$$

by Jensen's inequality and then using that $\sum_{z\in\mathcal{Z}_{H,2}}\sqrt{\alpha_z} = \sqrt{m}\sum_{z\in\mathcal{Z}_{H,2}}\sqrt{p_Z(z)} \leq \sqrt{mn}$, and also that by definition of $\mathcal{Z}_{H,2}$ we have $\sum_{z\in\mathcal{Z}_{H,2}}\sqrt{\alpha_z} \leq \sqrt{m}\sum_{z\in\mathcal{Z}_{H,2}}\sqrt{p_Z(z)} \leq \sqrt{m}\frac{m}{\ell_1}\sqrt{\frac{\ell_1}{m}} = \frac{m}{\sqrt{\ell_1}}$. Again by Jensen's inequality after taking expectations on both sides,

$$\begin{aligned}\sum_{z\in\mathcal{Z}_{H,2}} \varepsilon_z'^2\alpha_z\beta_z &\geq \sqrt{\ell_2}\frac{\left(\sum_{z\in\mathcal{Z}_{H,2}}\varepsilon_z'\alpha_z\right)^2}{\min(\sqrt{mn}, m/\sqrt{\ell_1})} > \sqrt{\ell_2}\frac{1}{36}\frac{m^2\varepsilon'^2}{\min(\sqrt{mn}, m/\sqrt{\ell_1})} \\ &= \frac{1}{36}\frac{m^{3/2}\varepsilon^2}{\ell_1\sqrt{\ell_2}}\max\left(\frac{1}{\sqrt{n}}, \sqrt{\frac{\ell_1}{m}}\right). \end{aligned}\tag{28}$$

However, note that since $\sum_{z\in\mathcal{Z}_{H,2}}\varepsilon_z'\alpha_z \leq \sqrt{2}\sum_{z\in\mathcal{Z}_{H,2}}\alpha_z \leq \sqrt{2}\,|\mathcal{Z}_{H,2}|\,\ell_1 \leq \sqrt{2}n\ell_1$ (as $\alpha_z \leq \ell_1$ for $z \in \mathcal{Z}_{H,2}$), the second sub-case cannot happen if $m\varepsilon' \geq 2\sqrt{2}n\ell_1$.

– In the third sub-case:

$$\sum_{z\in\mathcal{Z}_{H,3}} \varepsilon_z'^2\alpha_z\beta_z = \sum_{z\in\mathcal{Z}_{H,3}} \varepsilon_z'^2\alpha_z^2 \geq \frac{\left(\sum_{z\in\mathcal{Z}_{H,3}}\varepsilon_z'\alpha_z\right)^2}{\sum_{z\in\mathcal{Z}_{H,3}}1} \geq \frac{\left(\sum_{z\in\mathcal{Z}_{H,3}}\varepsilon_z'\alpha_z\right)^2}{\min(n,m)}$$

by Jensen's inequality and recalling that $|\mathcal{Z}_{H,3}| \leq \min(n,m)$; and again by Jensen's inequality after taking expectations on both sides,

$$\sum_{z\in\mathcal{Z}_{H,3}} \varepsilon_z'^2\alpha_z \geq \frac{\left(\sum_{z\in\mathcal{Z}_{H,3}}\varepsilon_z'\alpha_z\right)^2}{\min(n,m)} > \frac{1}{36}\frac{m^2}{\min(n,m)}\varepsilon'^2 = \frac{1}{36}\max\left(\frac{m^2}{n}, m\right)\frac{\varepsilon^2}{\ell_1\ell_2}\,. \tag{29}$$

However, note that since $\sum_{z\in\mathcal{Z}_{H,3}}\delta_z\alpha_z \leq \sqrt{2}\sum_{z\in\mathcal{Z}_{H,3}}\alpha_z \leq \sqrt{2}\,|\mathcal{Z}_{H,3}|\,\ell_2 \leq \sqrt{2}n\ell_2$ (as $\alpha_z < \ell_2$ for $z \in \mathcal{Z}_{H,3}$), the third sub-case cannot happen if $m\varepsilon' \geq 2\sqrt{2}n\ell_2$.

- In the second case, we want to lower bound $\sum_{z\in\mathcal{Z}_L}\varepsilon_z'^2\alpha_z^4$. We then use the same chain of (in)-equalities as in the second case of Proposition 3.1, to obtain

$$\sum_{z\in\mathcal{Z}_L}\varepsilon_z'^2\alpha_z^4 \geq \frac{\left(\sum_{z\in\mathcal{Z}_L}\varepsilon_z'\alpha_z\right)^4}{\left(\sum_{z\in\mathcal{Z}_L}\varepsilon_z'^{2/3}\right)^3}\,,$$

and recall that $\varepsilon_z' = \frac{\varepsilon_z}{2\sqrt{\ell_1\ell_2}} \leq \frac{1}{\sqrt{\ell_1\ell_2}}$ to conclude

$$\sum_{z\in\mathcal{Z}_L}\varepsilon_z'^2\alpha_z^4 \geq \frac{\ell_1\ell_2}{4n^3}\Big(\sum_{z\in\mathcal{Z}_L}\varepsilon_z'\alpha_z\Big)^4 = \frac{\ell_1\ell_2}{4n^3}\Big(\frac{1}{2\sqrt{\ell_1\ell_2}}\sum_{z\in\mathcal{Z}_L}\varepsilon_z\alpha_z\Big)^4 > \frac{1}{8}\frac{m^4\varepsilon^4}{n^3\ell_1\ell_2}\,. \tag{30}$$

However, note that since $\sum_{z\in\mathcal{Z}_L}\delta_z\alpha_z \leq \sqrt{2}\sum_{z\in\mathcal{Z}_L}\alpha_z \leq \sqrt{2}\,|\mathcal{Z}_L| \leq \sqrt{2}n$ (as $\alpha_z \leq 1$ for $z\in\mathcal{Z}_L$), the second case cannot happen if $m\varepsilon' \geq 2\sqrt{2}n$.

It remains to use Eqs. (27) to (30) and our setting of m to show that $\mathbb{E}[D] \geq C\sqrt{\min(n,m)}$ (where the constant $C>0$ depends on the choice of the constant in the definition of m).

- From Eq. (27) and the fact that $m \geq \zeta\min(\ell_1\ell_2/\varepsilon^4, \sqrt{n\ell_1\ell_2}/\varepsilon^2)$, we get

$$\sum_{z\in\mathcal{Z}_{H,1}}\varepsilon_z'^2\alpha_z\beta_z \gg \sqrt{\zeta\min(n,m)}$$

in the first sub-case of the first case.

- From Eq. (28) and the fact that $m \geq \zeta\min(n^{2/3}\ell_1^{2/3}\ell_2^{1/3}/\varepsilon^{4/3}, \ell_1\ell_2/\varepsilon^4, \sqrt{n}\ell_1\sqrt{\ell_2}/\varepsilon^2, n\ell_1^{3/2}\ell_2^{1/2}/\varepsilon)$, we get

$$\sum_{z\in\mathcal{Z}_{H,2}}\varepsilon_z'^2\alpha_z\beta_z \gg \sqrt{\zeta\min(n,m)}$$

in the second sub-case of the first case (depending on whether $\min(n,m)\min\left(n,\frac{m}{\ell_1}\right)$ is equal to n^2, m^2/ℓ_1, or mn). (The last term in the $\min$ enforcing the condition that this sub-case can only happen whenever $m\varepsilon' = O(n\ell_1)$.)

- From Eq. (29) and the fact that $m \geq \zeta\min(n^{3/4}\ell_1^{1/2}\ell_2^{1/2}/\varepsilon, \ell_1^2\ell_2^2/\varepsilon^4, n\ell_1^{1/2}\ell_2^{3/2}/\varepsilon)$, we get

$$\sum_{z\in\mathcal{Z}_{H,3}}\varepsilon_z'^2\alpha_z\beta_z \gg \sqrt{\zeta\min(n,m)}$$

in the third sub-case of the first case (depending on whether $\sqrt{\min(n,m)}\min\left(\frac{1}{m},\frac{n}{m^2}\right)$ is equal to $n^{3/2}/m^2$ or $1/m^{1/2}$). (The last term in the $\min$ enforcing the condition that this sub-case can only happen whenever $m\varepsilon' = O(n\ell_2)$.)

- From Eq. (30) and the fact that $m \geq \zeta\min(n^{7/8}\ell_1^{1/4}\ell_2^{1/4}/\varepsilon, n^{6/7}\ell_1^{2/7}\ell_2^{2/7}/\varepsilon^{8/7}, n\ell_1^{1/2}\ell_2^{1/2}/\varepsilon)$, we get

$$\sum_{z\in\mathcal{Z}_L}\varepsilon_z'^2\alpha_z\beta_z \gg \zeta^2\sqrt{\min(n,m)}$$

in the second case (depending on whether $\min(n,m)$ is equal to n or m). (The last term in the $\min$ enforcing the condition that this sub-case can only happen whenever $m\varepsilon' = O(n)$.)

This completes the proof of Proposition 5.1. □

5.5.2 Variances of D and A

First we bound the variance of D:

Lemma 5.3.

$$\mathrm{Var}[D] \le O(\mathbb{E}[D])\,.$$

Proof. Recall that $D = \sum_{z\in\mathcal{Z}} \sigma_z\omega_z\varepsilon_z'^2\mathbb{1}_{\{\sigma_z\ge 4\}}$. Since the σ_z's are independent and D_z is a function of σ_z, the D_z's are independent as well and so

$$\mathrm{Var}[D] = \sum_{z\in\mathcal{Z}} \mathrm{Var}[\sigma_z\omega_z\varepsilon_z'^2\mathbb{1}_{\{\sigma_z\ge 4\}}] = \sum_{z\in\mathcal{Z}} \varepsilon_z'^4\,\mathrm{Var}_\sigma[\sigma_z\omega_z\mathbb{1}_{\{\sigma_z\ge 4\}}]\,.$$

As σ_z is distributed as $\mathrm{Poisson}(\alpha_z)$, we can use Claim 2.2 to bound this

$$\begin{aligned}\mathrm{Var}[D] &\le C'\sum_{z\in\mathcal{Z}} \varepsilon_z'^4\mathbb{E}_\sigma[\sigma_z\omega_z\mathbb{1}_{\{\sigma_z\ge 4\}}]\\ &\le C'\sum_{z\in\mathcal{Z}} \varepsilon_z'^2\mathbb{E}_\sigma[\sigma_z\omega_z\mathbb{1}_{\{\sigma_z\ge 4\}}]\\ &= C'\mathbb{E}[D]\ ,\end{aligned}$$

for some absolute constant $C' > 0$. □

Since our statistic A is a linear combination of the Φ_{S_z}'s and all the S_z's are independent by Poissonization, we get the analogue of Proposition 3.2:

Proposition 5.2. *Let* $E := \sum_{z\in\mathcal{Z}} \omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\ge 4\}}$. *Then,*

$$\mathrm{Var}[A \mid \sigma, T] \le C\left(E + E^{1/2}\mathbb{E}[\,A \mid \sigma, T\,] + \mathbb{E}[\,A \mid \sigma, T\,]^{3/2}\right)\ , \tag{31}$$

where $\mathbb{E}[\,E \mid \sigma\,] = O(\min(n, M))$ *and* $C > 0$ *is some absolute constant.*

Proof. Since $\mathrm{Var}[A_z \mid \sigma_z, T_z] = \sigma_z^2\omega_z^2\mathbb{1}_{\{\sigma_z\ge 4\}}\,\mathrm{Var}[\Phi(S_z) \mid \sigma_z, T_z]$, we have by Eq. (22) that, for some absolute constant $C > 0$,

$$\mathrm{Var}[A_z \mid \sigma_z, T_z] \le C\left(\sigma_z^2\omega_z^2\left(\frac{\|p_{T_z} - q_{T_z}\|_2^2\sqrt{B_{T_z}}}{\sigma_z} + \frac{B_{T_z}}{\sigma_z^2} + \frac{\|p_{T_z} - q_{T_z}\|_2^3}{\sigma_z}\right)\mathbb{1}_{\{\sigma_z\ge 4\}}\right)\ .$$

We will handle the three terms of the RHS separately. First, by Cauchy–Schwarz and monotonicity of ℓ_p norms we get that

$$\begin{aligned}\sum_{z\in\mathcal{Z}} \sigma_z\omega_z^2\|p_{T_z} - q_{T_z}\|_2^2\sqrt{B_{T_z}}\mathbb{1}_{\{\sigma_z\ge 4\}} &\le \left(\sum_{z\in\mathcal{Z}} \omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\ge 4\}}\right)^{1/2}\left(\sum_{z\in\mathcal{Z}}\left(\sigma_z\omega_z\|p_{T_z} - q_{T_z}\|_2^2\right)^2\mathbb{1}_{\{\sigma_z\ge 4\}}\right)^{1/2}\\ &\le \left(\sum_{z\in\mathcal{Z}} \omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\ge 4\}}\right)^{1/2}\sum_{z\in\mathcal{Z}} \sigma_z\omega_z\|p_{T_z} - q_{T_z}\|_2^2\mathbb{1}_{\{\sigma_z\ge 4\}}\\ &= E^{1/2}\mathbb{E}[\,A \mid \sigma, T\,]\ ,\end{aligned} \tag{32}$$

the last equality from Eq. (25).

Moreover, for the second term $\sigma_z^2\omega_z^2\frac{B_{T_z}}{\sigma_z^2}\mathbb{1}_{\{\sigma_z\geq 4\}} = \omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\geq 4\}}$, it is immediate that summing over all bins we get $\sum_{z\in\mathcal{Z}}\omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\geq 4\}} = E$.

Let us now turn to the last term of our upper bound on the variance. We can write

$$\begin{aligned}\sigma_z^2\omega_z^2\frac{\|p_{T_z}-q_{T_z}\|_2^3}{\sigma_z}\mathbb{1}_{\{\sigma_z\geq 4\}} &= \sigma_z\omega_z^2\|p_{T_z}-q_{T_z}\|_2^3\mathbb{1}_{\{\sigma_z\geq 4\}} = \sqrt{\frac{\omega_z}{\sigma_z}}\sigma_z^{3/2}\omega_z^{3/2}\|p_{T_z}-q_{T_z}\|_2^3\mathbb{1}_{\{\sigma_z\geq 4\}}\\ &\leq \Big(\sigma_z\omega_z\|p_{T_z}-q_{T_z}\|_2^2\mathbb{1}_{\{\sigma_z\geq 4\}}\Big)^{3/2} = \mathbb{E}[\,A_z\mid\sigma,T\,]^{3/2}\end{aligned}$$

recalling that $\omega_z\leq\sigma_z$ by definition. We may use the inequality between the ℓ_1 and $\ell_{3/2}$ norms to conclude that $\sum_{z\in\mathcal{Z}}\mathbb{E}[\,A_z\mid\sigma,T\,]^{3/2}\leq\mathbb{E}[\,A\mid\sigma,T\,]^{3/2}$, which leads by the above to

$$\mathbb{E}\left[\sum_{z\in\mathcal{Z}}\sigma_z\omega_z^2\|p_{T_z}-q_{T_z}\|_2^3\mathbb{1}_{\{\sigma_z\geq 4\}}\,\middle|\,\sigma,T\right]\,.\leq\mathbb{E}[\,A\mid\sigma,T\,]^{3/2} \tag{33}$$

Since the A_z's are independent conditioned on σ_z and T_z, we have

$$\mathrm{Var}[A\mid\sigma,T] = \sum_{z\in\mathcal{Z}}\mathrm{Var}[A\mid\sigma_z,T_z]\,,$$

and therefore by Eqs. (32) and (33) and the definition of E we obtain

$$\mathrm{Var}[A\mid\sigma,T]\leq O\Big(E^{1/2}\mathbb{E}[\,A\mid\sigma,T\,]+E+\mathbb{E}[\,A\mid\sigma,T\,]^{3/2}\Big)\,. \tag{34}$$

It remains to establish the further guarantee that $\mathbb{E}[\,E\mid\sigma\,]=O(\min(n,M))$. To do so, observe that we can write, as ω_z only depends on the randomness σ,

$$\mathbb{E}\Big[\,\omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\geq 4\}}\,\Big|\,\sigma\,\Big] = \omega_z^2\mathbb{E}[\,B_{T_z}\mid\sigma\,]\,\mathbb{1}_{\{\sigma_z\geq 4\}}\leq\frac{\omega_z^2}{(1+t_{1,z})(1+t_{2,z})}\mathbb{1}_{\{\sigma_z\geq 4\}}$$

by Eq. (21). Recalling that $t_{i,z}=\min((\sigma_z-4)/4,\ell_i)$ by the definition of the flattening (Section 5.1) and that $\omega_z^2=\min(\sigma_z,\ell_1)\min(\sigma_z,\ell_2)$, this leads to

$$\mathbb{E}\Big[\,\omega_z^2 B_{T_z}\mathbb{1}_{\{\sigma_z\geq 4\}}\,\Big|\,\sigma\,\Big]\leq O(1)\cdot\mathbb{1}_{\{\sigma_z\geq 4\}}\,.$$

In particular, by summing over all bins z this implies that

$$\mathbb{E}[\,E\mid\sigma\,]\leq O(1)\sum_{z\in\mathcal{Z}}\mathbb{1}_{\{\sigma_z\geq 4\}}\leq O(1)\sum_{z\in\mathcal{Z}}\mathbb{1}_{\{\sigma_z\geq 1\}} = O(\min(n,M))\,, \tag{35}$$

as claimed.

Lemma 5.4 (Soundness)**.** *If* $\mathrm{d_{TV}}\Big(p,\mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}\Big)>\varepsilon$*, then with probability at least* $99/100$ *over* σ,T *we have simultaneously* $\mathbb{E}[\,A\mid\sigma,T\,]=\Omega\Big(\sqrt{\zeta\min(n,m)}\Big)$ *and*

$$\mathrm{Var}[A\mid\sigma,T]\leq O\Big(\min(n,m)+\sqrt{\min(n,m)}\mathbb{E}[\,A\mid\sigma,T\,]+\mathbb{E}[\,A\mid\sigma,T\,]^{3/2}\Big).$$

Proof. By Lemma 5.1, we have that D is a lower bound on $\mathbb{E}[\,A \mid \sigma, T\,]$ for all σ, T. Since Proposition 5.1 and Lemma 5.3 further ensures that $\mathbb{E}[D] \geq \Omega(\sqrt{\zeta \min(n,m)})$ and $\mathrm{Var}[D] \leq O(\mathbb{E}[D])$, applying Chebyshev's inequality on D results in

$$\begin{aligned}\Pr_{\sigma,T}\left[\,\mathbb{E}[\,A \mid \sigma, T\,] < \kappa\sqrt{\zeta \min(n,m)}\,\right] &\leq \Pr_{\sigma,T}[\,D < O(\mathbb{E}[D])\,] = O\left(\frac{\mathrm{Var}[D]}{\mathbb{E}[D]^2}\right) \\ &= O\left(\frac{1}{\mathbb{E}[D]}\right) = O\left(\frac{1}{\sqrt{\zeta \min(n,m)}}\right) \leq \frac{1}{200}\end{aligned}$$

for some absolute constant $\kappa > 0$. This gives the first statement. For the second, we start from Eq. (31) and we apply Markov's inequality to E: as $\mathbb{E}[\,E \mid \sigma\,] = O(\min(n, M))$, with probability at least $399/400$ we have $E \leq 400\mathbb{E}[\,E \mid \sigma\,] = O(\min(n, M))$. Moreover, recalling that $M = \sum_{z \in \mathcal{Z}} \sigma_z$ is a Poisson random variable with parameter m, we have $\Pr[\,M > 2m\,] \leq 399/400$. Therefore, by a union bound

$$\Pr_{\sigma,T}\left[\,\mathrm{Var}[A \mid \sigma, T] \geq \kappa'\left(\min(n,m) + \sqrt{\min(n,m)}\mathbb{E}[\,A \mid \sigma, T\,] + \mathbb{E}[\,A \mid \sigma, T\,]^{3/2}\right)\right] \leq \frac{1}{400} + \frac{1}{400} = \frac{1}{200}$$

again for some absolute constant $\kappa' > 0$. This gives the second statement. A union bound over both events concludes the proof. □

Lemma 5.5 (Completeness)**.** *If $p \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$, then with probability at least $99/100$ over σ, T we have simultaneously $\mathbb{E}[\,A \mid \sigma, T\,] = 0$ and $\mathrm{Var}[A \mid \sigma, T] \leq O(\min(n,m))$.*

Proof. The first statement is obvious by the definition of A as sum of the A_z's, since $\varepsilon_z = 0$ for all $z \in \mathcal{Z}$. For the second, the proof is identical as that of Lemma 5.4, but having only to deal with the term E in the bound on the variance (as the others are zero). □

□

5.6 Completing the Proof

Let our threshold τ be set to $\zeta^{1/4}\sqrt{\min(n,m)}$. Gathering the above pieces we obtain the following:

Lemma 5.6 (Soundness)**.** *If p is ε-far from conditionally independent, then $\Pr[\,A < \tau\,] \leq \frac{1}{50}$.*

Proof. We apply Chebyshev's inequality once more, this time to $A' := (A \mid \sigma, T)$ and relying on the bounds on its expectation and variance established in Lemma 5.4. Specifically, let $\mathcal{E}$ denote the event that both bounds of Lemma 5.4 hold simultaneously; then

$$\begin{aligned}\Pr[\,A \leq \tau\,] = \Pr[\,A' \leq \tau\,] &\leq \Pr[\,A' \leq \tau \mid \mathcal{E}\,] + \Pr[\overline{\mathcal{E}}] \\ &\leq \Pr\left[\,|A' - \mathbb{E}[\,A \mid \sigma, T\,]| \geq \frac{1}{2}\mathbb{E}[\,A \mid \sigma, T\,] \;\middle|\; \mathcal{E}\,\right] + \frac{1}{100}\end{aligned}$$

where the second line is because, conditioned on $\mathcal{E}$, $\mathbb{E}[\,A \mid \sigma, T\,] \geq \Omega\left(\sqrt{\zeta \min(n,m)}\right) \geq \zeta^{1/4}\sqrt{\min(n,m)} = 2\tau$.

It only remains to bound the first term:

$$\begin{aligned}\Pr[\,|A' - \mathbb{E}[\,A \mid \sigma, T\,]| \geq \mathbb{E}[\,A \mid \sigma, T\,] \mid \mathcal{E}\,] &\leq O\left(\frac{\min(n,m) + \sqrt{\min(n,m)}\mathbb{E}[\,A \mid \sigma, T\,] + \mathbb{E}[\,A \mid \sigma, T\,]^{3/2}}{\mathbb{E}[\,A \mid \sigma, T\,]^2}\right) \\ &= O\left(\frac{\min(n,m)}{\mathbb{E}[\,A \mid \sigma, T\,]^2} + \frac{\sqrt{\min(n,m)}}{\mathbb{E}[\,A \mid \sigma, T\,]} + \frac{1}{\mathbb{E}[\,A \mid \sigma, T\,]^{1/2}}\right) \\ &\leq O\left(\frac{1}{\zeta^{1/4}}\right) \leq \frac{1}{100}\end{aligned}$$

for the choice of a sufficiently large constant ζ in the definition of m. □

Lemma 5.7 (Completeness)**.** *If p is conditionally independent, then* $\Pr[\,A \geq \tau\,] \leq \frac{1}{50}$.

Proof. Analogously to the proof in the soundness case, we apply Chebyshev's inequality to $A' := (A \mid \sigma, T)$ and relying on Lemma 5.5. Specifically, let $\mathcal{E}]$ denote the event that the bound of Lemma 5.5 holds; then

$$\Pr[\,A \geq \tau\,] = \Pr[\,A' \geq \tau\,] \leq \Pr[\,A' \geq \tau \mid \mathcal{E}]\,] + \Pr[\overline{\mathcal{E}'}]\,.$$

To conclude, we bound the first term:

$$\begin{aligned}\Pr[\,A' \geq \tau \mid \mathcal{E}']\,] \leq \frac{\mathrm{Var}[A' \mid \mathcal{E}']}{\tau^2} &\leq O\left(\frac{\min(n,m)}{\tau^2}\right) \\ &\leq O\left(\frac{1}{\zeta}\right) \leq \frac{1}{100}\end{aligned}$$

again for the choice of a sufficiently large constant ζ in the definition of m. □

6 Sample Complexity Lower Bounds: The Case of Constant $|\mathcal{X}|, |\mathcal{Y}|$

In this section, we prove our tight sample complexity lower bound of

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

for testing conditional independence in the regime that $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $\mathcal{Z} = [n]$. This matches the sample complexity of our algorithm in Section 3, up to constant factors. In the main body of this section, we prove each lower bound separately.

The following expression for the total variation distance will be useful in the analysis of the lower bound constructions:

Fact 6.1. *For any $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ for $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $\mathcal{Z} = [n]$. we have that:*

$$\mathrm{d}_{\mathrm{TV}}(p_z, q_z) = 2\,|\mathrm{Cov}[(X \mid Z = z), (Y \mid Z = z)]| = \|p_z - q_z\|_2\,. \tag{36}$$

Proof. We have the following:

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

□

6.1 First Lower Bound Regime: $\Omega\left(n^{6/7}/\varepsilon^{8/7}\right)$ for $\varepsilon > \frac{1}{n^{1/8}}$

Assume that we are in the regime where

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

i.e., $\varepsilon > 1/n^{1/8}$. Suppose there is an algorithm for ε-testing conditional independence drawing $m \leq cn^{6/7}/\varepsilon^{8/7}$ samples from p, for some sufficiently small universal constant $c > 0$. Note that in this regime, we have $m \ll n$, i.e., we can assume that $m < c'n$ for some small constant $c' > 0$.

The yes-instance A pseudo-distribution p is drawn from the yes-instances as follows: Independently for each value $z \in [n]$, we set:

- With probability $\frac{m}{n}$, we select $p_Z(z) = \frac{1}{m}$ and $p_z(i,j) = \frac{1}{4}$ for all $i, j \in \{0,1\}$. In other words, we select uniform marginals for the conditional distributions $p_z(i,j)$.
- With probability $1 - \frac{m}{n}$, we select $p_Z(z) = \frac{\varepsilon}{n}$ and $p_z(i,j)$ us defined by the 2×2 matrix:
  - With probability $1/2$, $\frac{1}{100}\begin{pmatrix} 16 & 24 \\ 24 & 36 \end{pmatrix} := Y_1$,
  - With probability $1/2$, $\frac{1}{100}\begin{pmatrix} 36 & 24 \\ 24 & 16 \end{pmatrix} := Y_2$

It is easy to see that the resulting distribution p satisfies

$$\mathbb{E}\left[\sum_{z=1}^{n} p_Z(z)\right] = n\left(\frac{m}{n}\cdot\frac{1}{m} + \left(1 - \frac{m}{n}\right)\cdot\frac{\varepsilon}{n}\right) \in [1, 1+\varepsilon] \ ,$$

i.e., the marginal for Z has mass roughly 1 in expectation, and that $p \in \mathcal{P}_{\{0,1\},\{0,1\}|[n]}$.

The no-instance A pseudo-distribution p is drawn from the no-instances as follows: Independently for each value $z \in [n]$, we set

- With probability $\frac{m}{n}$, we set $p_Z(z) = \frac{1}{m}$ and $p_z(i,j) = \frac{1}{4}$ for all $i, j \in \{0,1\}$.
- With probability $1 - \frac{m}{n}$, we set $p_Z(z) = \frac{\varepsilon}{n}$ and $p_z(i,j)$ be defined by the 2×2 matrix:
  - With probability $1/8$, $\frac{1}{100}\begin{pmatrix} 6 & 24 \\ 24 & 46 \end{pmatrix} := N_1$,
  - With probability $1/8$, $\frac{1}{100}\begin{pmatrix} 46 & 24 \\ 24 & 6 \end{pmatrix} := N_2$,
  - With probability $3/4$, $\frac{1}{100}\begin{pmatrix} 26 & 24 \\ 24 & 26 \end{pmatrix} := N_3$.

Similarly, we have that $\mathbb{E}[\sum_{z=1}^{n} p_Z(z)] \in [1, 1+\varepsilon]$. Furthermore, the expected total variation distance between such a p and the corresponding $q := \sum_{z=1}^{n} p_Z(z) q_z \in \mathcal{P}_{\{0,1\},\{0,1\}|[n]}$ is

$$\mathbb{E}[\mathrm{d_{TV}}(p,q)] = \frac{1}{2}\left(n\left(\frac{m}{n}\cdot\frac{1}{m}\cdot 0 + \left(1-\frac{m}{n}\right)\frac{\varepsilon}{n}\left(\frac{1}{8}\cdot\frac{12}{100} + \frac{1}{8}\cdot\frac{12}{100} + \frac{3}{4}\cdot\frac{4}{100}\right)\right)\right) = \frac{3}{100}\left(1-\frac{m}{n}\right)\varepsilon > \frac{\varepsilon}{100} \ ,$$

where

$$\frac{12}{100} = \left|\frac{1}{100}\begin{pmatrix} 46 & 24 \\ 24 & 6 \end{pmatrix} - \frac{1}{100}\begin{pmatrix} 7 \\ 3 \end{pmatrix}\begin{pmatrix} 7 & 3 \end{pmatrix}\right|, \qquad \frac{4}{100} = \left|\frac{1}{100}\begin{pmatrix} 26 & 24 \\ 24 & 26 \end{pmatrix} - \frac{1}{100}\begin{pmatrix} 5 \\ 5 \end{pmatrix}\begin{pmatrix} 5 & 5 \end{pmatrix}\right|.$$

Thus, $\mathbb{E}[\mathrm{d}_{\mathrm{TV}}(p,q)] = \Omega(\varepsilon)$, which by Lemma 2.2 implies that $\mathbb{E}\Big[\mathrm{d}_{\mathrm{TV}}\Big(p, \mathcal{P}_{\{0,1\},\{0,1\}|[n]}\Big)\Big] = \Omega(\varepsilon)$.

The next claim shows that, for each $z \in [n]$, the first three norms of the conditional distribution $p_z(i,j)$ match, hence do not provide any information towards distinguishing between the yes- and no-cases. Therefore, we need to get at least 4 samples (X,Y,Z) with the same value of Z — that is a 4-collision with regard to Z — in order to have useful information.

Notation. Given a 4-variable function $R = R[X_1, X_2, X_3, X_4]$ and a real 2×2 matrix $M \in \mathcal{M}_2(\mathbb{R})$, we will denote $R(M) := R(M_{1,1}, M_{1,2}, M_{2,1}, M_{2,2})$.

We have the following:

Claim 6.1. *Let Y_1, Y_2, N_1, N_2, N_3 the probability matrices in the definition of the* yes *and* no*-instances. For every 4-variable polynomial $R \in \mathbb{R}[X_1, X_2, X_3, X_4]$ of degree at most 3, the following holds:*

$$\frac{1}{8}R(N_1) + \frac{1}{8}R(N_2) + \frac{3}{4}R(N_3) = \frac{1}{2}R(Y_1) + \frac{1}{2}R(Y_2)\ .$$

Proof. The first crucial observation is that the associated matrices can be expressed in the form

$$(N_1, Y_1, N_3, Y_2, N_2) = (A + kB)_{0 \le k \le 4}\ ,$$

where A, B are the following matrices:

$$A = \frac{1}{100}\begin{pmatrix} 6 & 24 \\ 24 & 46 \end{pmatrix}, \qquad B = \frac{1}{100}\begin{pmatrix} 10 & 0 \\ 0 & -10 \end{pmatrix}\ .$$

Therefore, for any function 4-variable function R (not necessarily a polynomial), we have

$$\begin{aligned}\left(\frac{1}{8}R(N_1) + \frac{1}{8}R(N_2) + \frac{3}{4}R(N_3)\right) - \left(\frac{1}{2}R(Y_1) + \frac{1}{2}R(Y_2)\right) \\ = \frac{1}{8}R(N_1) - \frac{1}{2}R(Y_1) + \frac{3}{4}R(N_3) - \frac{1}{2}R(Y_2) + \frac{1}{8}R(N_2) \\ = \frac{1}{8}\sum_{k=0}^{4}(-1)^k\binom{4}{k}R(A+kB) = \frac{1}{8}\sum_{k=0}^{4}(-1)^{4-k}\binom{4}{k}R(A+kB)\ ,\end{aligned}$$

which is the 4^{th}-order forward difference of R at A (more precisely, the fourth finite difference of $f(k) = R(A+kB)$). Using the fact that the $(d+1)^{\text{th}}$-order forward difference of a degree-d polynomial is zero, we get that the above RHS is zero for every degree-3 polynomial R. □

For the sake of simplicity and without loss of generality, we can use the Poissonization trick for the analysis of our lower bound construction (cf. [DK16, CDKS17]). Specifically, instead of drawing m independent samples from p, we assume that our algorithm is provided with m_z samples from the conditional distribution p_z (i.e., conditioned on $Z = z$), where the (m_z)'s are independent Poisson random variables with $m_z \sim \mathrm{Poisson}(mp_Z(z))$.

Consider the following process: we let $U \sim \mathrm{Bern}\left(\frac{1}{2}\right)$ be a uniformly random bit, and choose p to be selected as follows: (i) If $U = 0$, then p is drawn from the yes-instances, (ii) If $U = 1$, then p is drawn from the no-instances. For every $z \in [n]$, let $a_z = (a_z^{00}, a_z^{01}, a_z^{10}, a_z^{11})$ be the 4-tuple of counts of $(i,j)_{i,j\in\{0,1\}}$ among the m_z samples $(X,Y) \sim p_z$. Accordingly, we will denote $A = (a_z)_{z\in[n]}$.

Following the mutual information method used in [DK16], to show the desired sample complexity lower bound of $\Omega\left(n^{6/7}/\varepsilon^{8/7}\right)$, it suffices to show that $I\left(U;A\right) = o(1)$, unless $m = \Omega\left(n^{6/7}/\varepsilon^{8/7}\right)$. Since the $(a_z)_{z\in[n]}$'s are independent conditioned on U, we have that $I\left(U;A\right) \leq \sum_{z=1}^{n} I\left(U;a_z\right)$, and therefore it suffices to bound from above separately $I\left(U;a_z\right)$ for every z. We proceed to establish such a bound in the following lemma:

Lemma 6.1. *For any $z \in [n]$, we have $I\left(U;a_z\right) = O\left(\frac{\varepsilon^8 m^7}{n^7}\right)$.*

Before proving the lemma, we show that it implies the desired lower bound. Indeed, assuming Lemma 6.1, we get that

$$I\left(U;A\right) \leq \sum_{z=1}^{n} I\left(U;a_z\right) = \sum_{z=1}^{n} O\left(\frac{\varepsilon^8 m^7}{n^7}\right) = O\left(\frac{\varepsilon^8 m^7}{n^6}\right) ,$$

which is $o(1)$ unless $m = \Omega\left(n^{6/7}/\varepsilon^{8/7}\right)$. It remains to prove the lemma.

Proof of Lemma 6.1. By symmetry, it is sufficient to show the claim for $z = 1$. To simplify the notation, let $a := a_1$. We first bound $I\left(U;a\right)$ from above using [CDKS17, Fact 4.12] (see also [DK16]) as follows:

$$I\left(U;a\right) \leq \sum_{\alpha\in\mathbb{N}^4} \Pr[\, a = \alpha \,] \left(1 - \frac{\Pr[\, a = \alpha \mid U = 1\,]}{\Pr[\, a = \alpha \mid U = 0\,]}\right)^2 := \Phi(n,m,\varepsilon) \, .$$

Our next step is to get a hold on the conditional probabilities $\Pr[\, a = \alpha \mid U = 0\,]$ and $\Pr[\, a = \alpha \mid U = 1\,]$. For notational convenience, we set

$$p_1 := \frac{16}{100}, \quad p_2 := \frac{24}{100}, \quad p_3 := \frac{36}{100}, \quad q_1 := \frac{6}{100}, q_2 := \frac{26}{100}, \quad q_3 := \frac{46}{100} \, ,$$

and let Ξ denote the event that the bin is "heavy", i.e., that p_Z puts probability mass $\frac{1}{m}$ on it. Note that by construction this event happens with probability $\frac{m}{n}$ and that Ξ is independent of U.

We start with the yes-case. Recall that with probability m/n, Ξ holds: the probability $p_Z(1)$ equals $1/m$, in which case we draw $\mathrm{Poisson}\left(m \cdot \frac{1}{m}\right)$ samples from $Z = 1$ and each sample is uniformly random on $\{0,1\} \times \{0,1\}$. That is, each of the four outcomes is an independent $\mathrm{Poisson}\left(m \cdot \frac{1}{m} \cdot \frac{1}{4}\right)$ random variable. With probability $1 - m/n$, $\bar{\Xi}$ holds: $p_Z(1)$ equals ε/n and we draw $\mathrm{Poisson}(m \cdot \frac{\varepsilon}{n})$ samples from $Z = 1$. With probability $1/2$, all samples follow the first case, and with probability $1/2$ all samples follow the second case.

For any $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^4$, we can explicitly calculate the associated probabilities. Specifically, we can write:

$$\Pr\left[\, a = \alpha \;\middle|\; U = 0, \bar{\Xi}\,\right] = \frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!} \left(\frac{1}{2} p_1^{\alpha_1} p_2^{\alpha_2} p_2^{\alpha_3} p_3^{\alpha_4} + \frac{1}{2} p_3^{\alpha_1} p_2^{\alpha_2} p_2^{\alpha_3} p_1^{\alpha_4}\right) \, , \tag{37}$$

and

$$\begin{aligned}
\Pr[\, a = \alpha \mid U = 0\,] &= \Pr[\, a = \alpha \mid U = 0, \Xi\,] \cdot \Pr[\,\Xi\,] + \Pr\left[\, a = \alpha \;\middle|\; U = 0, \bar{\Xi}\,\right] \cdot \Pr\left[\,\bar{\Xi}\,\right] \\
&= \frac{m}{n} \cdot e^{-\cdot\frac{1}{4}\cdot 4} \frac{4^{-\sum_{\ell=1}^{4}\alpha_\ell}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!} + \left(1 - \frac{m}{n}\right) \cdot \Big(\\
&\quad \frac{1}{2} \frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!} p_1^{\alpha_1} p_2^{\alpha_2} p_2^{\alpha_3} p_3^{\alpha_4} + \\
&\quad \frac{1}{2} \frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!} p_3^{\alpha_1} p_2^{\alpha_2} p_2^{\alpha_3} p_1^{\alpha_4}\Big) \, .
\end{aligned}$$

Similarly, for the no-case, we have

$$\Pr\Big[\, a=\alpha \;\Big|\; U=1,\bar{\Xi}\,\Big] = \frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!}\left(\frac{1}{8}q_1^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_3^{\alpha_4}+\frac{1}{8}q_3^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_1^{\alpha_4}+\frac{3}{4}q_2^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_2^{\alpha_4}\right)\,, \tag{38}$$

$$\Pr[\, a=\alpha \mid U=0,\Xi\,] = \Pr[\, a=\alpha \mid U=1,\Xi\,]\,,$$

and

$$\begin{aligned}\Pr[\, a=\alpha \mid U=1\,] &= \Pr[\, a=\alpha \mid U=1,\Xi\,]\cdot\Pr[\,\Xi\,] + \Pr\Big[\, a=\alpha \;\Big|\; U=1,\bar{\Xi}\,\Big]\cdot\Pr\Big[\,\bar{\Xi}\,\Big]\\ &= \frac{m}{n}\cdot e^{-\frac{1}{4}\cdot 4}\frac{4^{-\sum_{\ell=1}^{4}\alpha_\ell}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!} + \left(1-\frac{m}{n}\right)\cdot\\ &\quad\left(\frac{1}{8}\frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!}q_1^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_3^{\alpha_4}+\frac{1}{8}\frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!}q_3^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_1^{\alpha_4}+\frac{3}{4}\frac{e^{-\frac{\varepsilon m}{n}}}{\alpha_1!\alpha_2!\alpha_3!\alpha_4!}q_2^{\alpha_1}p_2^{\alpha_2}p_2^{\alpha_3}q_2^{\alpha_4}\right).\end{aligned}$$

With these formulas in hand, we can write

$$\begin{aligned}\Phi(n,m,\varepsilon) &:= \sum_{\alpha\in\mathbb{N}^4}\Pr[\, a=\alpha\,]\left(\frac{\Pr[\, a=\alpha \mid U=0\,]-\Pr[\, a=\alpha \mid U=1\,]}{\Pr[\, a=\alpha \mid U=0\,]}\right)^2\\ &= \left(1-\frac{m}{n}\right)^2\sum_{\alpha\in\mathbb{N}^4}\Pr[\, a=\alpha\,]\left(\frac{\Pr\Big[\, a=\alpha \;\Big|\; U=0,\bar{\Xi}\,\Big]-\Pr\Big[\, a=\alpha \;\Big|\; U=1,\bar{\Xi}\,\Big]}{\Pr[\, a=\alpha \mid U=0\,]}\right)^2\\ &\le \sum_{\alpha\in\mathbb{N}^4}\Pr[\, a=\alpha\,]\left(\frac{\Pr\Big[\, a=\alpha \;\Big|\; U=0,\bar{\Xi}\,\Big]-\Pr\Big[\, a=\alpha \;\Big|\; U=1,\bar{\Xi}\,\Big]}{\Pr[\, a=\alpha \mid U=0\,]}\right)^2.\end{aligned}$$

By Eqs. (37) and (38) and Claim 6.1, we observe that the difference $\Pr\Big[\, a=\alpha \;\Big|\; U=0,\bar{\Xi}\,\Big]-\Pr\Big[\, a=\alpha \;\Big|\; U=1,\bar{\Xi}\,\Big]$

is zero for any $|\alpha| := \sum_i \alpha_i \leq 3$. We thus obtain

$$\begin{aligned}
\Phi(n,m,\varepsilon) &\leq \sum_{\substack{\alpha\in\mathbb{N}^4\\|\alpha|\geq 4}} \Pr[\,a=\alpha\,]\left(\frac{\Pr\Big[\,a=\alpha\;\Big|\;U=0,\bar{\Xi}\,\Big]-\Pr\Big[\,a=\alpha\;\Big|\;U=1,\bar{\Xi}\,\Big]}{\Pr[\,a=\alpha\mid U=0\,]}\right)^2\\
&= \sum_{k=4}^{\infty}\sum_{\substack{\alpha\in\mathbb{N}^4\\|\alpha|=k}} \Pr[\,a=\alpha\,]\cdot\Pr\Big[\,|a|=k\;\Big|\;\bar{\Xi}\,\Big]^2\cdot \qquad (\dagger)\\
&\qquad\left(\frac{\Pr\Big[\,a=\alpha\;\Big|\;U=0,\bar{\Xi},|a|=k\,\Big]-\Pr\Big[\,a=\alpha\;\Big|\;U=1,\bar{\Xi},|a|=k\,\Big]}{\Pr[\,a=\alpha\mid U=0\,]}\right)^2\\
&= \sum_{k=4}^{\infty}\frac{e^{-\frac{2\varepsilon m}{n}}}{k!^2}\left(\frac{\varepsilon m}{n}\right)^{2k}\sum_{\substack{\alpha\in\mathbb{N}^4\\|\alpha|=k}}\Pr[\,a=\alpha\,]\cdot\\
&\qquad\left(\frac{\Pr\Big[\,a=\alpha\;\Big|\;U=0,\bar{\Xi},|a|=k\,\Big]-\Pr\Big[\,a=\alpha\;\Big|\;U=1,\bar{\Xi},|a|=k\,\Big]}{\Pr[\,a=\alpha\mid U=0\,]}\right)^2\\
&\leq \sum_{k=4}^{\infty}\frac{e^{-\frac{2\varepsilon m}{n}}}{k!^2}\left(\frac{\varepsilon m}{n}\right)^{2k}\sum_{\substack{\alpha\in\mathbb{N}^4\\|\alpha|=k}}\Pr[\,a=\alpha\,]\cdot\left(\frac{2}{\Pr[\,a=\alpha\mid U=0\,]}\right)^2,
\end{aligned}$$

where for $(\dagger)$ we used the fact that, $|a|$ is independent of U to write

$$\Pr\Big[\,|a|=k\;\Big|\;U=1,\bar{\Xi}\,\Big]=\Pr\Big[\,|a|=k\;\Big|\;U=0,\bar{\Xi}\,\Big]=\Pr\Big[\,|a|=k\;\Big|\;\bar{\Xi}\,\Big]=\frac{e^{-\frac{\varepsilon m}{n}}}{k!}\left(\frac{\varepsilon m}{n}\right)^k.$$

To conclude the proof, we will handle the denominator using the bound

$$\Pr[\,a=\alpha\mid U=0\,]\geq\Pr[\,a=\alpha\mid\Xi,U=0\,]\cdot\Pr[\,\Xi\mid U=0\,]=\frac{m}{n}e^{-1}\frac{4^{-k}}{k!},$$

and rewrite $\Pr[\,a=\alpha\,]=\Pr[\,a=\alpha\mid|a|=|\alpha|\,]\cdot\Pr[\,|a|=|\alpha|\,]$. Using $x:=\frac{\varepsilon m}{n}$ in the following expressions for conciseness, we now get:

$$\begin{aligned}
\Phi(n,m,\varepsilon) &\leq 4e^2\frac{n^2}{m^2}e^{-2x}\sum_{k=4}^{\infty}(4x)^{2k}\Pr[\,|a|=k\,]\sum_{\substack{\alpha\in\mathbb{N}^4\\|\alpha|=k}}\Pr[\,a=\alpha\mid|a|=k\,]=4e^2\frac{n^2}{m^2}e^{-2x}\sum_{k=4}^{\infty}(4x)^{2k}\Pr[\,|a|=k\,]\\
&= 4e^2\frac{n^2}{m^2}e^{-2x}\sum_{k=4}^{\infty}(4x)^{2k}\left(\frac{m}{n}\frac{e^{-1}}{k!}+\left(1-\frac{m}{n}\right)e^{-x}\frac{x^k}{k!}\right)\\
&\leq 40\frac{n^2}{m^2}\sum_{k=4}^{\infty}(4x)^{2k}\left(\frac{m}{n}\frac{1}{k!}+\frac{x^k}{k!}\right)=40\frac{n}{m}\sum_{k=4}^{\infty}\frac{1}{k!}(4x)^{2k}+40\frac{n^2}{m^2}\sum_{k=4}^{\infty}\frac{1}{k!}\left(4^{2/3}x\right)^{3k}\\
&= 40\frac{n}{m}\frac{(4x)^8}{24}+o\left(\frac{n}{m}x^8\right)+40\frac{n^2}{m^2}\frac{1}{24}\left(4^{2/3}x\right)^{12}+o\left(\frac{n^2}{m^2}x^{12}\right) \qquad (\ddagger)\\
&= 2^{16}\frac{5}{3}\cdot\frac{\varepsilon^8m^7}{n^7}+o\left(\frac{\varepsilon^8m^7}{n^7}\right),
\end{aligned}$$

where for (‡) we relied on the Taylor series expansion of exp, recalling that $\frac{\varepsilon m}{n} \ll 1$. This completes the proof of Lemma 6.1. □

6.2 Second Lower Bound Regime: $\Omega\left(n^{7/8}/\varepsilon\right)$ for $\frac{1}{n^{3/8}} \leq \varepsilon \leq \frac{1}{n^{1/8}}$

Assume we are in the regime where

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

i.e., $\frac{1}{n^{3/8}} \leq \varepsilon \leq \frac{1}{n^{1/8}}$. Suppose there is a testing algorithm for conditional independence using $m \leq cn^{7/8}/\varepsilon$ samples, for a sufficiently small universal constant $c > 0$. In this regime, we also have $m \ll n$.

Our construction of yes- and no- instances in this case is similar to those of the previous lower bound, although some specifics about how p_Z is generated will change.

The yes-instance. A pseudo-distribution p is drawn from the yes-instances as follows: Independently for each value $1 \leq z \leq n-1$, we set:

- With probability $\frac{1}{2}$, $p_Z(z) = \frac{1}{m}$ and $p_z(i,j) = \frac{1}{4}$ for all $i,j \in \{0,1\}$. That is, we select uniform marginals for $p_z(i,j)$.
- With probability $\frac{1}{2}$, $p_Z(z) = \frac{\varepsilon}{n}$ and $p_z(i,j)$ is defined by the same 2×2 matrices as in the previous case:
  - With probability $1/2$, Y_1,
  - With probability $1/2$, Y_2.

Furthermore, we set $p_Z(n) = 1$, and $p_n(i,j) = \frac{1}{4}$ for all $i,j \in \{0,1\}$. The last condition ensures that $\|p_Z\|_1 = \Theta(1)$. It is clear that the resulting pseudo-distribution p satisfies $\|p_Z\|_1 = \Theta(1)$ and that $p \in \mathcal{P}_{\{0,1\},\{0,1\}|[n]}$.

The no-instance. A pseudo-distribution p is drawn from the no-instances as follows: Independently for each value $z \in [n]$, we set:

- With probability $\frac{m}{n}$, $p_Z(z) = \frac{1}{m}$ and $p_z(i,j) = \frac{1}{4}$ for all $i,j \in \{0,1\}$, as before.
- With probability $1 - \frac{m}{n}$, $p_Z(z) = \frac{\varepsilon}{n}$, and $p_z(i,j)$ is defined by the same 2×2 matrix as before:
  - With probability $1/8$, N_1,
  - With probability $1/8$, N_2,
  - With probability $3/4$, N_3.

Furthermore, we set as before $p_Z(n) = 1$, and $p_n(i,j) = \frac{1}{4}$ for all $i,j \in \{0,1\}$. This construction ensures that $\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(p, \mathcal{P}_{\{0,1\},\{0,1\}|[n]}\right)\right] = \Omega(\varepsilon)$.

The bulk of the proof of the $\Omega\left(n^{6/7}/\varepsilon^{8/7}\right)$ remains the same. In particular, setting $a := a_1$, we can bound from above as before the mutual information $I(U;a)$ by

$$I(U;a) \leq \sum_{k=4}^{\infty} \frac{e^{-\frac{2\varepsilon m}{n}}}{k!^2} \left(\frac{\varepsilon m}{n}\right)^{2k} \sum_{\substack{\alpha \in \mathbb{N}^4 \\ |\alpha| = k}} \Pr[\, a = \alpha \,] \cdot \left(\frac{2}{\Pr[\, a = \alpha \mid U = 0 \,]}\right)^2 .$$

In the current setting, we use the fact that

$$\Pr[\,a=\alpha\mid U=0\,]\geq\Pr[\,a=\alpha\mid\Xi,U=0\,]\cdot\Pr[\,\Xi\mid U=0\,]=\frac{1}{2}e^{-1}\frac{4^{-k}}{k!}\ ,$$

to obtain that

$$I\left(U;a\right)\leq 16e^{2}e^{-\frac{2\varepsilon m}{n}}\sum_{k=4}^{\infty}\left(\frac{4\varepsilon m}{n}\right)^{2k}\sum_{\substack{\alpha\in\mathbb{N}^{4}\\|\alpha|=k}}\Pr[\,a=\alpha\,]=16e^{2}e^{-\frac{2\varepsilon m}{n}}\sum_{k=4}^{\infty}\left(\frac{4\varepsilon m}{n}\right)^{2k}\Pr[\,|a|=k\,]\ .$$

Recalling that $\Pr[\,|a|=k\,]=\frac{1}{2}e^{-1}\frac{1}{k!}+\frac{1}{2}e^{-\frac{\varepsilon m}{n}}\frac{1}{k!}\left(\frac{\varepsilon m}{n}\right)^{2k}$ and that $\frac{\varepsilon m}{n}=\Theta\Big(\frac{1}{n^{1/8}}\Big)\ll 1$, with a Taylor series expansion of the first term of the sum, we finally get that

$$I\left(U;a\right)=O\left(\frac{\varepsilon^{8}m^{8}}{n^{8}}\right)\ .$$

Therefore, $I\left(U;A\right)\leq\sum_{z=1}^{n-1}O\Big(\frac{\varepsilon^{8}m^{8}}{n^{8}}\Big)=O\Big(\frac{\varepsilon^{8}m^{8}}{n^{7}}\Big)$, which is $o(1)$ unless $m=\Omega\Big(n^{7/8}/\varepsilon\Big)$. This completes the proof of this branch of the lower bound.

6.3 Third Lower Bound Regime: $\Omega\Big(\frac{\sqrt{n}}{\varepsilon^{2}}\Big)$ for $\varepsilon<n^{-3/8}$

Finally, assume we are in the regime where

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

i.e., $\varepsilon<n^{-3/8}$. In this case, we can show the desired lower bound by a simple reduction from the known hard instances for uniformity testing. Let N be an even positive integer. It is shown in [Pan08] that $\Omega\Big(\sqrt{N}/\varepsilon^{2}\Big)$ samples are required to distinguish between (a) the uniform distribution on $[N]$, and (b) a distribution selected at random by pairing consecutive elements $2i,2i+1$ and setting the probability mass of each pair to be either $\Big(\frac{1+2\varepsilon}{N},\frac{1-2\varepsilon}{N}\Big)$ or $\Big(\frac{1-2\varepsilon}{N},\frac{1+2\varepsilon}{N}\Big)$ independently and uniformly at random.

We map these instances to our conditional independence setting as follows: Let $N=4n$. We map $[N]$ to the set $\{0,1\}\times\{0,1\}\times[n]$ via the mapping $\Phi:[N]\to\{0,1\}\times\{0,1\}\times[n]$ defined as follows: $\Phi(2i)=(0,0,i)$, $\Phi(2i+1)=(0,1,i)$, $\Phi(2i+2)=(1,0,i)$, and $\Phi(2i+3)=(1,1,i)$.

For a distribution $p\in\Delta([N])$ adversarially selected as described above, the following conditions are satisfied: (1) In case (a), $\Phi(p)$ is the uniform distribution on $\{0,1\}\times\{0,1\}\times[n]$ and therefore $\Phi(p)\in\mathcal{P}_{\{0,1\},\{0,1\}|[n]}$. (2) In case (b), it is easy to see that for each fixed value of the third coordinate, the conditional distribution on the first two coordinates is one of the following:

$$\frac{1}{4}\begin{pmatrix}1+2\varepsilon & 1-2\varepsilon\\ 1+2\varepsilon & 1-2\varepsilon\end{pmatrix},\quad\frac{1}{4}\begin{pmatrix}1+2\varepsilon & 1-2\varepsilon\\ 1-2\varepsilon & 1+2\varepsilon\end{pmatrix},\quad\frac{1}{4}\begin{pmatrix}1-2\varepsilon & 1+2\varepsilon\\ 1-2\varepsilon & 1+2\varepsilon\end{pmatrix},\ \text{or}\quad\frac{1}{4}\begin{pmatrix}1-2\varepsilon & 1+2\varepsilon\\ 1+2\varepsilon & 1-2\varepsilon\end{pmatrix}\ .$$

It is clear that each of these four distributions is $\Omega(\varepsilon)$-far from independent. Therefore, by Lemma 2.2 it follows that $\mathrm{d_{TV}}\Big(\Phi(p),\mathcal{P}_{\{0,1\},\{0,1\}|[n]}\Big)=\Omega(\varepsilon)$.

In conclusion, we have established that any testing algorithm for $\mathcal{P}_{\{0,1\},\{0,1\}|[n]}$ can be used (with parameter ε) to test uniformity over $[N]$ (with parameter $\varepsilon'=O(\varepsilon)$). This implies a lower bound of $\Omega\Big(\sqrt{N}/\varepsilon'^{2}\Big)=\Omega(\sqrt{n}/\varepsilon^{2})$ on the sample complexity of ε-testing conditional independence.

7 Sample Complexity Lower Bound for $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = [n]$

In this section, we outline the proof of the (tight) sample complexity lower bound of $\Omega\left(n^{7/4}\right)$ for testing conditional independence in the regime that $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = [n]$, and $\varepsilon = \Omega(1)$.

Specifically, we will show that, when $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = n$ and $\varepsilon = 1/20$, it is impossible to distinguish between conditional independence and ε-far from conditional independence with $s = o(n^{7/4})$ samples. To do this, we begin by producing an adversarial ensemble of distributions. The adversarial distribution will be designed to match the cases where the upper bound construction will be tight. In particular, each conditional marginal distribution will have about $n^{3/4}$ heavy bins and the rest of the bins light. The difference between our distribution and the product of the marginals (if it exists) will be uniformly distributed about the light bins.

First we will come up with an ensemble where X and Y are conditionally independent and then we will tweak it slightly. For the first distribution, we let the distribution over Z be uniform. For each value of $Z = z$, we pick random subsets $A_z, B_z \subset [n]$ of size $n^{3/4}$ (which we assume to be an integer), which will be the heavy bins of the conditional distribution. We then let the conditional probabilities be defined by

$$\Pr[\, X = j \mid Z = z \,] = \begin{cases} n^{-3/4}/2 & \text{if } j \in A_z \\ 1/(2(n - n^{3/4})) & \text{else} \end{cases}$$

and

$$\Pr[\, Y = j \mid Z = z \,] = \begin{cases} n^{-3/4}/2 & \text{if } j \in B_z \\ 1/(2(n - n^{3/4})) & \text{else.} \end{cases}$$

We then let X and Y be conditionally independent on Z, defining the distribution over (X, Y, Z).

Finally, we introduce a new one bit variable W. In ensemble $\mathcal{D}_0$, W is an independent random bit. In ensemble $\mathcal{D}_1$, the conditional distribution on W given (X, Y, Z) is uniform random if $X \in A_Z$ or $Y \in B_Z$, but otherwise is given by a uniform random function $f\colon [n] \times [n] \times [n] \to \{0, 1\}$. In particular, in this case, W is determined by the values of X, Y, Z, though different elements of $\mathcal{D}_1$ will give different functions. Note that elements of $\mathcal{D}_0$ have XW and Y conditionally independent on Z, whereas elements of $\mathcal{D}_1$ are ε-far from any such distribution. We show that no algorithm that takes s samples from a random distribution from one of these families can reliably distinguish which family the samples came from.

In particular, let F be a uniform random bit. Let S be a sequence of s quadruples (W, X, Y, Z) obtained by picking a random element p from $\mathcal{D}_F$ and taking s independent samples from p. It suffices to show that one cannot reliably recover F from S. Note that with high probability S contains at most $t := 2s/n + \log n = o(n^{3/4})$ samples for each value of Z. Therefore of we let T_z be a sequence of t independent samples from p conditioned on $Z = z$ for each z, it suffices to show that F cannot be reliably recovered from $T_1, \ldots, T_n$. For this it suffices to show that $I(F; T_1, \ldots, T_n) = o(1)$. Since the T_z are conditionally independent on F, we have that $I(F; T_1, \ldots, T_n) \leq \sum_{z=1}^{n} I(F; T_z)$ and thus it suffices to show that $I(F; T_z) = o(1/n)$ for every z. Since this shared information is clearly the same for each z, we will suppress the subscript.

We say that two distinct elements of T *collide* if they have the same values of X and Y. We note that if we condition on the values of X and Y in the elements of T, that the values of W for the elements that do not collide are uniform random bits independent of the other values of W and of F. Therefore, no information can be gleaned from these W's. This means that all of the information comes from W's associated with collisions. Unfortunately, most collisions (as we will see) come from heavy values of either X or Y (or both), and these cases will also provide no extra information.

More formally, note that

$$I(F;T) = \mathbb{E}_{M\sim T}\left[O\left(\min\left(1,\left(1-\frac{\Pr[T=M\mid F=0]}{\Pr[T=M\mid F=1]}\right)^2\right)\right)\right].$$

We claim that if M has C pairs that collide, then

$$\left(1-\frac{\Pr[T=M\mid F=0]}{\Pr[T=M\mid F=1]}\right)^2 = O(C^2/n).$$

Our result will then follow from the observation that $\mathbb{E}[C^2] = O(t^2/n^{3/2}) = o(1)$.

Given M, call a value of X (resp. Y) *extraneous* if it

- occurs as an X- (resp. Y-) coordinate of an element of M; and
- does not occur as an X- (resp. Y-) coordinate of an element of M involved in a collision.

We claim that our bound

$$\left(1-\frac{\Pr[T=M\mid F=0]}{\Pr[T=M\mid F=1]}\right)^2 = O(C^2/n). \tag{39}$$

holds even after conditioning on which extraneous X are in A and which extraneous Y are in B. This will be sufficient since which X and Y are extraneous, and which of them are in A or B, are both independent of F. Let M_E be the X and Y values of M along with the information on which extraneous X and Y are in A or B. It is enough to bound

$$\left(1-\frac{\Pr[T=M\mid F=0]}{\Pr[T=M\mid F=1]}\right)^2 = \left(\frac{\Pr[T=M\mid F=1]-\Pr[T=M\mid F=0]}{\Pr[T=M\mid F=1]}\right)^2.$$

It thus suffices to bound

$$\left(\mathrm{d}_{\mathrm{TV}}((T\mid F=0,M_E),(T\mid F=1,M_E))\right)^2.$$

Say that a pair of colliding elements of M is *light* if neither the corresponding X nor the corresponding Y are in A or B. Observe that if we condition on no collision in M being light, the conditional distributions of T on $F=0$ and on $F=1$ are the same. Therefore, the expression above is bounded by

$$\Pr[\text{ There exists a light collision in } M]^2.$$

Thus, it suffices to bound the probability that M contains a light collision given its values of X and Y. If M (ignoring the values of W and Z) is $((x_1,y_1),\ldots,(x_t,y_t))$ the probability of seeing this M conditioned on the values of A and B is $\Theta(n^{-2t+|\{\ i\in[t]\ :\ x_i\in A\ \}|/4+|\{\ i\in[t]\ :\ y_i\in B\ \}|/4})$. Now given some set of a of values of X appearing in M, the prior probability that those are the values of X in M appearing in A is $n^{-a/4}\cdot\phi$, where ϕ is some quantity that changes by only a $1+o(1)$ factor if a single element is added or removed from the set of X's. A similar relation holds for Y's. Given this, and conditioning on whether the extraneous X's and Y's are in A and B, we note that each non-extraneous X or Y that is in A or B contributes a factor of roughly $n^{-1/4}$ to the prior probability of having that configuration of elements in A or B, but contributes a factor of at least $n^{1/2}$ to the conditional probability of seeing the M that we saw given those values of A and B. Therefore, even conditioned on the previously determined lightness of other collisions, each collision has only a $O(n^{-1/2})$ probability of being light. Therefore the probability that there is a light collision is $O(C/n^{1/2})$. This implies Eq. (39) and completes the proof of our lower bound.

References

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Proceedings of NIPS'15*, 2015. 1.1, 1.2, 5.3

[Agr92] A. Agresti. A survey of exact inference for contingency tables. *Statist. Sci.*, 7(1):131–153, 02 1992. 1.1

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001. 1.2, 5.3

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000. 1.1

[BH07] R. Blundell and J. L. Horowitz. A non-parametric test of exogeneity. *The Review of Economic Studies*, 74(4):1035–1058, 2007. 1.1

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004. 1.1

[BT14] T. Bouezmarni and A. Taamouti. Nonparametric tests for conditional independence using conditional distributions. *Journal of Nonparametric Statistics*, 26(4):697–719, 2014. 1.1

[Can15] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015. 1.1

[CDGR16] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016. See also [CDGR17] (full version). 1.1

[CDGR17] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, pages 1–59, 2017. 7

[CDKS17] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing Bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 370–448, 2017. 1.1, 6.1, 6.1

[CDS17] C. L. Canonne, I. Diakonikolas, and A. Stewart. Fourier-based testing for families of distributions. *CoRR*, abs/1706.05738, 2017. 1.1

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014. (document), 1.1, 1.2

[Coc54] W. G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451, 1954. 1.1

[Daw79] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. 1.1

[DDK18] C. Daskalakis, N. Dikkala, and G. Kamath. Testing Ising models. In *SODA*, 2018. To appear. 1.1

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013. 1.1

[DGPP16] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016. 1.1

[DGPP17] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Sample-optimal identity testing with high probability. *CoRR*, abs/1708.02728, 2017. 1.1

[DK16] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557. (document), 1.1, 1.2, 1.2, 1.2, 1.3.1, 1.3.2, 1.3.3, 2.2, 2.2, 2.3, 5.3, 6.1, 6.1, B.3

[DKN15a] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015. 1.1

[DKN15b] I. Diakonikolas, D. M. Kane, and V. Nikishkin. 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*, 2015. 1.1

[DKN17] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, pages 8:1–8:15, 2017. 1.1

[DM01] M. A. Delgado and W. G. Manteiga. Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics*, 29(5):1469–1507, 2001. 1.1

[dMASdBP14] P. de Morais Andrade, J. M. Stern, and C. A. de Braganca Pereira. Bayesian test of significance for conditional independence: The multinomial model. *Entropy*, 16(3):1376–1395, 2014. 1.1

[Dob59] R. L. Dobrušin. A general formulation of the fundamental theorem of Shannon in the theory of information. *Uspehi Mat. Nauk*, 14(6 (90)):3–104, 1959. 2.1

[DP17] C. Daskalakis and Q. Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 697–703, 2017. 1.1

[EO87] D. Easley and M. O'Hara. Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19(1):69 – 90, 1987. 1.1

[Fis24] R. A. Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924. 1.1

[Gol17] O. Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. 1.1

[Gra80] C.W.J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(Supplement C):329 – 352, 1980. 1.1

[GS10] G. Geenens and L. Simar. Nonparametric tests for conditional independence in two-way contingency tables. *Journal of Multivariate Analysis*, 101(4):765–788, 2010. 1.1

[HPS16] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 3315–3323, 2016. 1.1

[Hua10] T.-M. Huang. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.*, 38(4):2047–2091, 08 2010. 1.1

[LG96] O. Linton and P. Gozalo. Conditional Independence Restrictions: Testing and Estimation. Cowles Foundation Discussion Papers 1140, Cowles Foundation for Research in Economics, Yale University, 1996. 1.1

[LRR11] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011. 1.2, 5.3

[MH59] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, April 1959. PMID: 13655060. 1.1, 1.1

[Nea03] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., 2003. 1.1

[NUU17] K. Natori, M. Uto, and M. Ueno. Consistent learning Bayesian networks with thousands of variables. In *Proceedings of The 3rd International Workshop on Advanced Methodologies for Bayesian Networks*, volume 73 of *Proceedings of Machine Learning Research*, pages 57–68. PMLR, 20–22 Sep 2017. 1.1

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008. 1.1, 6.3

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. 1.1

[Pin05] M. S. Pinsker. On the estimation of information via variation. *Problemy Peredachi Informatsii*, 41(2):3–8, 2005. A

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012. 1.1

[SGS00] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000. 1.1, 1.1

[Son09] K. Song. Testing conditional independence via Rosenblatt transforms. *Ann. Statist.*, 37(6B):4011–4045, 12 2009. 1.1

[SW07] L. Su and H. White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807 – 834, 2007. 1.1

[SW08] L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008. 1.1

[SW14] L. Su and H. White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27 – 44, 2014. Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions. 1.1

[TBA06] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct 2006. 1.1

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014. 1.1

[WH17] X. Wang and Y. Hong. Characteristic function based testing for conditional independence: a nonparametric regression approach. *Econometric Theory*, pages 1–35, 2017. 1.1, 1.1

[Wyn78] A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Inform. and Control*, 38(1):51–59, 1978. 2.1

[ZPJS11] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813. AUAI Press, 2011. 1.1

A 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 A.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*

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

Proof. This follows immediately upon noting that by Lemma A.1 (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 A.2 (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 $\Omega(\varepsilon)$.[1] 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 A.1. *Assume $(X,Y,Z) \sim p$, where $p \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ with $|\mathcal{X}| = \ell_1$, $|\mathcal{Y}| = \ell_2$, 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 $I(X;Y|Z) \leq O(\varepsilon \log(\ell_1\ell_2/\varepsilon))$;*
- *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 [Pin05] (and Lemma 2.2). □

[1] 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.

B Deferred Proofs from Section 2

B.1 Proof of Lemma 2.1

The proof follows from the following chain of (in-)equalities:

$$\begin{aligned}
2\mathrm{d}_{\mathrm{TV}}(p,p') &= \sum_{(i,j,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} |p(i,j,z) - p'(i,j,z)| \\
&= \sum_{(i,j,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} |p_Z(z)\cdot p_z(i,j) - p'_Z(z)\cdot p'_z(i,j)| \\
&= \sum_{(i,j,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} |p_Z(z)\cdot (p_z(i,j) - p'_z(i,j)) + (p_Z(z) - p'_Z(z))\cdot p'_z(i,j)| \\
&\le \sum_{(i,j,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p_Z(z)\cdot |p_z(i,j) - p'_z(i,j)| + \sum_{(i,j,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} |p_Z(z) - p'_Z(z)|\cdot p'_z(i,j) \\
&= \sum_{z\in\mathcal{Z}} \left(p_Z(z)\cdot \sum_{(i,j)\in\mathcal{X}\times\mathcal{Y}} |p_z(i,j) - p'_z(i,j)| \right) + \sum_{z\in\mathcal{Z}} \left(|p_Z(z) - p'_Z(z)|\cdot \sum_{(i,j)\in\mathcal{X}\times\mathcal{Y}} p'_z(i,j) \right) \\
&= 2\sum_{z\in\mathcal{Z}} p_Z(z)\cdot \mathrm{d}_{\mathrm{TV}}(p_z,p'_z) + 2\mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z)\ ,
\end{aligned}$$

where the fourth line used the triangle inequality and the last line used the fact that $\sum_{(i,j)\in\mathcal{X}\times\mathcal{Y}} p'_z(i,j) = 1$. This completes the proof of the first part of the lemma. For the second part, we note that the equality in (2) holds if and only if the triangle inequality in the fourth line above holds with equality, i.e., when $p_Z = p'_Z$. This completes the proof of Lemma 2.1. □

B.2 Proof of Lemma 2.2

Let $p' \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$ be such that $\mathrm{d}_{\mathrm{TV}}(p,p') \le \varepsilon$ and $q = \sum_{z\in\mathcal{Z}} p_Z(z)q_z$. Since $\mathrm{d}_{\mathrm{TV}}(p,q) \le \mathrm{d}_{\mathrm{TV}}(p,p') + \mathrm{d}_{\mathrm{TV}}(p',q) \le \varepsilon + \mathrm{d}_{\mathrm{TV}}(p',q)$, it suffices to show that $\mathrm{d}_{\mathrm{TV}}(p',q) \le 3\varepsilon$. By Lemma 2.1, we have that

$$\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(q,p') &\le \sum_{z\in\mathcal{Z}} q_Z(z)\cdot \mathrm{d}_{\mathrm{TV}}(q_z,p'_z) + \mathrm{d}_{\mathrm{TV}}(q_Z,p'_Z) \\
&= \sum_{z\in\mathcal{Z}} p_Z(z)\cdot \mathrm{d}_{\mathrm{TV}}(q_z,p'_z) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) \\
&= \sum_{z\in\mathcal{Z}} p_Z(z)\cdot \mathrm{d}_{\mathrm{TV}}\left(p_{z,X}\otimes p_{z,Y}, p'_{z,X}\otimes p'_{z,Y}\right) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) \\
&\le \sum_{z\in\mathcal{Z}} p_Z(z)\cdot \left(\mathrm{d}_{\mathrm{TV}}\left(p_{z,X},p'_{z,X}\right) + \mathrm{d}_{\mathrm{TV}}\left(p_{z,Y},p'_{z,Y}\right)\right) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) \\
&= \sum_{z\in\mathcal{Z}} p_Z(z)\mathrm{d}_{\mathrm{TV}}\left(p_{z,X},p'_{z,X}\right) + \sum_{z\in\mathcal{Z}} p_Z(z)\mathrm{d}_{\mathrm{TV}}\left(p_{z,Y},p'_{z,Y}\right) + \mathrm{d}_{\mathrm{TV}}(p_Z,p'_Z) \\
&\le 3\varepsilon\ ,
\end{aligned}$$

where the second line uses the fact that $q_Z = p_Z$, the third line uses the fact that $q_z = p_{z,X}\otimes p_{z,Y}$ (Definition 2.1) and that $p'_z = p'_{z,X}\otimes p'_{z,Y}$ (since $p' \in \mathcal{P}_{\mathcal{X},\mathcal{Y}|\mathcal{Z}}$), the fourth line uses the sub-additivity of total variation distance for product distributions, and the last line uses the fact that each of the three terms in the fifth line is bounded from above by $\mathrm{d}_{\mathrm{TV}}(p,p')$. This completes the proof of Lemma 2.2. □

B.3 Proof of Lemma 2.3

This lemma is essentially shown in [DK16]. The only difference is that we require a proof for (ii) when S is a set of m independent samples (as opposed to $\mathrm{Poi}(m)$ samples from p in [DK16]). We show this by an explicit calculation below.

Let a_i equal one plus the number of copies of i in S, i.e. $a_i := 1 + \sum_{j \in S} \mathbb{1}_{\{i=j\}}$. We note that the expected squared ℓ_2-norm of p_S is $\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^{a_i} p_i^2/a_i^2\right] = \sum_{i=1}^n p_i^2 \mathbb{E}[1/a_i]$. Further, a_i is distributed as $1 + X$ where X is a $\mathrm{Bin}(m, p_i)$ random variable. Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{1+X}\right] &= \sum_{k=0}^{m} \frac{1}{k+1}\binom{m}{k} p_i^k (1-p_i)^{m-k} = \frac{1}{(m+1)p_i}\sum_{k=0}^{m}\binom{m+1}{k+1} p_i^{k+1}(1-p_i)^{(m+1)-(k+1)} \\ &= \frac{1}{(m+1)p_i}\sum_{\ell=1}^{m+1}\binom{m+1}{\ell} p_i^{\ell}(1-p_i)^{(m+1)-\ell} = \frac{1-(1-p_i)^{m+1}}{(m+1)p_i} \le \frac{1}{(m+1)p_i}\,.\end{aligned}$$

This implies $\mathbb{E}\left[\|p_S\|_2^2\right] \le \sum_{i=1}^n p_i^2/(mp_i) = (1/m)\sum_{i=1}^n p_i = 1/m$. which completes the proof. □

B.4 Proof of Claim 2.1

Recalling that $\mathbb{E}[N] = \lambda$ and $\mathbb{E}[N^2] = \lambda + \lambda^2$, we get

$$\mathbb{E}\left[N\mathbb{1}_{\{N\ge 4\}}\right] = e^{-\lambda}\sum_{k=4}^{\infty} k\frac{\lambda^k}{k!} = \lambda - e^{-\lambda}\left(\lambda + \lambda^2 + \frac{1}{2}\lambda^3\right) := f(\lambda)\,,$$

and

$$\mathrm{Var}\, N\mathbb{1}_{\{N\ge 4\}} = \left(\lambda + \lambda^2 - e^{-\lambda}\left(\lambda + 2\lambda^2 + \frac{3}{2}\lambda^3\right)\right) - \left(\lambda - e^{-\lambda}\left(\lambda + \lambda^2 + \frac{1}{2}\lambda^3\right)\right)^2 := g(\lambda)\,.$$

From these expressions, it is easy to check that (i) $\lim_{x\to 0}\frac{f(x)}{g(x)} = \frac{1}{4}$, and (ii) $\lim_{x\to\infty}\frac{f(x)}{g(x)} = 1$. From the definition as a variance of a non-constant random variable, it follows that $g(x) > 0$ for all $x > 0$, from which we get that (iii) $\frac{f}{g}$ is continuous and positive on $[0,\infty)$. Combining these three statements, we get that $\frac{f}{g}$ achieves a minimum c on $[0,\infty)$, and that this minimum is positive. This implies the result with $C := 1/c$. The value 4.22 comes from studying numerically this ratio, whose minimum is achieved for $x \simeq 1.1457$.

B.5 Proof of Claim 2.2

Let $a, b \ge 0$, $\lambda > 0$, and assume $X \sim \mathrm{Poisson}(\lambda)$. Without loss of generality, suppose $0 < a \le b$ (the case $a = 0$ being trivial). We can rewrite $X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\ge 4\}}$ as Y with

$$Y := X^2\mathbb{1}_{\{X\le a\}} + \sqrt{a}X^{3/2}\mathbb{1}_{\{a<X\le b\}} + \sqrt{ab}X\mathbb{1}_{\{X>b\}}$$

which implies

$$Y^2 = X^4\mathbb{1}_{\{X\le a\}} + aX^3\mathbb{1}_{\{a<X\le b\}} + abX^2\mathbb{1}_{\{X>b\}}\,.$$

By linearity of expectation, the original claim boils down to proving there exists $C > 0$ such that

$$\begin{aligned}&\mathbb{E}\left[X^4\mathbb{1}_{\{4\le X\le a\}}\right] + \mathbb{E}\left[aX^3\mathbb{1}_{\{a<X\le b\}}\mathbb{1}_{\{X\ge 4\}}\right] + \mathbb{E}\left[abX^2\mathbb{1}_{\{X>b\}}\mathbb{1}_{\{X\ge 4\}}\right] \\ &\qquad \le C\left(\mathbb{E}\left[X^2\mathbb{1}_{\{4\le X\le a\}}\right] + \mathbb{E}\left[\sqrt{a}X^{3/2}\mathbb{1}_{\{a<X\le b\}}\mathbb{1}_{\{X\ge 4\}}\right] + \mathbb{E}\left[\sqrt{ab}X\mathbb{1}_{\{X>b\}}\mathbb{1}_{\{X\ge 4\}}\right]\right) \\ &\qquad\quad + \left(\mathbb{E}\left[X^2\mathbb{1}_{\{4\le X\le a\}}\right] + \mathbb{E}\left[\sqrt{a}X^{3/2}\mathbb{1}_{\{a<X\le b\}}\mathbb{1}_{\{X\ge 4\}}\right] + \mathbb{E}\left[\sqrt{ab}X\mathbb{1}_{\{X>b\}}\mathbb{1}_{\{X\ge 4\}}\right]\right)^2\end{aligned}$$

and since $(x+y+z)^2 \geq x^2+y^2+z^2$ for $x,y,z \geq 0$, it is enough to show

$$\mathbb{E}\Big[\beta^2 X^{2\alpha}\mathbb{1}_{\{X\in S\}}\Big] \leq C\beta\mathbb{E}\Big[X^{\alpha}\mathbb{1}_{\{X\in S\}}\Big] + \beta^2\mathbb{E}\Big[X^{\alpha}\mathbb{1}_{\{X\in S\}}\Big]^2$$

for $\alpha, \beta > 0$, and $S \subseteq \mathbb{R}_+$ an interval. This in turn follows from arguments similar to that of the proof of Claim 2.1.

B.6 Proof of Claim 2.3

For $\lambda < 8$, we can take bound the expectation by the contribution of $X = 4$ as $\mathbb{E}\Big[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\Big] \geq 4\sqrt{\min(4,a)\min(4,b)}\lambda^4/4! \geq \lambda^4/3$. Then $\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)} \geq \lambda\sqrt{\min(\lambda,2)\min(\lambda,2)} \geq \lambda^2/4 \geq \lambda^4/256$. Thus $\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4) \leq 256\lambda^4$. Putting this together we have that for $\lambda < 8$,

$$\mathbb{E}\Big[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\Big] \geq (1/768)\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4)\ .$$

To deal with the $\lambda \geq 8$ case, we claim that $\Pr[\,X \geq \lfloor\lambda/2\rfloor\,] \geq 1/2$ in this case. To see this, we just need to expand $1 = \exp(-\lambda)\sum_{k=0}^{\infty} f(k)\lambda^k/k!$ and note that for $1 \leq k \leq \lambda/2$, the ratio of the k term to the $k-1$ term is at least $\lambda/k \geq 2$. Thus the sum of the first $\lfloor\lambda/2\rfloor$ terms is smaller than the $k = \lfloor\lambda/2\rfloor$ term and so

$$\exp(-\lambda)\sum_{k=0}^{\lfloor\lambda/2\rfloor-1}\lambda^k/k! \leq \exp(-\lambda)\sum_{k=\lfloor\lambda/2\rfloor}^{\infty}\lambda^k/k\ .$$

The RHS is $\Pr[\,X \geq \lfloor\lambda/2\rfloor\,]$ and the LHS is $\Pr[\,X < \lfloor\lambda/2\rfloor\,] = 1 - \Pr[\,X \geq \lfloor\lambda/2\rfloor\,]$ and so we have that $\Pr[\,X \geq \lfloor\lambda/2\rfloor\,] \geq 1/2$ as claimed.

For $8 \leq \lambda \leq 2\min a, b$, we have

$$\begin{aligned}
\mathbb{E}\Big[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\Big] &\geq \mathbb{E}\Big[X^2\mathbb{1}_{\{X\geq\lfloor\lambda/2\rfloor\}}\Big] \\
&\geq (1/2)(\lfloor\lambda/2\rfloor)^2 \geq \lambda^2/3 \\
&\geq \min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}/6, \lambda^4/3) \\
&\geq (1/6)\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4)
\end{aligned}$$

For $\lambda \geq 2\max a, b, 4$, noting that for $X \geq \lambda/2$, $\sqrt{\min(X,a)\min(X,b)} = \sqrt{ab}$, we have

$$\begin{aligned}
\mathbb{E}\Big[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\Big] &\geq \mathbb{E}\Big[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq\lfloor\lambda/2\rfloor\}}\Big] \\
&\geq \lfloor\lambda/2\rfloor\sqrt{ab} \\
&\geq \sqrt{ab}\lambda/3 \\
&\geq (1/6)\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4)
\end{aligned}$$

The final case we need to consider is when λ is between $2a$ and $2b$ and the maximum of those is over 4

Supposing without loss of generality that $a \leq b$, for $\max 2a, 4 \leq \lambda \leq 2b$, we have

$$\begin{aligned}
\mathbb{E}\left[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq 4\}}\right] &\geq \mathbb{E}\left[X\sqrt{\min(X,a)\min(X,b)}\mathbb{1}_{\{X\geq \lfloor\lambda/2\rfloor\}}\right] \\
&\geq \lfloor\lambda/2\rfloor^{3/2}\sqrt{a} \\
&\geq \sqrt{ab}\lambda^{3/2}/4 \\
&\geq (1/8)\min(\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4)\ .
\end{aligned}$$