Conditional Independence Testing for Discrete Distributions: Beyond χ^2- and G-tests

Ilmun Kim[†] and coauthors

[†]Departments of Statistics and Data Science, Yonsei University

September 18, 2022

Abstract

This paper is concerned with the problem of conditional independence testing for discrete data. In recent years, the computer science community has shed new light on this fundamental problem, emphasizing finite-sample optimality. Their non-asymptotic viewpoint has led to novel conditional independence tests that enjoy certain optimality under various regimes. Despite their attractive theoretical properties, the considered tests are not necessarily practical, relying on a Poissonization trick and unspecified constants in their critical values. In this work, we attempt to bridge the gap between theory and practice by reproving optimality without Poissonization and calibrating test statistics based on Monte Carlo permutations. Along the way, we also prove that classical asymptotic χ^2- and G-tests are notably sub-optimal in a high-dimensional regime, which justifies a demand for new tools. Our theoretical results are complemented by experiments based on both simulated and real-world datasets. Accompanying this paper is an `R` package `UCI` that implements the proposed tests.

1 Introduction

Conditional independence (CI) is the backbone of diverse fields in statistics, including graphical models (De Campos and Huete, 2000; Koller and Friedman, 2009) and causal inference (Spohn, 1994; Pearl, 2014; Imbens and Rubin, 2015). Among several benefits, this fundamental assumption allows us to simplify the structure of a model, thereby increasing interpretability and reducing computational costs. To justify the use of CI assumption, it is of considerable interest to test whether X and Y are independent after accounting for the effect of Z. Due to its important role, the problem of CI testing has received much attention in the past decade, resulting in numerous exciting new developments (Li and Fan, 2020, for a recent review). However, most of the recent work is dedicated to continuous data and the importance of discrete CI testing is relatively overlooked.

In discrete settings, two commonly used methods are χ^2-test (Pearson, 1900) and G-test (McDonald, 2014), and their asymptotic equivalence is well-known under regularity conditions (Chapter 14 of Bishop et al., 2007). When X and Y are binary, Cochran–Mantel–Haenszel test (Agresti, 2003) is another popular method for CI testing. Despite their popularity, these methods are asymptotic in nature, frequently calibrated by their limiting null distributions. Therefore their validity remains questionable in finite-sample scenarios. This miscalibration issue becomes more serious in high-dimensional regimes where the number of categories can be significantly larger than the sample size. Besides, the power of these methods is not well-understood except in classical fixed-dimensional settings.

Over the past few years, the theoretical computer science community has revisited statistical testing problems under the name of property testing (Canonne, 2020, 2022, for a survey). Unlike asymptotic approaches taken by statisticians, computer scientists put a great emphasis on finite-sample guarantees, leading to innovative testing procedures with non-asymptotic optimality. Equipped with different tools such as concentration inequalities, their new viewpoint has revealed a fine-grained picture of the statistical complexity of problems, which was concealed by the traditional asymptotic analysis. For discrete CI testing, Canonne et al. (2018) put forward two testing algorithms and analyze their sample complexity from a non-asymptotic perspective. Their sample complexity results are further complemented by matching lower bounds, demonstrating optimality of their procedures in various regimes. In spite of these technical advances, their approach poses several practical challenges. First, their results rely on a Poissonization trick where the sample size is treated as a Poisson random variable. This assumption greatly simplifies the theoretical analysis, but practitioners may not necessarily take Poisson sampling to ensure theoretical guarantees. Another issue worth highlighting is the dependence of unspecified constants in their critical values. In many statistical applications, the type I error is more of concern than the type II error. It is therefore desirable to set a critical value in such a way as to maximize the power, while tightly controlling the type I error. Unfortunately, it is unclear from Canonne et al. (2018) how to set their critical values to meet this criteria, thereby leaving room for improvement from a practical perspective. Indeed, this issue was the main motivation of recent work of Kim et al. (2021, 2022) that advocates the use of permutation methods in two-sample and independence testing problems.

With these issues in mind, our work makes the following contributions: (i) In Theorem 1, we depoissonize the sample complexity results of Canonne et al. (2018) and establish the same theoretical guarantees under the standard sampling setting. On a technical level, the challenge lies in dealing with the complicated dependence structure of multinomial samples. We overcome this difficulty through the negative association property of multinomial distributions, which is seemingly unappreciated in the literature of property testing. (ii) We further make the algorithms of Canonne et al. (2018) practical by leveraging the permutation method to calibrate test statistics. This resampling approach completely removes the issue arising from unspecified constants, and provably controls the type I error in any finite sample scenarios. In Theorem 2, we prove that Monte Carlo permutation tests achieve the same sample complexity as the theoretical tests of Canonne et al. (2018). (iii) The considered test statistics are linear combinations of fourth order U-statistics, which can be daunting computationally. We address this computational concern by presenting alternative linear time expressions in Proposition 1. (iv) We also prove an independent result that demonstrates sub-optimality of asymptotic χ^2- and G-tests in their power performance. This negative result naturally inspires efforts to develop new CI tests that perform better than the classical ones. (v) Finally, we provide extensive simulation results that demonstrate the practical value of the proposed methods in Section 5, and the algorithms are available from the `R` package `UCI`.

Our work is related to Tsamardinos and Borboudakis (2010) who warn the risk of asymptotic calibration for χ^2- and G-tests, and further highlight benefits of the permutation procedure in type I error control. We support their claim by proving the negative result of asymptotic χ^2- and G-tests, and demonstrate attractive properties of the permutation method both in type I and II error control. Another related work is Berrett and Samworth (2021) where the authors propose a permutation test based on a U-statistic for unconditional independence testing. Concurring with our view, Berrett and Samworth (2021) put an emphasize on the permutation approach for practical calibration and demonstrate the competitive performance of their proposal, coined `USP` test, over χ^2- and G-tests. In fact, when the conditional variable is degenerate (i.e. Z takes a single value), one of our practical proposals becomes exactly the same as that of Berrett and Samworth (2021). In this sense, our work can be considered as an extension of Berrett and Samworth (2021) to CI testing. We also refer to Agresti (1992); Yao and Tritchler (1993) that discuss exact inference methods for contingency tables. It is worth pointing out that the current paper builds on our prior work Kim et al. (2021) that proves the sample complexity results of Canonne et al. (2018) using permutation tests. However, the analysis of Kim et al. (2021) relies on Poissonization and also makes use of (computationally expensive) full permutation tests. The current work deviates from Kim et al. (2021) by removing Poissonization and employing a more computationally efficient permutation test via Monte Carlo sampling. We also propose a new permutation test, called `wUCI`-test, that avoids sample splitting and illustrate its competitive finite sample performance under a variety of settings.

Organization. The rest of this paper is organized as follows. In Section 2, we set the stage by presenting some background information on sample complexity and Poissonization. Section 3 describes the test statistics that we study, and verifies that they can be computed in a linear time complexity. Section 4 contains our main theoretical results including depoissonization and sub-optimality of χ^2- and G-tests. In Section 5, we demonstrate the empirical performance of the proposed methods based on simulated and real-world datasets, before concluding in Section 6. All the proofs of our results are relegated to the Appendix.

Notation. For a positive integer a, we use the shorthand $[a] = \{1, \ldots, a\}$. The conditional independence of X and Y given Z is symbolized as $X \perp\!\!\!\perp Y \mid Z$. Given two discrete distributions p and q, we write the L_1 distance between p and q as $\|p - q\|_1$. Following the convention, we say that random variables $X_1, \ldots, X_n$ are i.i.d. when they are independent and identically distributed. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if it holds that $C_1 \leq a_n/b_n \leq C_2$ for some positive constants C_1 and C_2, and for all n. We also write $a_n = O(b_n)$ or $a_n \lesssim b_n$ to indicate that $a_n \leq C b_n$ for some positive constant C independent of n.

2 Background

Before presenting our main results, we start by building some background knowledge on sample complexity and Poissonization.

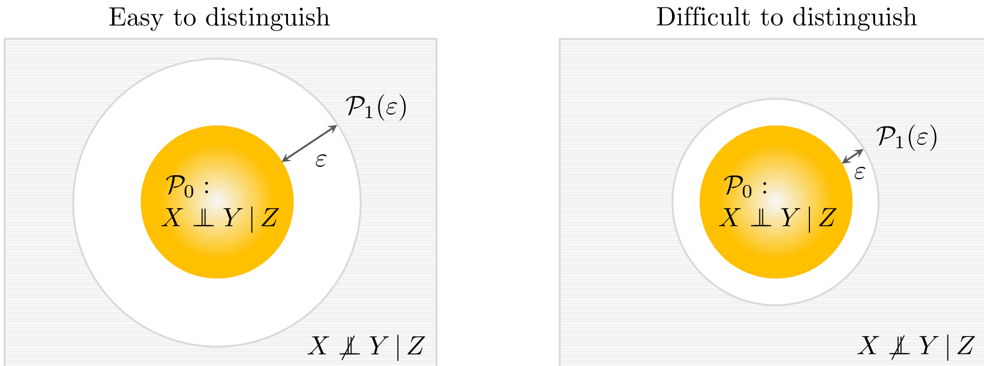

Figure 1: Schematic of the hypotheses of CI testing. The space of null distributions is ε far from the space of alternative distributions where the distance parameter ε controls the difficulty of the problem.

2.1 Setting the stage

Consider the set of discrete distributions of (X, Y, Z) on a domain $[\ell_1] \times [\ell_2] \times [d]$, denoted by $\mathcal{P}$. Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ such that $X \perp\!\!\!\perp Y \mid Z$. Given n i.i.d. random vectors $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ drawn from $p_{X,Y,Z} \in \mathcal{P}$, our goal is to distinguish

$$H_0 : p_{X,Y,Z} \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : p_{X,Y,Z} \in \mathcal{P}_1(\varepsilon) = \Big\{ p \in \mathcal{P} : \inf_{q \in \mathcal{P}_0} \|p - q\|_1 \geq \varepsilon \Big\}, \tag{1}$$

where $\varepsilon > 0$ is a distance parameter (Figure 1 for a pictorial description). As in the prior work of Canonne et al. (2018), we are interested in characterizing the minimum number of samples that allows us to address th above problem of CI testing. When the sample size n is too small, no valid test can reliably differentiate the null from the alternative. On the other hand, when the sample size n is too large, the problem becomes trivial, resulting in many successful tests. The smallest n that makes the problem feasible is called the optimal sample complexity formally defined below.

Optimal sample complexity. We define a test ϕ which maps from the samples to a binary outcome as $\phi : \{(X_i, Y_i, Z_i)\}_{i=1}^n \mapsto \{0, 1\}$. For some fixed $\alpha \in (0, 1)$, let Φ_α denote the set of level α tests such that for each $\phi \in \Phi_\alpha$,

$$\sup_{p \in \mathcal{P}_0} \mathbb{P}_p(\phi = 1) \leq \alpha.$$

The minimax risk is the worst-case type II error of an optimal level α test defined as

$$R_n(\varepsilon, \alpha) = \inf_{\phi \in \Phi_\alpha} \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\phi = 0).$$

The difficulty of the problem can be characterized as the minimum number of samples that makes the minimax risk bounded by some fixed constant $\beta \in (0, 1-\alpha)$[1]. Such minimum number of samples is called the optimal sample complexity given as

$$n^\star = \inf \Big\{ n : R_n(\varepsilon, \alpha) \leq \beta \Big\}.$$

We say that a level α test ϕ is rate optimal in sample complexity if

$$n^\star \asymp \inf \bigg\{ n : \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\phi = 0) \leq \beta \bigg\}.$$

In practice, an optimal test whose risk is exactly equal to $R_n(\varepsilon, \alpha)$ is mostly inaccessible. For this reason, we make a compromise and seek to find a rate optimal test.

Remark 1 (Choice of metrics)**.** As is well-known, one half of the L_1 distance between two distributions p and q is the same as their total variation (TV) distance:

$$\frac{1}{2}\|p - q\|_1 = \mathrm{TV}(p, q) = \sup_A |p(A) - q(A)|,$$

where the supremum is taken over all possible measurable sets. The optimal sample complexity depend crucially on the choice of metrics in the definition of the alternative hypothesis (1). Since our aim is simply to depoissonize the previous results, we focus on the L_1 distance or equivalently the TV distance throughout this paper. We refer to the recent work of Neykov et al. (2022+) that investigates the optimal sample complexity for CI testing in terms of a Wasserstein distance.

2.2 Poissonization

Poissonization is now a standard technique to study the sample complexity in distribution testing (e.g. Valiant, 2011; Chan et al., 2014; Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019). The idea itself is old in statistics and probability theory, dating back at least to Kac (1949). See also Chapter 3.5 of Vaart and Wellner (1996). Its application to distribution testing has started in relatively recent years (e.g. Valiant, 2011) as a trick to simplify several calculations in dealing with categorical data. In particular, it is well known that when the data are generated from Poisson sampling (Algorithm 1), the number of samples falling into disjoint sets are mutually independent. This independent property lies at the heart of deriving various existing results of the sample complexity in distribution testing. See Canonne (2022) for a recent review.

2.3 General recipe and related issue

If a constant factor is not of main concern, there are straightforward ways of transferring the sample complexity obtained from Poisson sampling to the usual sampling scenario with a fixed sample size.

[1] Throughout this paper, we treat the target type I and II errors α and β as universal constants, e.g. $\alpha = \beta = 0.05$.

Algorithm 1 Poisson Sampling

Input: For a fixed $n \in \mathbb{N}$ and a distribution $p_{X,Y,Z}$

1. Draw $N \sim \text{Poisson}(n)$.

2. Generate i.i.d. $(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$ random variables from $p_{X,Y,Z}$.

Return: $\{(X_i, Y_i, Z_i)\}_{i=1}^N$

One concrete procedure, described in Neykov et al. (2021), is as follows. Given n i.i.d. copies of (X, Y, Z),

1. Draw $N \sim \text{Poisson}\big(\frac{n}{2}\big)$.

2. If $N > n$, then accept the null hypothesis.

3. If $N \leq n$, perform a test based on $(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$ and return a result.

Let ϕ be a generic test function using $(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$ and then the resulting test from the above procedure can be concretely written as $\phi^* = \mathbb{1}(N \leq n)\phi$. Suppose that the test ϕ has the type I error as well as the type II error bounded by α and β, respectively. Then the corresponding test ϕ^* has the type I and II error bounds as

$$\sup_{p \in \mathcal{P}_0} \mathbb{E}_p[\phi^*] \leq \alpha \quad \text{and} \quad \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{E}_p[1 - \phi^*] \leq \beta + \mathbb{P}(N > n).$$

Provided that the Poisson distribution is tightly concentrated around its mean, the additional term in the type II error of ϕ^* can be made small when n is relatively large in comparison to β.[2] Therefore one can transfer the sample complexity obtained under Poissonization to the usual sampling scheme up to a constant factor.

Nevertheless, due to an inefficient use of the data as well as a non-trivial chance of accepting the null irrelevant to the data, practitioners may not necessarily follow this general recipe. To address this concern, there has been an effort to depoissonize the results of distribution testing under Poisson sampling. For instance, Kim (2020) revisits the truncated χ^2 test for goodness-of-fit testing proposed by Balakrishnan and Wasserman (2019) and establishes the same optimality without Poissonization. We also refer to Jacquet and Szpankowski (1998) and Chapter 2.5 of Penrose (2003) that present useful depoissonization tools, but mostly for asymptotic results. In this work, we analyze the tests proposed by Canonne et al. (2018) under the usual i.i.d. sampling with a fixed sample size and show that the same sample complexity can be achieved without any further assumption. At the heart of our technique is the negative association of multinomial distributions, which would be potentially useful for depoissonizing other sample complexity results.

[2]More precisely, an exponential tail bound for a Poisson random variable ensures that $\mathbb{P}(N > n) \leq e^{-n/8}$ (e.g. Theorem A.0.8 of Canonne, 2022) so that if $n \geq 8\log(1/\beta)$, the type II error is bounded above by 2β.

Multinomial sampling. To fix terminology, we refer to the usual sampling with fixed sample size n as *multinomial sampling* in what follows.

3 Test statistics

To describe our main results, we first need to recall the test statistics introduced by Canonne et al. (2018). As noted by Neykov et al. (2021), their test statistics can be viewed as linear combinations of U-statistics constructed using the observations of (X, Y) in the same category of Z. We describe two kinds of U-statistics considered in Canonne et al. (2018).

3.1 Unweighted U-statistic

Suppose that we observe $\sigma \geq 4$ i.i.d. observations of (X, Y) supported on $[\ell_1] \times [\ell_2]$. We let $p_{X,Y}$ denote the joint discrete distribution of (X, Y) and let p_X and p_Y denote the marginal distribution of X and Y, respectively. The first U-statistic is an unbiased estimator of the squared L_2 distance between p_{XY} and $p_X p_Y$. In more detail, borrowing the notation from Neykov et al. (2021), let

$$\phi_{ij}(qr) = \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r) - \mathbb{1}(X_i = q)\mathbb{1}(Y_j = r).$$

For four distinct observations indexed by $i, j, k, l \in [\sigma]$, the kernel of the unweighted U-statistic is defined as

$$h_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4) \in \Pi} \sum_{q \in [\ell_1], r \in [\ell_2]} \phi_{\pi_1\pi_2}(qr)\phi_{\pi_3\pi_4}(qr),$$

where Π is the set of all possible permutations of (i, j, k, l). By the linearity of the expectation, we see that h_{ijkl} is an unbiased estimator of the squared L_2 distance between p_{XY} and $p_X p_Y$. Given this kernel and the dataset $D = \{(X_1, Y_1), \ldots, (X_\sigma, Y_\sigma)\}$, the unweighted U-statistic is computed as

$$U(D) = \binom{\sigma}{4}^{-1} \sum_{i<j<k<l:(i,j,k,l) \in [\sigma]} h_{ijkl}. \tag{2}$$

It is worth pointing out that $U(D)$ is equivalent to the U-statistic for unconditional independence testing considered in Berrett et al. (2021); Berrett and Samworth (2021); Kim et al. (2022). Denote the datasets $D_m = \{(X_i, Y_i) : Z_i = m, i \in [n]\}$ and write the sample size within D_m by σ_m. The final test statistic for CI testing based on $U(D)$ is then constructed as

$$T = \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U(D_m).$$

3.2 Weighted U-statistic

The previous unweighted U-statistic may suffer from high variance especially when the L_2 norms of p_{XY} and $p_X p_Y$ are large, resulting in a sub-optimal performance (see the simulation result in Figure 2 under Scenario 1). To address this issue, Canonne et al. (2018) employ the flattening idea proposed by Diakonikolas and Kane (2016) where heavier bins are artificially broken into several pieces in a way to reduce the L_2 norms of the transformed distributions. This leads to a smaller variance of the test statistic. As made clear in Neykov et al. (2021), the flattening procedure is equivalent to using a carefully designed weighted kernel for a U-statistic.

To describe the weighted U-statistic, assume that the sample size $\sigma = 4 + 4t$ for some $t \in \mathbb{N}$, and let $t_1 = \min(t, \ell_1)$ and $t_2 = \min(t, \ell_2)$. The construction of the weighted U-statistic involves sample splitting where the dataset $D = \{(X_1, Y_1), \ldots, (X_\sigma, Y_\sigma)\}$ is split into $D_X = \{X_i : i \in [t_1]\}$, $D_Y = \{Y_i : t_1 + 1 \leq i \leq t_1 + t_2\}$ and $D_{X,Y} = \{(X_i, Y_i) : 2t + 1 \leq i \leq \sigma\}$ of sizes t_1, t_2 and $4 + 2t$, respectively. Define a_q (and a'_r) as the number of observations equal to q (and r) in D_X (and D_Y). For four distinct observations indexed by i, j, k, l in $D_{X,Y}$, consider a weighted kernel given as

$$h^{\boldsymbol{a}}_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4) \in \Pi} \sum_{q \in [\ell_1], r \in [\ell_2]} \frac{\phi_{\pi_1\pi_2}(qr)\phi_{\pi_3\pi_4}(qr)}{(1 + a_q)(1 + a'_r)}.$$

Due to the independence among the splitted datasets, the expectation of $h^{\boldsymbol{a}}_{ijkl}$ is the square of the L_2 distance between p_{XY} and $p_X p_Y$ weighted by $(1 + a_q)(1 + a'_r)$. Given the kernel, the weighted U-statistic is computed as

$$U^{\boldsymbol{a}}_W(D) = \binom{2t+4}{4}^{-1} \sum_{i<j<k<l : (i,j,k,l) \in D_{X,Y}} h^{\boldsymbol{a}}_{ijkl}, \tag{3}$$

where $(i, j, k, l) \in D_{X,Y}$ indicates taking four observations indexed by (i, j, k, l) from the dataset $D_{X,Y}$. As before, denote the datasets $D_m = \{(X_i, Y_i) : Z_i = m, i \in [n]\}$ and write the sample size within D_m by σ_m. By further writing $\omega_m = \sqrt{\min(\sigma_m, \ell_1)\min(\sigma_m, \ell_2)}$, the final test statistic for CI testing based on $U^{\boldsymbol{a}}_W(D)$ is then constructed as

$$T_W = \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \omega_m U^{\boldsymbol{a}}_W(D_m).$$

Practical considerations. As shown in Canonne et al. (2018), the test based on T_W achieves the optimal sample complexity in broader regimes than that based on T. Despite its attractive theoretical properties, the resulting test may experience a loss of practical power due to an inefficient use of the data arising from sample splitting. To mitigate this issue, we introduce another weighted U-statistic without sample splitting. First, it is worth trying to understand the weight $(1+a_q)(1+a'_r)$ at a population level. In particular, the conditional expectation of the weight given the random sample size t is $\big(1 + \min\{t, \ell_1\}p_X(q)\big)\big(1 + \min\{t, \ell_1\}p_Y(r)\big)$. This motivates us to consider a weight kernel

$$h_{ijkl}^{\boldsymbol{b}} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi} \sum_{q\in[\ell_1],r\in[\ell_2]} \frac{\phi_{\pi_1\pi_2}(qr)\phi_{\pi_3\pi_4}(qr)}{(1+b_q)(1+b'_r)},$$

where $b_q = \min\{\sigma,\ell_1\}\sigma^{-1}\sum_{i=1}^{\sigma}\mathbb{1}(X_i=q)$ and $b'_r = \min\{\sigma,\ell_2\}\sigma^{-1}\sum_{i=1}^{\sigma}\mathbb{1}(Y_i=r)$. The resulting weighted U-statistic is then computed as

$$U_W^{\boldsymbol{b}} = \binom{\sigma}{4}^{-1} \sum_{i<j<k<l:(i,j,k,l)\in[\sigma]} h_{ijkl}^{\boldsymbol{b}}. \tag{4}$$

Similarly as before, the final CI test statistic based on $U_W^{\boldsymbol{b}}(D)$ is given as

$$T_W^{\dagger} = \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m\omega_m U_W^{\boldsymbol{b}}(D_m).$$

We coin the permutation test based on $T_W^{\dagger}$ as `wUCI`-test, and formally introduce the procedure in Algorithm 4 of Appendix A.

Remark 2 (Connection with the truncated χ^2-test)**.** As pointed out by several authors (Haberman, 1988; Balakrishnan and Wasserman, 2019; Kim, 2020), the classical χ^2-test for goodness-of-fit testing can easily break down for sparse multinomial data. To address this problem, Balakrishnan and Wasserman (2019) introduce a modification of the χ^2-test by using a truncated weight function and prove its minimax optimality. Interestingly, the weight $(1+b_q)(1+b'_r)$ that we consider is closely connected to the truncated weight of Balakrishnan and Wasserman (2019) and may be regarded as an empirical counterpart for independence testing. To explain, let us simply focus on the first term in the product weight and notice that

$$\ell_1^{-1}(1+b_q) = \frac{1}{\ell_1} + \min\left(\frac{\sigma}{\ell_1},1\right)\widehat{p}_X(q) \asymp \max\left\{\frac{1}{\ell_1}, \min\left(\frac{\sigma}{\ell_1},1\right)\widehat{p}_X(q)\right\},$$

where $\widehat{p}_X(q) = \sigma^{-1}\sum_{i=1}^{\sigma}\mathbb{1}(X_i=q)$. When $\sigma\ell_1^{-1}$ is large, the right-hand side of the above equation approximates $\max\{\ell_1^{-1}, p_X(q)\}$, which is exactly the same as the truncated weight in Balakrishnan and Wasserman (2019) for goodness-of-fit testing.

3.3 Linear time expression

The original forms of the aforementioned U-statistics take $O(\sigma^4\ell_1\ell_2)$ time to compute, which can be prohibitive for large σ,ℓ_1,ℓ_2. Luckily, this computational complexity can be reduced to $O(\sigma)$ by exploiting a contingency table representation. A computationally convenient form of the unweighted U-statistic $U(D)$ is already given by Canonne et al. (2018); Berrett et al. (2021); Berrett and Samworth (2021). We also note that an alternative form of $U_W^{\boldsymbol{a}}(D)$ is provided in Canonne et al. (2018), but a naive calculation of their expression takes at least $O(\sigma\ell_1^2\ell_2^2)$ time. Here we present a general expression for the U-statistics and explain that it can be run in linear time independent of ℓ_1 and ℓ_2.

To this end, let us set some notation. Let $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_{\ell_1}\}$ and $\boldsymbol{\upsilon} = \{\upsilon_1, \ldots, \upsilon_{\ell_2}\}$ be some weight vectors with non-zero components. Given the dataset $D = \{(X_1, Y_1), \ldots, (X_\sigma, Y_\sigma)\}$, consider four distinct observations indexed by $i, j, k, l \in [\sigma]$ and define the weighted kernel associated with $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ as

$$h_{ijkl}^{\boldsymbol{\eta},\boldsymbol{\upsilon}} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi} \sum_{q\in[\ell_1],r\in[\ell_2]} \frac{\phi_{\pi_1\pi_2}(qr)\phi_{\pi_3\pi_4}(qr)}{\eta_q \upsilon_r}.$$

It is clear that the above kernel is equivalent to h_{ijkl} when $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ are one-vectors. Similarly, when $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ are defined with $1 + a_q$ and $1 + a_r'$, respectively, then the above kernel corresponds to $h_{ijkl}^{\boldsymbol{a}}$. The U-statistic based on $h_{ijkl}^{\boldsymbol{\eta},\boldsymbol{\upsilon}}$ is denoted by $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$, which is similarly computed as in (2). For $q \in [\ell_1]$ and $r \in [\ell_2]$, define $o_{qr} = \sum_{i=1}^n \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r)$, $o_{q+} = \sum_{r=1}^{\ell_2} o_{qr}$ and $o_{+r} = \sum_{q=1}^{\ell_1} o_{qr}$. With this notation in place, we give an alternative expression of $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ as follows.

Proposition 1 (Alternative expression)**.** *The U-statistic $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ can be written as*

$$U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D) \; = \; \frac{1}{\sigma(\sigma-3)}\left[A_1 + \frac{1}{(\sigma-1)(\sigma-2)}A_2 - \frac{2}{\sigma-2}A_3\right]$$

where

$$A_1 = \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\left(\frac{o_{qr}^2 - o_{qr}}{\eta_q \upsilon_r}\right), \quad A_2 = \sum_{q=1}^{\ell_1}\left(\frac{o_{q+}^2 - o_{q+}}{\eta_q}\right) \cdot \sum_{r=1}^{\ell_2}\left(\frac{o_{+r}^2 - o_{+r}}{\upsilon_r}\right),$$

$$A_3 = \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}(o_{q+}o_{+r} - o_{q+} - o_{+r} + 1)}{\eta_q \upsilon_r}.$$

We now discuss the average time complexity of $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ by assuming that the weight vectors $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ are given in advance. First of all, we note that the $\ell_1 \times \ell_2$ contingency table of D is sparse in a sense that it has at most σ non-zero entries. Importantly, the zero entries do not affect the calculation of A_1, A_2, A_3. Hence we only focus on the non-zero entries of the contingency table, which can be computed in linear time, for instance, by using hash tables (e.g. Cormen et al., 2022). Similarly, the non-zero row sums and the non-zero column sums of the contingency table can be computed in linear time independent of ℓ_1 and ℓ_2. Given the non-zero entries of $\{o_{qr}, o_{q+}, o_{+r} : q \in [\ell_1], r \in [\ell_2]\}$, the computational complexity of the terms A_1, A_2, A_3 is linear as their non-zero summands are at most $O(\sigma)$.

For the unweighted U-statistic $U(D)$, there is no additional cost for computing $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ as they are equal to one-vectors. For the weighted U-statistics $U_W^{\boldsymbol{a}}(D)$ and $U_W^{\boldsymbol{b}}(D)$, the weight vectors $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ are functions of o_{q+} and o_{+r} (computed on a separate dataset for $U_W^{\boldsymbol{a}}(D)$) and they only require an additional $O(\sigma)$ time to compute. Thus the overall average time complexity is still linear.

4 Main theoretical results

Having introduced the test statistics, we are now ready to provide the main theoretical results of this paper. In Section 4.1, we establish the same sample complexity of the tests using T and T_W under multinomial sampling. In Section 4.2, we provide and analyze more practical tests based on permutation procedures.

4.1 Sample complexity without Poissonization

In this subsection, we revisit two main results of Canonne et al. (2018), namely Theorem 1.1 and Theorem 1.3 concerning with the sample complexity of a test using T and T_W, respectively.

A. Sample complexity of a test based on T: Suppose that the test statistic T is constructed using N i.i.d. samples from $p_{X,Y,Z}$ where $N \sim \text{Poisson}(n)$. We reject the null of CI when $T > \zeta\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta > 0$. Then for $\ell_1 = \ell_2 = 2$, Theorem 1.1 of Canonne et al. (2018) proves that the resulting test has the sample complexity

$$O\left(\max\left\{\frac{d^{1/2}}{\varepsilon^2},\ \min\left\{\frac{d^{7/8}}{\varepsilon},\ \frac{d^{6/7}}{\varepsilon^{8/7}}\right\}\right\}\right). \tag{5}$$

They also prove that this sample complexity is rate optimal by presenting a matching lower bound.

B. Sample complexity of a test based on T_W: The test based on T is not necessarily optimal in the high-dimensional regime where ℓ_1 and ℓ_2 can vary with other parameters. As shown in Theorem 1.3 of Canonne et al. (2018), a more general result of the sample complexity can be derived by using T_W. In particular, given N i.i.d. samples from $p_{X,Y,Z}$ where $N \sim \text{Poisson}(n)$, we reject the null of CI when $T_W > \zeta'\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta' > 0$. The sample complexity of the resulting test satisfies

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

Moreover, this upper bound is shown to be optimal, up to constant factors, in a number of regimes. See Canonne et al. (2018) for a discussion. We note in passing that there was an error in the original proof of Theorem 1.3 in Canonne et al. (2018), which has been recently fixed by Kim (2022).

We now depoissonize the previous results and establish the same sample complexity under multinomial sampling.

Theorem 1 (Multinomial sampling)**.** *Suppose that we observe* $D_n = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$ *i.i.d. samples from* $p_{X,Y,Z}$ *with fixed sample size* n*. Then the following two statements hold:*

1. *Compute* T *based on* D_n *and reject the null if* $T > \zeta\sqrt{\min(n,d)}$ *for a sufficiently large constant* $\zeta > 0$*. Then the resulting test has the sample complexity as in* (5) *when* $\ell_1, \ell_2 = O(1)$*.*

2. *Compute* T_W *based on* D_n *and reject the null if* $T_W > \zeta'\sqrt{\min(n,d)}$ *for a sufficiently large constant* $\zeta' > 0$*. Then the resulting test has the sample complexity as in* (6)*.*

A few remarks are in order.

Remark 3.

- The above theorem essentially says that the tests based on T and T_W have the same performance in sample complexity under Poisson sampling and multinomial sampling. This result may not come as a surprise given that a Poisson random variable is strongly concentrated around its mean. See empirical evidence in Kim et al. (2021). However, the proof under multinomial sampling turns out to be highly non-trivial, requiring a delicate analysis.

- One of the main technical hurdles in the proof is to overcome a lack of independence between random variables in different bins when the sample size is no longer Poisson. The independence property is useful in analyzing the variance of the sum of U-statistics as it leads to zero covariance terms. We address the lack of independence by employing the negative association (NA) property of multinomial random vectors (Joag-Dev and Proschan, 1983). This NA property ensures that the covariance terms are non-positive, which turns out to be enough to make the same theoretical guarantee under multinomial sampling.

- Even though we remove Poissonization, the resulting tests are not necessarily practical. In particular, their critical values depend on unspecified constants ζ and ζ'. The choice of these constants, resulting in tight control of the type I error, is challenging in practice. We take a further step to address this issue via the permutation method in Section 4.2, and demonstrate their empirical performance in Section 5.

4.2 Calibration via permutations

As mentioned earlier, the tests used in Theorem 1 depend on unspecified constants, which raises the issue of practicality. This section attempts to address this problem by presenting more practical tests calibrated by the permutation method, and examine their sample complexity under multinomial sampling. In particular, we prove that their sample complexity remains the same as the corresponding (theoretical) tests in Theorem 1. As briefly mentioned earlier, a similar result was established in Theorem 5 of Kim et al. (2021) but under Poisson sampling. In contrast, we do not assume that the sample size follows a Poisson distribution and therefore reduce the gap between theory and practice. We start by describing the testing procedures that we analyze.

Algorithm 2 `UCI`: U-statistic permutation CI test

Input: Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α

For $j \in [B]$ **do**

For $m \in [d]$ **do**

- Generate $\pi \sim \text{Uniform}(\Pi_{\sigma_m})$ independent of everything else.
- Compute $U(D_m^\pi)$ as in (2) based on the permuted dataset D_m^π.

End

Set $T_j \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U(D_m^\pi)$.

End

- Set $T \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U(D_m)$ computed without permutations.
- Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_j \geq T) + 1\right].$$

Output: Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

Permutation test using T. This first test compares the test statistic T with its permutation correspondences, and rejects the null if the resulting permutation p-value is less than or equal to significance level α. To further explain, let Π_{σ_m} denote the set of all permutations of $[\sigma_m]$ for each $m \in [d]$. Given π drawn from Π_{σ_m}, we define D_m^π by rearranging Y values in D_m according to π. More specifically, suppose that we have $D_m = \{(X_1, Y_1), \ldots, (X_{\sigma_m}, Y_{\sigma_m})\}$. Then the corresponding permuted dataset becomes $D_m^\pi = \{(X_1, Y_{\pi_1}), \ldots, (X_{\sigma_m}, Y_{\pi_{\sigma_m}})\}$. Equipped with this notation, we implement Algorithm 2 and make a decision based on the output.

Permutation test using T_W. The second test that we analyze calculates its p-value by comparing T_W with its permutation correspondences. The overall procedure is similar to the previous one except that it utilizes the half-permutation procedure for a technical reason (see Remark 6 Kim et al., 2021). To explain the procedure, we decompose D_m into $D_{X,m}$, $D_{Y,m}$ and $D_{X,Y,m}$ of size $t_{1,m}$, $t_{2,m}$ and $2t_m + 4$, respectively, as in Section 3.2. Unlike Algorithm 2, we only permute Y values within $D_{X,Y,m}$ for each $m \in [d]$ and then evaluate the significance of T_W. A more detailed procedure is described in Algorithm 3.

Having outlined the permutation procedures, we now discuss their sample complexity. First of all, it is noteworthy that both permutation tests are exact level α in any finite sample scenarios. This simply follows by the fact that the original test statistic and their permutation correspondences

Algorithm 3 `wUCI_split`: weighted U-statistic permutation CI test using sample splitting

Input: Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α

For $j \in [B]$ **do**

For $m \in [d]$ **do**

- Generate $\pi \sim \text{Uniform}(\Pi_{2t_m+4})$ independent of everything else.
- Define $D_m^\pi = D_{X,m} \cup D_{Y,m} \cup D_{X,Y,m}^\pi$.
- Compute $U_W^{\boldsymbol{a}}(D_m^\pi)$ as in (3) based on the permuted dataset D_m^π.

End

Set $T_{j,W} \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \omega_m U_W^{\boldsymbol{a}}(D_m^\pi)$.

End

- Set $T_W \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \omega_m U_W^{\boldsymbol{a}}(D_m)$ computed without permutations.
- Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_{j,W} \geq T_W) + 1\right].$$

Output: Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

are exchangeable under the null. Using this observation, it can be seen that that the resulting p-values in Algorithm 2 and 3 are super-uniform (e.g. Lemma 1 of Romano and Wolf, 2005). The next theorem turns to the type II error and establishes their sample complexity.

Theorem 2 (Permutation tests)**.** *Suppose that we observe $D_n = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$ i.i.d. samples from $p_{X,Y,Z}$ with fixed sample size n. We also assume that the number of random permutations B satisfies $B \geq \max\{4(1-\alpha)\alpha^{-1}, 8\alpha^{-2}\log(4\beta^{-1})\}$ where α and β are pre-specified type I and II errors, respectively. Then the following two statements hold:*

1. *The test from Algorithm 2 has the sample complexity as in (5) when $\ell_1, \ell_2 = O(1)$.*
2. *The test from Algorithm 3 has the sample complexity as in (6).*

We highlight that the above theorem studies the random permutation tests where B is not required to increase with the sample size n. This is in contrast to full permutation tests considered in Kim et al. (2021) that enumerate all possible permutations. We also note that the constant factors in the condition of B is not tight and can be improved with more effort.

We next turn our attention to χ^2- and G-tests and discuss their sub-optimality.

4.3 Sub-optimality of χ^2- and G-tests

Practitioners frequently use χ^2- and G-tests for conditional independence, which have nice asymptotic properties in fixed dimensional settings. In this subsection, we move beyond the fixed dimensional setting and prove that these classical tests are markedly sub-optimal in terms of sample complexity. This negative result naturally gives a ground for the development of new tools. To define χ^2-test and G-test formally, let us write $o_{qrs} = \sum_{i=1}^n \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r)\mathbb{1}(Z_i = s)$ and $e_{qrs} = o_{q+s}o_{+rs}/o_{++s}$ where $o_{q+s} = \sum_{r\in[\ell_2]} o_{qrs}$, $o_{+rs} = \sum_{q\in[\ell_1]} o_{qrs}$ and $o_{++s} = \sum_{q\in[\ell_1],r\in[\ell_2]} o_{qrs}$, respectively, for $q \in [\ell_1], r \in [\ell_2], s \in [d]$. Given this notation, χ^2-test and G-test are based on the following test statistics

$$\chi^2 = \sum_{q\in[\ell_1],r\in[\ell_2],s\in[d]} \frac{(o_{qrs} - e_{qrs})^2}{e_{qrs}} \quad \text{and} \quad G = 2 \sum_{q\in[\ell_1],r\in[\ell_2],s\in[d]} o_{qrs} \log\left(\frac{o_{qrs}}{e_{qrs}}\right). \tag{7}$$

These test statistics converge to a χ^2 distribution with $(\ell_1 - 1) \times (\ell_2 - 1) \times d$ degrees of freedom under the null of conditional independence and under some regularity conditions (Tsamardinos and Borboudakis, 2010). Based on this asymptotic result, χ^2-test and G-test reject the null when χ^2 and G are larger than the $1-\alpha$ quantile of the χ^2 distribution with $(\ell_1 - 1) \times (\ell_2 - 1) \times d$ degrees of freedom. We first emphasize that these classical tests do not control the type I error uniformly over the null distributions and their validity guarantee requires that the sample size go to infinity. This is even true for the simplest case where $d = 1$, which corresponds to the unconditional independence problem (Berrett and Samworth, 2021, for details). Moreover, their asymptotic power can be exactly equal to zero in some regime where the sample size is much larger than the bound in (5) as shown below.

Proposition 2 (Sub-optimality of χ^2- and G-tests)**.** *Assume that $\ell_1 = \ell_2 = 2$, $\varepsilon = 0.5$ and $\alpha \in (0, 1)$ is some fixed constant. Further assume that $d = n \times r_n$ where r_n is an arbitrary positive sequence that increases to infinity as $n \to \infty$. In this scenario, the worse case power of χ^2- and G-tests approach zero as*

$$\lim_{n\to\infty} \inf_{p\in\mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\chi^2 > q_{1-\alpha,d}) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \inf_{p\in\mathcal{P}_1(\varepsilon)} \mathbb{P}_p(G > q_{1-\alpha,d}) = 0,$$

where $q_{1-\alpha,d}$ denotes the $1 - \alpha$ quantile of the χ^2 distribution with d degrees of freedom.

We provide some remarks on this result.

Remark 4.

- As shown in Theorem 2, the proposed tests can achieve significant power (indeed rate optimal when $\ell_1 = \ell_2 = 2$) under the same scenario as long as $r_n \lesssim n^{1/6}$, highlighting sub-optimality of χ^2- and G-tests. Moreover, the proposed tests are valid over the entire class of null distributions unlike asymptotic χ^2- and G-tests.

- At a high-level, the reason behind this negative result is that the critical values of χ^2- and G-tests are not adaptive to the underlying distribution. More specifically, we can think of a setting where most of conditional bins are empty with high probability, i.e. an intrinsic dimension of Z is much smaller than d. In this case, it is more natural to use the critical value that reflects the intrinsic dimension rather than the ambient dimension. However, χ^2- and G-test do not take this intuition into account, and thereby their test statistics can be much smaller than $q_{1-\alpha,d}$ under the alternative. This leads to asymptotically zero power as we formally prove in Appendix C.4.
- This issue can be alleviated by using the permutation approach where empty bins are ignored in calibration (Tsamardinos and Borboudakis, 2010). Nevertheless, it is unknown whether the permutation-based χ^2- and G-tests are optimal in terms of sample complexity. We leave this important question for future work.

5 Numerical analysis

In this section, we provide numerical results that compare the proposed tests (`UCI`-test in Algorithm 2 and `wUCI`-test in Algorithm 4) with χ^2-test and G-test under various scenarios. For a fair comparison, we calibrate both χ^2-test and G-test using the permutation method (as in Algorithm 2) and reject the null when their permutation p-values are less than or equal to significance level α. Throughout our simulations, we set $\alpha = 0.05$ and the number of permutations $B = 199$. All the power values reported in this section are estimated by Monte Carlo simulation with 10,000 repetitions.

5.1 Simulated data examples

We start by comparing the power of the considered tests based on synthetic datasets. We only focus on the power results given that all of the tests are calibrated by the permutation procedure, resulting in valid type I error control in any finite sample sizes. There are eight different scenarios that we consider under the alternative where the domain sizes of X, Y, Z are set to $\ell_1 = \ell_2 = 20$ and $d = 10$, respectively. The considered scenarios are described as follows.

- **Scenario 1.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) first let $p_{X,Y|Z}(x, y \,|\, z) \propto x^{-2}y^{-2}$, (ii) then replace $p_{X,Y|Z}(\ell_1, \ell_2 \,|\, z)$ with 0.015, and (iii) finally normalize $p_{X,Y|Z}$ to have its sum to be one. This setting results in a strong signal in χ^2 divergence but relatively weaker signal in L_2 distance over bins.
- **Scenario 2.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) let $p_{X,Y|Z}(x, y \,|\, z) \propto x^{-2}y^{-2}$, (ii) set $\delta = \min\{p_{X,Y|Z}(x, y \,|\, z) : x \in [2], y \in [2]\}$, and (iii) perturb $p_{X,Y|Z}$ by replacing $p_{X,Y|Z}(x, y \,|\, z)$ with $p_{X,Y|Z}(x, y \,|\, z) + (-1)^{x+y}\delta$ for $x \in [2]$ and $y \in [2]$. The resulting probability vector has a small signal in χ^2 divergence, but relatively stronger signal in L_2 distance over bins.

- **Scenario 3.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) set $p_{X,Y|Z}(x, y \,|\, z) = 0$ for all $x \in [\ell_1]$ and $y \in [\ell_2]$, (ii) set $p_{X,Y|Z}(1, 1 \,|\, z) = (1-q)^2$, $p_{X,Y|Z}(1, y \,|\, z) = (1-q)q(\ell_1-1)^{-1}$ for $y \in [\ell_2]\backslash\{1\}$, $p_{X,Y|Z}(x, 1 \,|\, z) = (1-q)q(\ell_1-1)^{-1}$ for $x \in [\ell_1]\backslash\{1\}$, $p_{X,Y|Z}(x, y \,|\, z) = q^2(\ell_1-1)^{-1}$ for $x = y \in [\ell_1]\backslash\{1\}$ where $q = 0.2$. This simulation setting is borrowed from Zhang and Zhang (2022).
- **Scenario 4.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, let $p_{X,Y|Z}(x, y \,|\, z) = \{1 + (-1)^{x+y}\}\ell_1^{-1}\ell_2^{-1}$ be a perturbed uniform distribution. This is the setting where χ^2-test, G-test and `UCI`-test perform similarly for unconditional independence testing. See Figure 5 of Berrett and Samworth (2021).
- **Scenario 5.** Set p_Z to be uniform over $[d]$. For $z = 1$, let $p_{X,Y|Z}(x, y \,|\, z) = 0.25$ for $x \in [2], y \in [2]$ and zero otherwise. In other words, there is no signal in the first category of Z. On the other hand, for $z \in [d]\backslash\{1\}$, set $p_{X,Y|Z}(x, y \,|\, z) = \{1 + (-1)^{x+y}\}\ell_1^{-1}\ell_2^{-1}$ as in Scenario 4.
- **Scenario 6.** Set p_Z to be uniform over $[d]$. For $z = 1$, $p_{X,Y|Z}(1, 1 \,|\, z) = p_{X,Y|Z}(2, 2 \,|\, z) = 0.4$, $p_{X,Y|Z}(1, 2 \,|\, z) = p_{X,Y|Z}(2, 1 \,|\, z) = 0.1$ and zero otherwise. On the other hand, for $z \in [d]\backslash\{1\}$, set $p_{X,Y|Z}(x, y \,|\, z) = \ell_1^{-1}\ell_2^{-1}$, i.e. $X \perp\!\!\!\perp Y$ for $z \in [d]\backslash\{1\}$, resulting in a sparse alternative.
- **Scenario 7.** Set $p_Z(z) \propto z^{-1}$. For $z \in [d]$, let $p_{X,Y|Z}(x, y \,|\, z) = \{1 + (-1)^{x+y}z^{-1}\}\ell_1^{-1}\ell_2^{-1}$. By construction, the signal becomes weaker as z increases and the sample size σ_z tends to be smaller as z increases.
- **Scenario 8.** $p_Z(z) \propto z^{-1}$. For $z \in [d]$, let $p_{X,Y|Z}(x, y \,|\, z) = \{1+(-1)^{x+y}(d-z+1)^{-1}\}\ell_1^{-1}\ell_2^{-1}$. Note that the signal becomes stronger as z increases, and the sample size σ_z tends to be smaller as z increases. To put it in another way, this is the reverse setting of Scenario 7.

The results are presented in Figure 2 and Figure 3. It is clear from the results that no test outperforms the others over all scenarios. In particular, χ^2-test has the highest power when the underlying distribution has a strong signal in χ^2 divergence such as Scenario 1, and similarly, `UCI`-test performs well when there is a strong signal in L_2 distance such as Scenario 2. It is also interesting to observe that `wUCI`-test shows an impressive performance compared to the others in Scenario 3, and all of the tests perform similarly in Scenario 4 (except χ^2-test with large sample sizes) where the corresponding null distribution of $(X, Y) \,|\, Z$ is uniform over bins. In Scenario 5 and Scenario 6, we are essentially in a situation where $\ell_1 = \ell_2 = 2$ for $z = 1$ and $\ell_1 = \ell_2 = 20$ for $z \in [d]\backslash\{1\}$. In these scenarios, the first bin of Z plays a less important role than the other bins in determining the value of χ^2 and G statistics (these statistics have higher variance when ℓ_1 and ℓ_2 are large). In contrast, the test statistic for `UCI`-test is mostly dominated by the data from the first bin of Z in the same scenarios. This explains an outstanding performance of χ^2-test and G-test in Scenario 5 where signals are spread out over bins except the first one. On the other hand, χ^2-test and G-test have low power in Scenario 6 where only the first bin of Z has a signal. Under the same scenarios, `UCI`-test behaves in the opposite way, attaining high power when the first bin

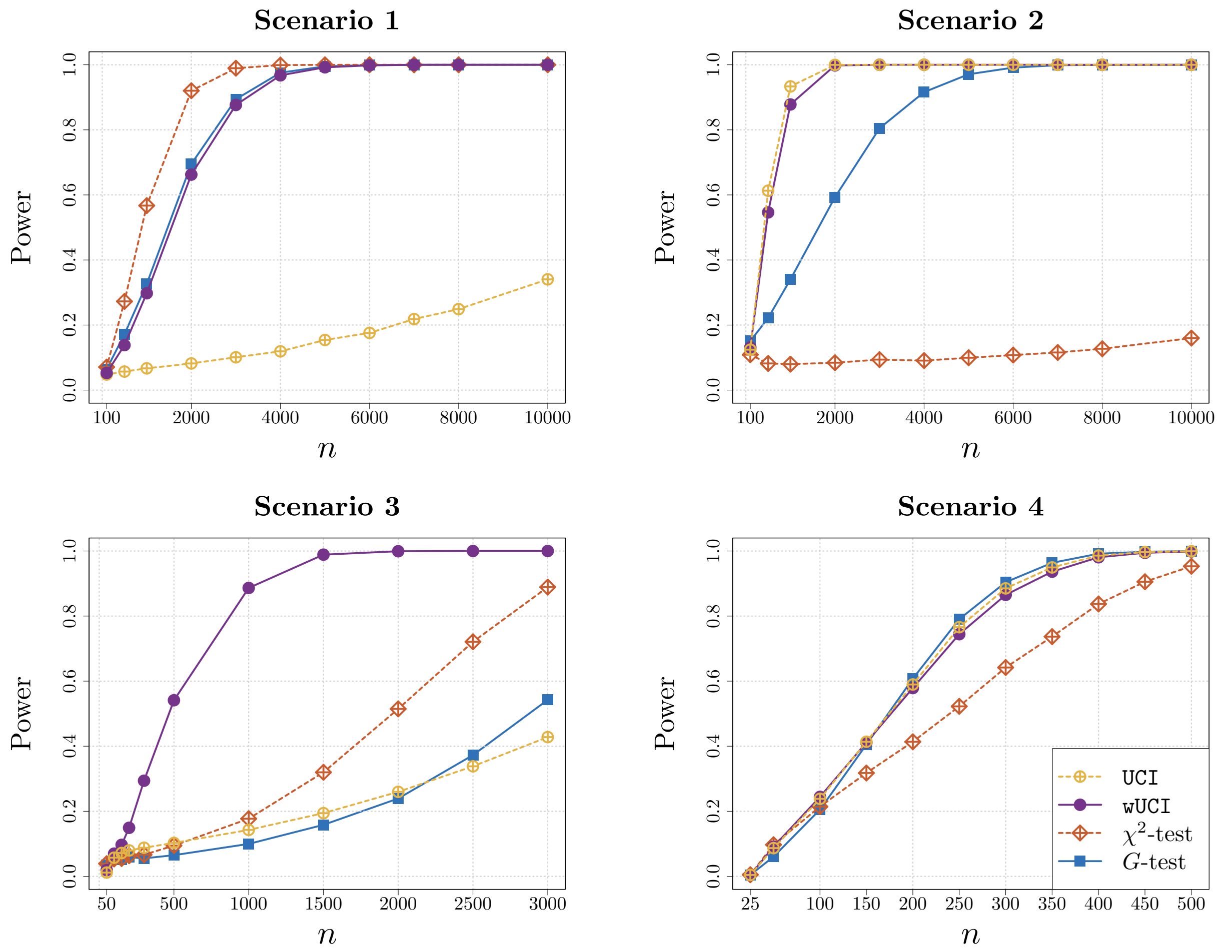

Figure 2: Power comparisons of the considered tests in Scenario 1–Scenario 4 described in Section 5.1.

is significant. Another interesting observation is that `wUCI`-test performs as powerful as G-test in Scenario 5 whereas it outperforms both χ^2- and G-tests in Scenario 6 when the sample size is large. This may be explained by the choice of weights in its statistic that roughly interpolate χ^2 weights and uniform weights as explained in Remark 2. We also note that the test statistic for `UCI`-test is a linear combination of U-statistics weighted by sample sizes over bins. This explains the relatively lower power of `UCI`-test than χ^2- and G-tests in Scenario 7 where the bins with smaller sample sizes tend to have stronger signals. In contrast, we observe the opposite behavior in Scenario 8. On the other hand, `wUCI` performs the second best both in Scenario 7 and Scenario 8.

To summarize, we observe that different tests perform better than the others under different scenarios. The proposed tests often dominate the classical ones when there are strong signals, especially in L_2 distance, over bins with large sample sizes. On the other hand, it is possible to design situations such as Scenario 1 where the proposed tests attain lower power than the classical

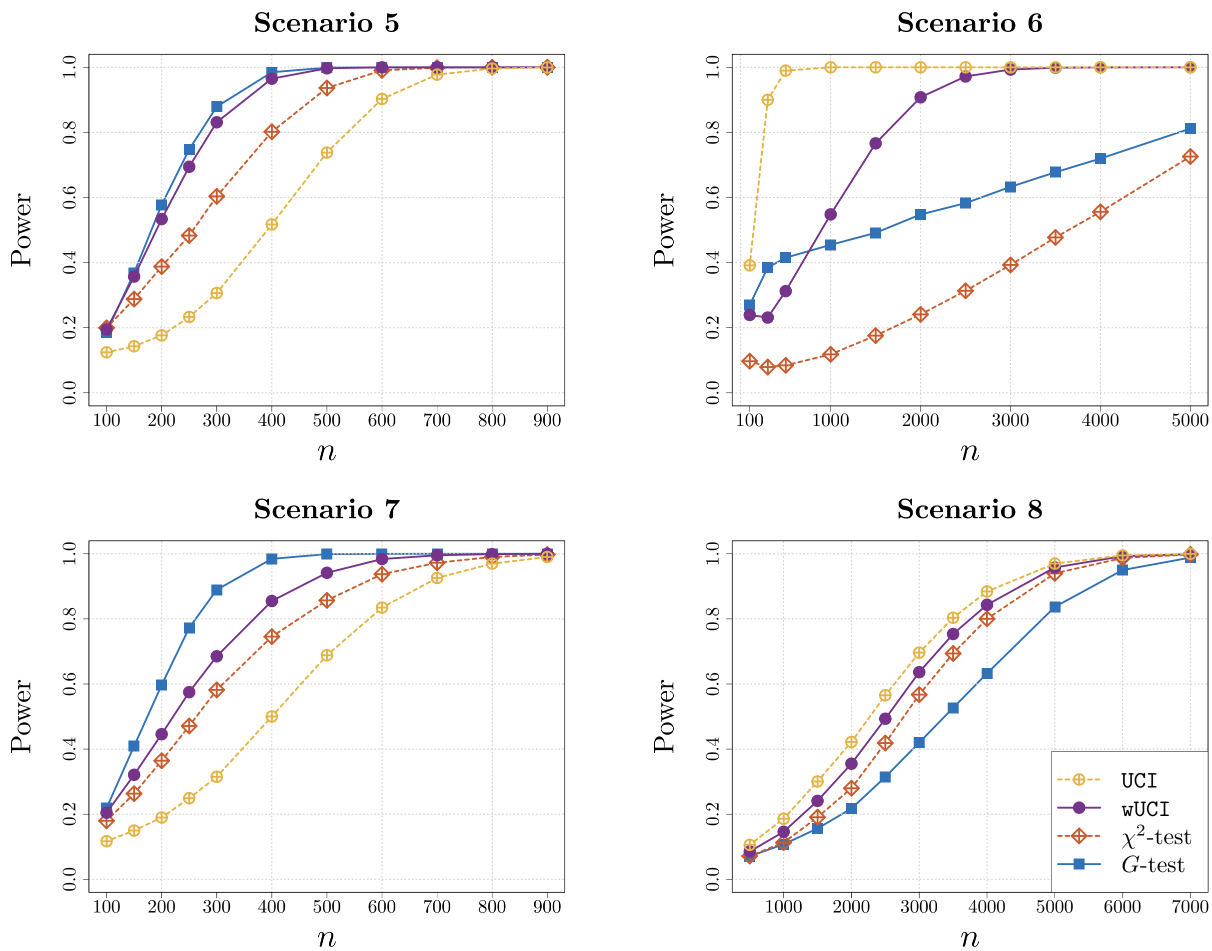

Figure 3: Power comparisons of the considered tests in Scenario 5–Scenario 8 described in Section 5.1.

ones. Nevertheless, our simulation results validate the claim that the proposed tests can work as practical tools that complement classical χ^2-test and G-test under various scenarios.

5.2 Real-world data examples

We next provide numerical illustrations based on real-world datasets.

Admission dataset. The first dataset that we look at is the Berkeley admissions dataset, which is a well-known example for Simpson's paradox (Bickel et al., 1975). As summarized in Table 1, the dataset consists of 4,526 applications with 3 variables (X, Y, Z) where X and Y are binary variables, representing the gender (male or female) and the admission status (admitted or rejected), respectively. The conditional variable Z takes the department name among {A, B, C, D, E, F}. When the dataset is aggregated over the departments, it appears that male applicants are more likely to be admitted than woman applicants. However, as reported by Bickel et al. (1975), there seems to exist a bias in favor of women when looking at the individual departments, indicating the existence of conditional dependence. We assess this claim of conditional dependence by implementing the considered permutation tests. For this dataset, the corresponding p-values are computed as 0.005 for both χ^2- and G-tests, 0.04 for `UCI`-test and 0.03 for `wUCI`-test, respectively. All the p-values are significant at level $\alpha = 0.05$, revealing evidence of conditional dependence.

Table 1: Admissions data at University of California, Berkeley from the six largest departments in 1973

	Men		Women	
Major	Applicants	Admitted	Applicants	Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Diamonds dataset. Next we consider the diamonds dataset available in `R` package `ggplot2`. The dataset contains the information of 53,940 diamonds including their price, clarity, color, quality of the cut, etc. In our analysis, the price variable is partitioned into 100 intervals of equal size. We set the corresponding categorized price variable as X and set the clarity variable as Y. The clarity variable has 8 categories (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF) and it measures the purity of a diamond. The conditional variable Z is chosen to be either the cut variable or the color variable in our analysis. Both variables are discrete with 5 (Fair, Good, Very Good, Premium, Ideal) and 7 (D, E, F, G, H, I, J) categories, respectively. In the experiments, we treat the entire dataset as the population (thereby the ground truth is known to us) from which we randomly draw n observations without replacement. Based on this sample, we compute the permutation p-values of the considered tests and estimate their power via Monte Carlo simulation. The results can be found in Figure 4 where we collect the power of $\{\texttt{UCI}, \texttt{wUCI}, \chi^2, G\}$-tests by changing the sample size n. The left panel of Figure 4 provides the power results when the conditional variable is set to be the color variable. As can be seen, χ^2-test has the significantly lower power than the others. Among the other three tests, `wUCI`-test has the highest power followed by `UCI`-test while the difference is minor. We can see a similar pattern from the right panel of Figure 4 where the conditional variable is set to be the cut variable. These results highlight the practical value of the proposed tests in analyzing real-world datasets where classical tests potentially suffer from low power.

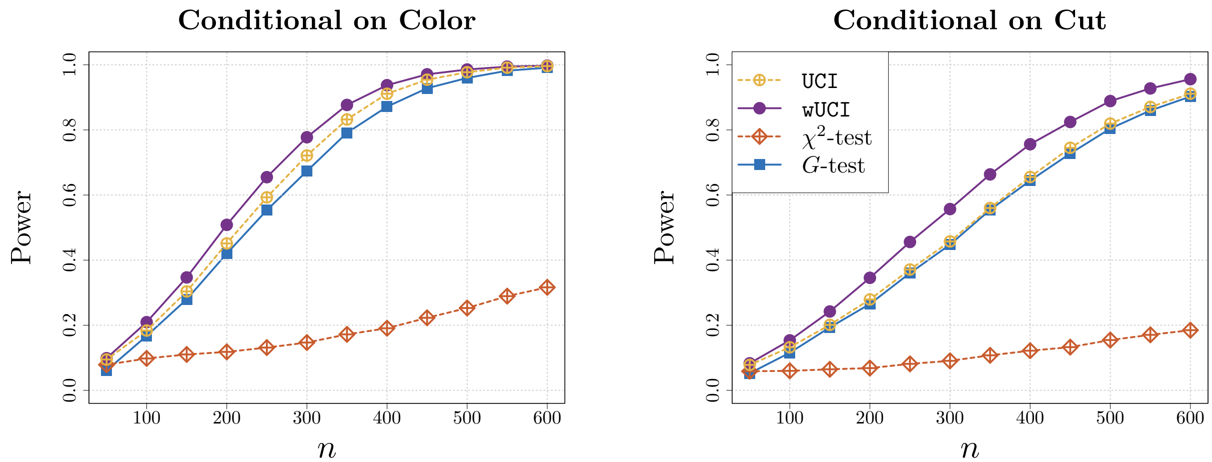

Figure 4: Power comparisons of the considered tests based on the diamonds dataset. Both panels analyze independence between the (categorized) price and clarity variables conditional on the color variable and the cut variable, respectively. All of the tests have increasing power as the sample size increases. Markedly, χ^2-test has significantly lower power than the other tests, whereas `wUCI`-test seems to perform the best in this dataset.

6 Discussion

In this paper, we have revisited recent developments of CI testing for discrete data put forward by the computer science community. Despite attractive theoretical properties, the considered tests in their analysis have limited practical value, relying on Poissonization and unspecified constants in their critical values. In this work, we have made an attempt to bridge the gap between theory and practice by removing Poissonization and utilizing the permutation method to calibrate test statistics. We have also complemented our theoretical results with a thorough numerical analysis and demonstrated certain benefits of the proposed tests over classical χ^2- and G-tests.

Our work leaves several important avenues for future research. One prominent direction is to depoissonize other sample complexity results in the literature using the tools developed in this paper. For instance, one can reproduce the results of Neykov et al. (2021); Kim et al. (2021) for continuous CI testing without Poissonization. Another direction which may be fruitful to pursue is to devise a CI test that further incorporates the prior information about alternative distributions. For example, suppose that we are in an alternative setting where only a handful of conditional bins are significant. In this case, it is possible to obtain a substantial power gain by using sparse weights in the proposed statistics. Additionally, it would be of interest to see whether `wUCI`-test or other tests without sample splitting can achieve the sample complexity (6) in a general regime. We leave these interesting questions for future work.

Acknowledgements.

References

Agresti, A. (1992). A Survey of Exact Inference for Contingency Tables. *Statistical Science*, 7(1):131–153.

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons.

Balakrishnan, S. and Wasserman, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893–1927.

Berrett, T. B., Kontoyiannis, I., and Samworth, R. J. (2021). Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, 49(5):2457–2490.

Berrett, T. B. and Samworth, R. J. (2021). USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A*, 477(2256):20210549.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Canonne, C. L. (2020). A survey on distribution testing: Your data is big. But is it blue? *Theory of Computing*, pages 1–100.

Canonne, C. L. (2022). Topics and techniques in distribution testing. https://ccanonne.github.io/files/misc/main-survey-fnt.pdf.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing Conditional Independence of Discrete Distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 735–748, New York, NY, USA. Association for Computing Machinery.

Chan, S.-O., Diakonikolas, I., Valiant, P., and Valiant, G. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). *Introduction to Algorithms*. MIT press.

De Campos, L. M. and Huete, J. F. (2000). A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning*, 24(1):11–37.

Diakonikolas, I. and Kane, D. M. (2016). A New Approach for Testing Properties of Discrete Distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Haberman, S. J. (1988). A Warning on the Use of Chi-squared Statistics with Frequency Tables with Small Expected Cell Counts. *Journal of the American Statistical Association*, 83(402):555–560.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacquet, P. and Szpankowski, W. (1998). Analytical Depoissonization and its Applications. *Theoretical Computer Science*, 201(1-2):1–62.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295.

Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Sciences*, 35(5):252–257.

Kim, I. (2020). Multinomial goodness-of-fit based on U-statistics: High-dimensional asymptotic and minimax optimality. *Journal of Statistical Planning and Inference*, 205:74–91.

Kim, I. (2022). Comments on "Testing Conditional Independence of Discrete Distributions". *arXiv preprint arXiv:2207.02819*.

Kim, I., Balakrishnan, S., and Wasserman, L. (2022). Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251.

Kim, I., Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Local permutation tests for conditional independence. *arXiv preprint arXiv:2112.11666*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Li, C. and Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489.

McDonald, J. H. (2014). G-test of goodness-of-fit. *Handbook of Biological Statistics*, 487:53–58.

Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177.

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Penrose, M. (2003). *Random Geometric Graphs*, volume 5. OUP Oxford.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13(5):1393–1434.

Spohn, W. (1994). On the Properties of Conditional Independence. In *Patrick suppes: Scientific philosopher*, pages 173–196. Springer.

Tsamardinos, I. and Borboudakis, G. (2010). Permutation Testing Improves Bayesian Network Learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 322–337. Springer.

Vaart, A. W. and Wellner, J. A. (1996). Weak Convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Valiant, G. and Valiant, P. (2017). An Automatic Inequality Prover and Instance Optimal Identity Testing. *SIAM Journal on Computing*, 46(1):429–455.

Valiant, P. (2011). Testing Symmetric Properties of Distributions. *SIAM Journal on Computing*, 40(6):1927–1968.

Yao, Q. and Tritchler, D. (1993). An exact analysis of conditional independence in several 2×2 contingency tables. *Biometrics*, pages 233–236.

Zhang, J. and Zhang, Z. (2022). A Normal Test for Independence via Generalized Mutual Information. *arXiv preprint arXiv:2207.09541*.

A Algorithm

The below is the algorithm for `wUCI`-test used in our simulation study.

Algorithm 4 `wUCI`: weighted U-statistic permutation CI test

Input: Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α

For $j \in [B]$ **do**

  For $m \in [d]$ **do**

  - Generate $\pi \sim \text{Uniform}(\Pi_{\sigma_m})$ independent of everything else.
  - Compute $U_W^{\boldsymbol{b}}(D_m^\pi)$ as in (4) based on the permuted dataset D_m^π.

  End

  Set $T_{W,j}^\dagger \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U_W^{\boldsymbol{b}}(D_m^\pi)$.

End

- Set $T_W^\dagger \leftarrow \sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U_W^{\boldsymbol{b}}(D_m)$ computed without permutations.
- Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_{W,j}^\dagger \geq T_W^\dagger) + 1\right].$$

Output: Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

B Lemmas

In this section, we prove several lemmas in Canonne et al. (2018) under multinomial sampling. These results are main building blocks that lead to Theorem 1 and Theorem 2. We start by showing that the result in Lemma 3.1 of Canonne et al. (2018) holds without Poissonization.

Lemma 1. *Suppose that* $X \sim \text{Binomial}(n, p)$ *where* $n \geq 4$*. Then there exists an absolute constant* $\gamma > 0.85$ *such that*

$$\mathbb{E}[X\mathbb{1}(X \geq 4)] \geq \gamma \min\{np, (np)^4\}.$$

The second lemma that we prove corresponds to Claim 2.1 of Canonne et al. (2018).

Lemma 2. *Suppose that* $X \sim \text{Binomial}(n, p)$ *where* $n \geq 4$*. Then there exists a constant* $\gamma > 0$ *such* *that*

$$\mathrm{Var}[X\mathbb{1}(X \geq 4)] \leq \gamma \mathbb{E}[X\mathbb{1}(X \geq 4)].$$

The following lemma proves Claim 2.2 of Canonne et al. (2018) for a binomial random variable with parameters n, p. It is worth pointing out that the the original statement of Claim 2.2 in Canonne et al. (2018) contains an error, which has been corrected by Kim (2022).

Lemma 3. *Suppose that* $X \sim \mathrm{Binomial}(n,p)$ *where* $n \geq 9$ *and let* $a, b \geq 2$. *Then there exists a constant* $C > 0$ *such that*

$$\begin{aligned}&\mathrm{Var}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big] \\ &\leq C\min\{np, \sqrt{ab}\}\mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big].\end{aligned}$$

The following lemma proves Claim 2.3 of Canonne et al. (2018) for a binomial random variable with parameters n, p.

Lemma 4. *Suppose that* $X \sim \mathrm{Binomial}(n,p)$ *where* $n \geq 9$. *Then there exists a constant* $C > 0$ *such that for any* $a, b \geq 2$ *and* $\lambda = np$,

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

The next lemma builds on the negative association of a multinomial random vector, which is proved in Appendix C.9.

Lemma 5. *Suppose that* $\boldsymbol{\sigma} := \{\sigma_1, \ldots, \sigma_d\}$ *has a multinomial distribution with parameters* n *and* $(p_1, \ldots, p_d)$. *Then for any* $1 \leq i \neq j \leq d$

$$\mathrm{Cov}\big\{\sigma_i\mathbb{1}(\sigma_i \geq 4), \sigma_j\mathbb{1}(\sigma_j \geq 4)\big\} \leq 0.$$

Moreover, let $\omega_i = \sqrt{\min\{\sigma_i, \ell_1\}\min\{\sigma_i, \ell_2\}}$ *for* $\ell_1, \ell_2 \geq 0$. *Then for any* $1 \leq i \neq j \leq d$

$$\mathrm{Cov}\big\{\sigma_i\mathbb{1}(\sigma_i \geq 4)\omega_i, \sigma_j\mathbb{1}(\sigma_j \geq 4)\omega_j\big\} \leq 0.$$

As shown in Lemma 6 of Kim et al. (2021), rejecting the null when the permutation p-value is less than or equal to α is equivalent to rejecting the null when the corresponding test statistic is greater than the $1-\alpha$ quantile of the permutation distribution. We state this result below for completeness, and its proof can be found in Appendix C.10.

Lemma 6 (Quantile)**.** *Let* $q_{1-\alpha}$ *be the* $1-\alpha$ *quantile of the empirical distribution of* $V, V_1, \ldots, V_B$. *Then it holds that*

$$\mathbb{1}\left(\frac{1}{B+1}\left[\sum_{i=1}^{B}\mathbb{1}(V_i \geq V) + 1\right] > \alpha\right) = \mathbb{1}(V \leq q_{1-\alpha}).$$

Finally, we state Efron–Stein inequality for completeness (e.g. Theorem 3.1 of Boucheron et al., 2013).

Lemma 7 (Efron–Stein inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X)$ be a square-integrable function of $X = (X_1, \ldots, X_n)$. Moreover, $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$ and define*

$$Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).$$

Then

$$\mathrm{Var}[Z] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\big[(Z - Z_i')^2\big].$$

C Proofs

This section collects the proof of the main theorems and lemmas.

C.1 Proof of Theorem 1

Given the lemmas in Appendix B together with the results in Canonne et al. (2018), the claims in Theorem 1 are almost immediate. We present more details below. Since the permutation test controls the type I error in any finite sample sizes, we only focus on the alternative case where the underlying distribution is ε-far from the null in L_1 distance, i.e. $\inf_{q \in \mathcal{P}_0} \|p - q\|_1 \geq \varepsilon$, for both T and T_W.

C.1.1 Claim for T

Starting with the first claim, we analyze the expectation and the variance of T under multinomial sampling. Conditional on σ_m, $U(D_m)$ is an unbiased estimator of $\delta_m^2 := \|p_{X,Y\,|\,Z=m} - p_{X|Z=m}p_{Y|Z=m}\|_2^2$. Moreover, since $\sigma_m \sim \mathrm{Binomial}\big(n, p_Z(m)\big)$, Lemma 1 proves that

$$\begin{aligned}
\mathbb{E}[T] &= \sum_{m \in [d]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m U(D_m)\big] \\
&= \sum_{m \in [d]} \mathbb{E}[\mathbb{1}(\sigma_m \geq 4)\sigma_m] \cdot \delta_m^2 \\
&\gtrsim \sum_{m \in [d]} \min\big\{np_Z(m), \big(np_Z(m)\big)^4\big\} \cdot \delta_m^2.
\end{aligned}$$

Set the sample size $n = \beta' \max\big(\sqrt{d}/\varepsilon'^2, \min(d^{7/8}/\varepsilon', d^{6/7}\varepsilon'^{8/7})\big)$ where $\varepsilon' := \varepsilon/\sqrt{\ell_1\ell_2}$ and $\beta' \geq 1$ is a sufficiently large constant. Using the inequality, we follow the same lines of the proof of Proposition

3.1 in Canonne et al. (2018) and show that

$$\mathbb{E}[T] \ \gtrsim \ \beta' \sqrt{\min(n,d)}.$$

Turning to the variance of T, the law of total variance yields

$$\mathrm{Var}[T] \ = \ \mathbb{E}\big[\mathrm{Var}[T \,|\, \boldsymbol{\sigma}]\big] + \mathrm{Var}\big[\mathbb{E}[T \,|\, \boldsymbol{\sigma}]\big].$$

The analysis of the first term is essentially the same as Proposition 3.2 of Canonne et al. (2018). First note that conditional on $\boldsymbol{\sigma}$, $D_1, \ldots, D_d$ are mutually independent (given $\sigma_1, \ldots, \sigma_d$, we independently generate $D_1, \ldots, D_d$), and therefore the following identity holds

$$\mathrm{Var}[T \,|\, \boldsymbol{\sigma}] \ = \ \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \mathrm{Var}[U(D_m) \,|\, \sigma_m].$$

We then use the variance bound for $U(D_m)$ in Equation (4) of Canonne et al. (2018) and see

$$\begin{aligned}
\mathrm{Var}[T \,|\, \boldsymbol{\sigma}] \ &\lesssim \ \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \bigg(\frac{\|p_{X,Y \,|\, Z=m} - p_{X|Z=m}p_{Y|Z=m}\|_2^2}{\sigma_m} + \frac{1}{\sigma_m^2} \bigg) \\
&\lesssim \ \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \|p_{X,Y \,|\, Z=m} - p_{X|Z=m}p_{Y|Z=m}\|_2^2 + \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4) \\
&\lesssim \ \mathbb{E}[T \,|\, \boldsymbol{\sigma}] + \min(n,d).
\end{aligned}$$

This reveals that $\mathbb{E}\big[\mathrm{Var}[T \,|\, \boldsymbol{\sigma}]\big] \lesssim \mathbb{E}[T] + \min(n,d)$. For the second term, the analysis is again similar to that in Proposition 3.2 of Canonne et al. (2018). The only difference is that $\sigma_1, \ldots, \sigma_d$ are no longer independent but negatively associated. In Canonne et al. (2018), the independence between $\sigma_1, \ldots, \sigma_d$ is used to show that

$$\mathrm{Var}\big[\mathbb{E}[T \,|\, \boldsymbol{\sigma}]\big] = \mathrm{Var}\Bigg[\sum_{m \in [d]} \sigma_m \mathbb{1}(\sigma_m \geq 4)\delta_m^2\Bigg] = \sum_{m \in [d]} \delta_m^4 \mathrm{Var}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big].$$

In our case, we make use of Lemma 5 (negative association) and prove that

$$\begin{aligned}
&\mathrm{Var}\Bigg[\sum_{m \in [d]} \sigma_m \mathbb{1}(\sigma_m \geq 4)\delta_m^2\Bigg] \\
= \ &\sum_{m \in [d]} \delta_m^4 \mathrm{Var}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big] + \sum_{\substack{m,m' \in [d] \\ m \neq m'}} \mathrm{Cov}\big\{\sigma_m \mathbb{1}(\sigma_m \geq 4), \sigma_{m'} \mathbb{1}(\sigma_{m'} \geq 4)\big\}
\end{aligned}$$

$$\leq \sum_{m\in[d]} \delta_m^4 \mathrm{Var}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big].$$

This does not change the overall conclusion as we are only interested in bounding the variance (up to constants). Additionally, Lemma 2 yields that $\mathrm{Var}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big] \lesssim \mathbb{E}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big]$ and thus

$$\mathrm{Var}\big[\mathbb{E}[T \mid \boldsymbol{\sigma}]\big] \lesssim \sum_{m\in[d]} \delta_m^4 \mathbb{E}\big[\sigma_m \mathbb{1}(\sigma_m \geq 4)\big] \lesssim \mathbb{E}[T].$$

Combining the previous bounds, we have $\mathrm{Var}[T] \lesssim \min(n,d) + \mathbb{E}[T]$. Given the bounds on the expectation and the variance, the claim for T follows by Chebyshev's inequality and adjusting the value of β' and ζ as in Section 3.1.3 of Canonne et al. (2018).

C.1.2 Claim for T_W

Next we prove the sample complexity result for T_W. The type II error result based on T_W under Poisson sampling is given in Lemma 5.6 of Canonne et al. (2018). This lemma builds upon several preliminary results proved under Poisson sampling. Once we depoissonize these preliminary results, then the rest of the proof remains the same. Instead of reproducing the entire proof, we only go over places where Poissonization trick is critical and reprove the corresponding results under multinomial sampling.

- The first place where Poissonization is used is Lemma 5.2 of Canonne et al. (2018). In the process of investigating the expectation of T_W, Canonne et al. (2018) introduce a quantity $D = \sum_{m\in[d]} D_m$ where

  $$D_m = \sigma_m \omega_m \frac{\varepsilon_m^2}{\ell_1 \ell_2} \mathbb{1}(\sigma_m \geq 4) \quad \text{and} \quad \varepsilon_m = \mathrm{TV}\big(p_{X,Y \mid Z=m}, p_{X|Z=m} p_{Y|Z=m}\big).$$

  Then their Lemma 5.2 claims that

  $$\mathbb{E}[D] \ \gtrsim \ \sum_{m\in[d]} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \min(\alpha_m \beta_m, \alpha_m^4),$$

  where $\alpha_m = n p_Z(m)$ and $\beta_m = \sqrt{\min(\alpha_m, \ell_1)\min(\alpha_m, \ell_2)}$. We note that the same result holds under multinomial sampling since

  $$\mathbb{E}[D] \ = \ \sum_{m\in[d]} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \mathbb{E}\big[\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\big] \ \gtrsim \ \sum_{m\in[d]} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \min(\alpha_m \beta_m, \alpha_m^4),$$

  where the inequality follows by Lemma 4.

- The second place where Poissonization is important is Lemma 5.3 of Canonne et al. (2018), which claims that

$$\mathrm{Var}[D] \lesssim \mathbb{E}[D].$$

This result holds under multinomial sampling as well. To see this, let $\varepsilon'_m := \varepsilon_m/\sqrt{\ell_1\ell_2}$ and observe

$$\begin{aligned}
\mathrm{Var}[D] &= \mathrm{Var}\Bigg[\sum_{m\in[d]} \sigma_m\omega_m\varepsilon_m'^2\mathbb{1}(\sigma_m \geq 4)\Bigg] \\
&= \sum_{m\in[d]} \mathrm{Var}\big[\sigma_m\omega_m\varepsilon_m'^2\mathbb{1}(\sigma_m \geq 4)\big] + \sum_{\substack{m,m'\in[d]\\ m\neq m'}} \mathrm{Cov}\big\{\sigma_m\omega_m\mathbb{1}(\sigma_m \geq 4), \sigma_{m'}\omega_{m'}\mathbb{1}(\sigma_{m'} \geq 4)\big\} \\
&\leq \sum_{m\in[d]} \mathrm{Var}\big[\sigma_m\omega_m\varepsilon_m'^2\mathbb{1}(\sigma_m \geq 4)\big],
\end{aligned}$$

where the inequality uses the negative association result established in Lemma 5. We then use Lemma 3 and see that

$$\begin{aligned}
\sum_{m\in[d]} \mathrm{Var}\big[\sigma_m\omega_m\varepsilon_m'^2\mathbb{1}(\sigma_m \geq 4)\big] &\lesssim \sum_{m\in[d]} \varepsilon_m'^4\sqrt{\ell_1\ell_2}\mathbb{E}\big[\sigma_m\omega_m\mathbb{1}(\sigma_m \geq 4)\big] \\
&\lesssim \sum_{m\in[d]} \varepsilon_m'^2\mathbb{E}\big[\sigma_m\omega_m\mathbb{1}(\sigma_m \geq 4)\big] = \mathbb{E}[D].
\end{aligned}$$

Therefore, Lemma 5.3 of Canonne et al. (2018) holds under multinomial sampling as well.

- Lemma 5.4 of Canonne et al. (2018) provides a lower bound for the (conditional) expected value of the test statistic and an upper bound for the (conditional) variance of the test statistic, which simultaneously hold with high probability. In this result, the authors use a concentration property of $N \sim \mathrm{Poisson}(n)$ to its mean in order to connect $\min(N, d)$ with $\min(n, d)$. This step can be skipped under multinomial sampling as n is fixed in this case.

Other than Lemma 5.2, Lemma 5.3 and Lemma 5.4 of Canonne et al. (2018), the other steps of the proof of Lemma 5.6 in Canonne et al. (2018) are essentially the same under multinomial sampling. Hence we omit the details and finish the proof.

C.2 Proof of Theorem 2

Part 1. We begin with the first part that proves the sample complexity of `UCI`-test in Algorithm 2. By Lemma 6, the type II error of `UCI`-test is equal to

$$\mathbb{P}(p_{\mathrm{perm}} > \alpha) = \mathbb{P}(T \leq q_{1-\alpha}),$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of $T, T_1, \ldots, T_B$. Our main strategy is to show that $q_{1-\alpha}$ is upper bounded by the critical value used in Theorem 1. More specifically, suppose that there exists a sufficiently large $\zeta > 0$ (not necessarily the same ζ in Theorem 1) such that

$$\mathbb{P}\big(q_{1-\alpha} \geq \zeta \sqrt{\min(n,d)}\big) \leq \frac{\beta}{2}. \tag{8}$$

Then by denoting the event of $q_{1-\alpha} < \zeta\sqrt{\min(n,d)}$ by $\mathcal{A}$, the type II error of `UCI`-test is bounded by

$$\begin{aligned}
\mathbb{P}(p_{\mathrm{perm}} > \alpha) \ &\leq \ \mathbb{P}(T \leq q_{1-\alpha}, \, \mathcal{A}) + \mathbb{P}(\mathcal{A}^c) \\
&\leq \ \mathbb{P}\big(T \leq \zeta\sqrt{\min(n,d)}\big) + \frac{\beta}{2} \\
&\leq \ \beta,
\end{aligned}$$

where the last inequality uses the result of Theorem 1. Having this inequality, we only need to prove inequality (8). To this end, we consider the permutation distribution of T based on all possible local permutations. To explain it further, recall that Π_{σ_m} is the set of all permutations of $[\sigma_m]$ for each $m \in [d]$, and D_m^{π} is the set of the data with $Z = m$ whose Y values are permuted based on $\pi \in \Pi_{\sigma_m}$. Let $\pi_{(m)}$ be a random permutation drawn uniformly from Π_{σ_m} for each $m \in [d]$. Further assume that $\pi_{(1)}, \ldots, \pi_{(m)}$ are mutually independent and write $\boldsymbol{\pi} = \{\pi_{(1)}, \ldots, \pi_{(m)}\}$. Let $T^{\boldsymbol{\pi}}$ be the test statistic computed as T but based on $\{D_1^{\pi(1)}, \ldots, D_d^{\pi(d)}\}$. Then the permutation distribution of T based on all possible local permutations is given as

$$F_{\boldsymbol{\pi},T}(x) := \mathbb{E}_{\pi_{(1)},\ldots,\pi_{(d)}}\big[\mathbb{1}\big(T^{\boldsymbol{\pi}} \leq x\big) \,\big|\, \{(X_i, Y_i, Z_i)\}_{i=1}^n\big].$$

In other words, $F_{\boldsymbol{\pi},T}$ is the conditional distribution of T_j in Algorithm 2 given $\{(X_i, Y_i, Z_i)\}_{i=1}^n$. We denote the $1-\alpha$ quantile of $F_{\boldsymbol{\pi},T}$ by $q_{1-\alpha}^{\dagger}$. Unlike $q_{1-\alpha}$, the quantile of $F_{\boldsymbol{\pi},T}$ does not involve randomness from Monte Carlo simulation and thus is relatively easier to handle theoretically. In fact, when B is sufficiently large, $q_{1-\alpha}$ and $q_{1-\alpha}^{\dagger}$ are close to each other by the law of large numbers. For our purpose, it is enough to show that $q_{1-\alpha}^{\dagger}$ is larger than $q_{1-\alpha}$ with high probability. To begin with, note that the Dvoretzky–Kiefer–Wolfowitz inequality guarantees that

$$\sup_{x\in\mathbb{R}} \bigg| F_{\boldsymbol{\pi},T}(x) - \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}\big(T_i \leq x\big) \bigg| \leq \sqrt{\frac{1}{2B}\log\bigg(\frac{4}{\beta}\bigg)}$$

holds with probability at least $1 - \beta/4$. Under this good event, it holds that

$$q_{1-\alpha} \ = \ \inf\bigg\{x \in \mathbb{R} : 1 - \alpha \leq \frac{1}{B+1}\sum_{i=0}^{B} \mathbb{1}\big(T_i \leq x\big)\bigg\}$$

$$
\begin{aligned}
&\le \inf\Big\{x\in\mathbb{R}: 1-\alpha \le \frac{1}{B+1}\sum_{i=1}^{B}\mathbb{1}\big(T_i \le x\big)\Big\} \\
&= \inf\Big\{x\in\mathbb{R}: (1-\alpha)\frac{B+1}{B} \le \frac{1}{B}\sum_{i=1}^{B}\mathbb{1}\big(T_i \le x\big)\Big\} \\
&\le \inf\Big\{x\in\mathbb{R}: (1-\alpha)\frac{B+1}{B} + \sqrt{\frac{1}{2B}\log\Big(\frac{4}{\beta}\Big)} \le F_{\boldsymbol{\pi},T}(x)\Big\} \\
&\le q^{\dagger}_{1-\alpha/2},
\end{aligned}
$$

where we utilize the assumption on $B \ge \max\{4(1-\alpha)\alpha^{-1},\, 8\alpha^{-2}\log(4/\beta)\}$ to ensure that

$$
(1-\alpha)\frac{B+1}{B} + \sqrt{\frac{1}{2B}\log\Big(\frac{4}{\beta}\Big)} \le 1-\frac{\alpha}{2}.
$$

Here constant 2 is chosen for convenience and can be an arbitrary positive number greater than one. This result combined with the union bound yields that

$$
\mathbb{P}\big(q_{1-\alpha} \ge \zeta\sqrt{\min(n,d)}\big) \;\le\; \mathbb{P}\big(q^{\dagger}_{1-\alpha/2} \ge \zeta\sqrt{\min(n,d)}\big) + \frac{\beta}{4}.
$$

Hence condition (8) is ensured if

$$
\mathbb{P}\big(q^{\dagger}_{1-\alpha/2} \ge \zeta\sqrt{\min(n,d)}\big) \le \frac{\beta}{4}. \tag{9}
$$

Indeed, the above inequality is essentially proved in the proof of Theorem 5 in Kim et al. (2021). Strictly speaking, Kim et al. (2021) prove this under Poissonization, but this is not critical in the proof of (9) and the same lines of the proof go through under multinomial sampling. This completes the proof of the first part of Theorem 2.

Part 2. The proof of the second part is similar to the first part. Again, we consider the permutation distribution of T_W where the permutation only applies to $D_{X,Y,1},\ldots,D_{X,Y,d}$. More specifically, the permutation distribution function of T_W at $x\in\mathbb{R}$ is given as the conditional expectation of $\mathbb{1}(T_{j,W}\le x)$ given everything except permutations. We denote the $1-\alpha$ quantile of the resulting permutation distribution as $q^{\dagger}_{1-\alpha,W}$ and the $1-\alpha$ quantile of $T_W, T_{1,W},\ldots,T_{B,W}$ by $q_{1-\alpha,W}$. By following the same logic in the first part of the proof, when $B \ge \max\{4(1-\alpha)\alpha^{-1},\, 8\alpha^{-2}\log(4/\beta)\}$, it can be ensured that $q_{1-\alpha,W} \le q^{\dagger}_{1-\alpha,W}$ with probability at least $1-\beta$. Therefore, the type II error of the permutation test in Algorithm 3 is bounded by

$$
\mathbb{P}(T_W \le q_{1-\alpha,W}) \;\le\; \mathbb{P}(T_W \le q^{\dagger}_{1-\alpha,W}) + \frac{\beta}{4}.
$$

Under Poissonization, Theorem 5 of Kim et al. (2021) shows that

$$\mathbb{P}(T_W \leq q^{\dagger}_{1-\alpha,W}) \ \leq \ \frac{3}{4}\beta.$$

In fact, the same lines of the proof of Theorem 5 in Kim et al. (2021) combined with the results in Appendix C.1.2 prove that the same conclusion holds under multinomial sampling. Therefore we omit the details and finish the proof.

C.3 Proof of Proposition 1

We start by making an important observation that $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ is equivalent to the U-statistic of Hilbert–Schmidt Independence Criterion (Gretton et al., 2005) based on kernels

$$k(x,x') = \sum_{q=1}^{\ell_1} \frac{\mathbb{1}(x=q)\mathbb{1}(x'=q)}{\eta_q} \quad \text{and} \quad l(y,y') = \sum_{r=1}^{\ell_2} \frac{\mathbb{1}(y=r)\mathbb{1}(y'=r)}{\upsilon_r},$$

for $x,x' \in [\ell_1]$ and $y,y' \in [\ell_2]$. Leveraging this observation along with the computational trick introduced in Song et al. (2012), $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ can be computed via convenient matrix operations. Specifically, we let $\mathbf{K},\mathbf{L}$ be $\sigma \times \sigma$ kernel matrices whose entries are $\mathbf{K}_{i,j} = k(X_i,X_j)$ and $\mathbf{L}_{i,j} = l(Y_i,Y_j)$, respectively. Define $\widetilde{\mathbf{K}}$ by letting $\widetilde{\mathbf{K}}_{i,j} = \mathbf{K}_{i,j}$ if $i \neq j$ and $\widetilde{\mathbf{K}}_{i,j} = 0$ if $i = j$. That is, $\widetilde{\mathbf{K}}$ is the same as $\mathbf{K}$ except having zero diagonal components. We similarly define $\widetilde{\mathbf{L}}$ by relating it to $\mathbf{L}$. Denote the one-vector of size σ by $\mathbf{1}$. Then Equation (5) of Song et al. (2012) gives a computationally efficient form of $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D)$ as

$$U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(D) \ = \ \frac{1}{\sigma(\sigma-3)}\left[\mathrm{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}) + \frac{\mathbf{1}^\top\widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\widetilde{\mathbf{L}}\mathbf{1}}{(\sigma-1)(\sigma-2)} - \frac{2}{\sigma-2}\mathbf{1}^\top\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\mathbf{1}\right].$$

It is therefore enough to show that $A_1 = \mathrm{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}})$, $A_2 = \mathbf{1}^\top\widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\widetilde{\mathbf{L}}\mathbf{1}$ and $A_3 = \mathbf{1}^\top\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\mathbf{1}$. We prove these in order.

- For the claim $A_1 = \mathrm{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}})$, we observe that

$$\begin{aligned}
\mathrm{tr}(\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}) \ &= \ \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{K}_{i,j}\mathbf{L}_{i,j}\mathbb{1}(i\neq j) \\
&= \ \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\sum_{q=1}^{\ell_1}\frac{\mathbb{1}(X_i=q)\mathbb{1}(X_j=q)}{\eta_q}\right]\left[\sum_{r=1}^{\ell_2}\frac{\mathbb{1}(Y_i=r)\mathbb{1}(Y_j=r)}{\upsilon_r}\right]\mathbb{1}(i\neq j) \\
&= \ \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathbb{1}(X_i=q)\mathbb{1}(X_j=q)\mathbb{1}(Y_i=r)\mathbb{1}(Y_j=r)}{\eta_q\upsilon_r}\mathbb{1}(i\neq j)
\end{aligned}$$

$$= \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\left[\frac{o_{qr}^2 - o_{qr}}{\eta_q \upsilon_r}\right] = A_1.$$

- To prove the second claim $A_2 = \mathbf{1}^\top \widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\widetilde{\mathbf{L}}\mathbf{1}$, notice that

$$\begin{aligned}\mathbf{1}^\top\widetilde{\mathbf{K}}\mathbf{1} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{q=1}^{\ell_1}\frac{\mathbb{1}(X_i = q)\mathbb{1}(X_j = q)}{\eta_q}\mathbb{1}(i \neq j) \\ &= \sum_{q=1}^{\ell_1}\left[\frac{o_{q+}^2 - o_{q+}}{\eta_q}\right].\end{aligned}$$

Similarly, we have

$$\mathbf{1}^\top\widetilde{\mathbf{L}}\mathbf{1} = \sum_{r=1}^{\ell_2}\left[\frac{o_{+r}^2 - o_{+r}}{\upsilon_r}\right].$$

Combining these two proves the identity $A_2 = \mathbf{1}^\top \widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\widetilde{\mathbf{L}}\mathbf{1}$.

- Finally, the claim $A_3 = \mathbf{1}^\top\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\mathbf{1}$ can be proved by following a series of identities:

$$\begin{aligned}\mathbf{1}^\top\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\mathbf{1} &= \sum_{i=1}^{n}\left[\sum_{j=1}^{n}\mathbf{K}_{i,j}\mathbb{1}(i \neq j)\right]\left[\sum_{j'=1}^{n}\mathbf{L}_{i,j'}\mathbb{1}(i \neq j')\right] \\ &= \sum_{i=1}^{n}\left[\sum_{j=1}^{n}\sum_{q=1}^{\ell_1}\frac{\mathbb{1}(X_i = q)\mathbb{1}(X_j = q)}{\eta_q}\mathbb{1}(i \neq j)\right]\left[\sum_{j'=1}^{n}\sum_{r=1}^{\ell_2}\frac{\mathbb{1}(Y_i = r)\mathbb{1}(Y_{j'} = r)}{\upsilon_r}\mathbb{1}(i \neq j')\right] \\ &= \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\frac{\mathbb{1}(X_i = q)\mathbb{1}(X_j = q)}{\eta_q}\mathbb{1}(i \neq j)\right]\left[\sum_{j'=1}^{n}\frac{\mathbb{1}(Y_i = r)\mathbb{1}(Y_{j'} = r)}{\upsilon_r}\mathbb{1}(i \neq j')\right] \\ &= \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}o_{q+}o_{+r}}{\eta_q\upsilon_r} - \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}o_{q+}}{\eta_q\upsilon_r} - \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}o_{+r}}{\eta_q\upsilon_r} + \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}}{\eta_q\upsilon_r} \\ &= \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}(o_{q+}o_{+r} - o_{q+} - o_{+r} + 1)}{\eta_q\upsilon_r} \\ &= A_3.\end{aligned}$$

This completes the proof of Proposition 1.

C.4 Proof of Proposition 2

Our strategy is to find a sequence of distributions under the alternative for which the asymptotic power of χ^2- and G-tests are zero under the given conditions. This is enough to prove the claim as we are concerned with the worse case power. Let us start by forming a sequence of discrete distribution on $[2] \times [2] \times [d]$.

Worst case scenario. Consider the setting where $p_{X,Y|Z}(1,1 \,|\, z) = p_{X,Y|Z}(2,2 \,|\, z) = 1/2$ and $p_{X,Y|Z}(1,2 \,|\, z) = p_{X,Y|Z}(2,1 \,|\, z) = 0$ for all $z \in [d]$. For each n, set $p_Z(1) = (1-1/n)^{1/n}$ so that the probability of having all of $Z_1, \ldots, Z_n$ equal to one is $1 - 1/n$. We then set $p_Z(2) = \ldots = p_Z(d) = \big(1 - p_Z(1)\big)/(d-1)$. Notice that by construction, p_Z takes non-zero values over $[d]$ and thus the dimension of Z is d. We then let $p = p_{X,Y|Z}p_Z$ and claim that

$$\inf_{q \in \mathcal{P}_0} \|p - q\|_1 \geq \frac{1}{2}. \tag{10}$$

Note that for any $q = q_{X|Z}q_{Y|Z}q_Z \in \mathcal{P}_0$,

$$\begin{aligned} \|p-q\|_1 &= \|p_{X,Y|Z}p_Z - q_{X|Z}q_{Y|Z}q_Z\|_1 \\ &\overset{\text{(i)}}{\geq} \|p_{X,Y|Z}p_Z - q_{X|Z}q_{Y|Z}p_Z\|_1 - \|q_{X|Z}q_{Y|Z}p_Z - q_{X|Z}q_{Y|Z}q_Z\|_1 \\ &= \sum_{z \in [d]} p_Z(z) \cdot \|p_{X,Y|Z=z} - q_{X|Z=z}q_{Y|Z=z}\|_1 - \|p_Z - q_Z\|_1 \\ &\overset{\text{(ii)}}{\geq} \sum_{z \in [d]} p_Z(z) \cdot \|p_{X,Y|Z=z} - q_{X|Z=z}q_{Y|Z=z}\|_1 - \|p-q\|_1, \end{aligned}$$

where step (i) uses the triangle inequality and step (ii) uses the following inequality (which can be proved using the triangle inequality):

$$\|p_Z - q_Z\|_1 = \sum_{z \in [d]} |p_Z(z) - q_Z(z)| = \sum_{z \in [d]} \Bigg| \sum_{x \in [\ell_1], y \in [\ell_2]} p(x,y,z) - q(x,y,z) \Bigg| \leq \|p-q\|_1.$$

Therefore we can conclude that

$$\inf_{q \in \mathcal{P}_0} \|p-q\|_1 \geq \inf_{q \in \mathcal{P}_0} \left\{ \sum_{z \in [d]} p_Z(z) \cdot \|p_{X,Y|Z=z} - q_{X|Z=z}q_{Y|Z=z}\|_1 \right\}$$

In order to further lower bound the above quantity, we first prove a preliminary result. Let $\mathcal{Q}_0$ be the collection of joint distributions of (X,Y) on $[2] \times [2]$ such that for any $q_{X,Y} \in \mathcal{Q}_0$, $q_{X,Y} = q_X q_Y$, i.e. $X \perp\!\!\!\perp Y$. Then for a joint distribution $p_{X,Y}$ such that $p_{X,Y}(1,1) = p_{X,Y}(2,2) = 1/2$ and

$p_{X,Y}(1,2) = p_{X,Y}(2,1) = 0$, we have

$$\inf_{q \in \mathcal{Q}_0} \|p_{X,Y} - q\|_1 \geq \inf_{q \in \mathcal{Q}_0} \|p_{X,Y} - q\|_2 = \frac{1}{2},$$

where the last equality can be shown by solving the following optimization problem:

$$\inf_{a,b \in [0,1]} \sqrt{\left(\frac{1}{2} - ab\right)^2 + a^2(1-b)^2 + (1-a)^2 b^2 + \left(\frac{1}{2} - (1-a)(1-b)\right)^2} = \frac{1}{2}.$$

Consequently, given that $(X,Y) \perp\!\!\!\perp Z$ in our construction,

$$\begin{aligned}\inf_{q \in \mathcal{P}_0} \|p - q\|_1 \ &\geq \ \inf_{q \in \mathcal{P}_0} \left\{ \sum_{z \in [d]} p_Z(z) \cdot \|p_{X,Y|Z=z} - q_{X|Z=z} q_{Y|Z=z}\|_1 \right\} \\ &\geq \ \frac{1}{2},\end{aligned}$$

which proves claim (10).

Main proof. Now let $\mathcal{A}_n$ be an event that all of $Z_1, \ldots, Z_n$ are equal to one, and $\mathcal{A}^c$ be its complement. Starting with χ^2-test, its worst case power is bounded by

$$\begin{aligned}\inf_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\chi^2 > q_{1-\alpha,d}) \ &\leq \ \mathbb{P}_{p^*}(\chi^2 > q_{1-\alpha,d}) \\ &\leq \ \mathbb{P}_{p^*}(\chi^2 > q_{1-\alpha,d} \,|\, \mathcal{A}_n)\mathbb{P}_{p^*}(\mathcal{A}) + \mathbb{P}_{p^*}(\mathcal{A}_n^c) \qquad (11) \\ &\leq \ \mathbb{P}_{p^*}(\chi^2 > q_{1-\alpha,d} \,|\, \mathcal{A}_n) + \frac{1}{n}.\end{aligned}$$

Conditional on $\mathcal{A}_n$, χ^2/n has the same distribution as

$$W_n^\star := \frac{(W - W^2/n)^2}{W^2} + \frac{2\big(W^2(n-W)^2/n^2\big)}{W(n-W)} + \frac{\big(n - W - (n-W)^2/n\big)^2}{(n-W)^2},$$

where $W \sim \text{Binomial}(n, 1/2)$. Using this notation, we have the identity

$$\mathbb{P}_{p^*}(\chi^2 > q_{1-\alpha,d} \,|\, \mathcal{A}_n) \ = \ \mathbb{P}\left(\frac{n}{d} \cdot W_n^\star - 1 > \frac{1}{\sqrt{d}} \cdot \frac{q_{1-\alpha,d} - d}{\sqrt{d}}\right).$$

Under our choice of $d = n \times r_n$ where $r_n \to \infty$, we apply the law of large number and continuous mapping theorem and see that

$$\frac{n}{d} \cdot W_n^\star - 1 \xrightarrow{p} -1,$$

where $X_n \xrightarrow{p} X$ means convergence of X_n to X in probability. In addition, by the central limit theorem along the fact that convergence in distribution implies convergence of the quantile function at all continuity points, we have

$$\frac{q_{1-\alpha} - d}{\sqrt{2d}} \to z_{1-\alpha} \quad \text{as } d \to \infty,$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$. Thus

$$\frac{1}{\sqrt{d}} \cdot \frac{q_{1-\alpha,d} - d}{\sqrt{d}} \to 0 \quad \text{as } n \to \infty.$$

Therefore by Slutsky's theorem, we have

$$\mathbb{P}_{p^*}(\chi^2 > q_{1-\alpha,d} \,|\, \mathcal{A}_n) = \mathbb{P}\biggl(\frac{n}{d} \cdot W_n^\star - 1 > \frac{1}{\sqrt{d}} \cdot \frac{q_{1-\alpha,d} - d}{\sqrt{d}}\biggr) \to 0.$$

This along with the upper bound in (11) implies that

$$\lim_{n\to\infty} \inf_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\chi^2 > q_{1-\alpha,d}) = 0.$$

The proof of G-test is essentially the same. The only difference is that conditional on $\mathcal{A}_n$, G/n has the same distribution as

$$V_n^\star := 2\biggl[\frac{W}{n}\log\biggl(\frac{n}{W}\biggr) + \biggl(1 - \frac{W}{n}\biggr)\log\biggl(\frac{1}{1-\frac{W}{n}}\biggr)\biggr],$$

from which it holds that

$$\mathbb{P}_{p^*}(G > q_{1-\alpha,d} \,|\, \mathcal{A}_n) \;=\; \mathbb{P}\biggl(\frac{n}{d} \cdot V_n^\star - 1 > \frac{1}{\sqrt{d}} \cdot \frac{q_{1-\alpha,d} - d}{\sqrt{d}}\biggr).$$

Again, it can be seen that $\frac{n}{d} \cdot V_n^\star - 1 \xrightarrow{p} -1$ and $\frac{1}{\sqrt{d}} \cdot \frac{q_{1-\alpha,d}-d}{\sqrt{d}} \to 0$. Following the same steps as in the case of χ^2-test, we conclude

$$\lim_{n\to\infty} \inf_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(G > q_{1-\alpha,d}) = 0.$$

This completes the proof of Proposition 2.

C.5 Proof of Lemma 1

We start by computing the expectation of $X\mathbb{1}(X \geq 4)$ explicitly as

$$\mathbb{E}[X\mathbb{1}(X \geq 4)] \;=\; \mathbb{E}[X] - \mathbb{E}[X\mathbb{1}(X \leq 3)]$$

$$= np - np(1-p)^{n-1} - n(n-1)p^2(1-p)^{n-2} - \frac{n(n-1)(n-2)}{2}p^3(1-p)^{n-3}.$$

Given this formula, we consider two cases: (i) $np > 1$ and (ii) $np \leq 1$, separately. Throughout the proof, we assume that $p > 0$ since the claim is trivial when $p = 0$.

Case (i) $np > 1$**.** First assume that $np > 1$, equivalently $p > 1/n$. At a high-level, when np is large, there is only a small possibility that $\mathbb{1}(X < 4) = 1$. In this case, $X\mathbb{1}(X \geq 4)$ behaves like X and we expect that $\mathbb{E}[X\mathbb{1}(X \geq 4)] \approx \mathbb{E}[X] = np$. We now make this statement more precise. To this end, let us define

$$\begin{aligned} r_n(p) &:= \frac{\mathbb{E}[X\mathbb{1}(X \geq 4)]}{np} \\ &= 1 - (1-p)^{n-1} - (n-1)p(1-p)^{n-2} - \frac{(n-1)(n-2)}{2}p^2(1-p)^{n-3}. \end{aligned}$$

We will show that $r_n(p)$ is an increasing function of p for $p > 1/n$ and thus the minimum is achieved at $r_n(n^{-1})$. To show this, we calculate its derivative with respect to p, which can be seen to be non-negative for all $n \geq 4$ as

$$r_n'(p) = \frac{(n-1)(n-2)(n-3)(1-p)^n p^2}{2(1-p)^4} \geq 0 \quad \text{for all } n \geq 4.$$

Therefore, the minimum of $r_n(p)$ is achieved at $p = 1/n$ as a function of p, which yields

$$\begin{aligned} r_n(n^{-1}) &= 1 - \left(1 - \frac{1}{n}\right)^{n-1} - \frac{n-1}{n}\left(1 - \frac{1}{n}\right)^{n-2} - \frac{(n-1)(n-2)}{2n^2}\left(1 - \frac{1}{n}\right)^{n-3} \\ &= \left(1 - \frac{1}{n}\right)^n \left\{\frac{n(5n-6)}{2(n-1)^2}\right\} \\ &\overset{(*)}{\geq} e^{-1}\left(\frac{n-1}{n}\right)\left\{\frac{n(5n-6)}{2(n-1)^2}\right\} = e^{-1}\left\{\frac{5}{2} - \frac{1}{2(n-1)}\right\}, \end{aligned}$$

where step $(*)$ follows by the inequality

$$\left(1 + \frac{x}{n}\right)^n \geq e^x\left(1 - \frac{x^2}{n}\right) \quad \text{for all } n \geq 1 \text{ and } |x| \leq n.$$

From this, we conclude that $r_n(n^{-1}) \geq e^{-1} \times \frac{7}{3} \approx 0.858$ for all $n \geq 4$ and thus

$$\mathbb{E}[X\mathbb{1}(X \geq 4)] \geq \gamma np \quad \text{if } np > 1.$$

Case (ii) $np \leq 1$**.** Next assume $np \leq 1$, i.e. $p \leq 1/n$. Then $\min\{np, (np)^4\} = (np)^4$ and the ratio function $r_n(p)$ is computed as

$$\begin{aligned} r_n(p) &:= \frac{\mathbb{E}[X\mathbb{1}(X \geq 4)]}{n^4 p^4} \\ &= \frac{1}{(np)^3} - \frac{(1-p)^{n-1}}{(np)^3} - \frac{(n-1)}{n^3 p^2}(1-p)^{n-2} - \frac{(n-1)(n-2)}{2n^3 p}(1-p)^{n-3}. \end{aligned}$$

Given this formula, we aim to prove that $r_n(p)$ is a decreasing function of p for $p \leq 1/n$ and hence the minimum is achieved at $r_n(n^{-1}) \geq \gamma$ as shown earlier. To prove this, by assuming $p \neq 0$, it can be seen that the derivative of $r_n(p)$ as a function of p is given by

$$r'_n(p) = \frac{(1-p)^n\{(n-4)p[(n-3)p((n-2)p+3)+6]+6\} - 6(1-p)^4}{2n^3(1-p)^4 p^4}.$$

Our goal is to show that $r'_n(p) \leq 0$ for any $p \in (0, 1/n]$, which is equivalent to the conditions

$$\begin{aligned} & r'_n(p) \leq 0 \\ \overset{\text{iff}}{\Longleftrightarrow}\ & 6(1-p)^n - 6(1-p)^4 + (1-p)^n(n-4)p\{(n^2-5n+6)p^2 + 3(n-3)p + 6\} \leq 0 \\ \overset{\text{iff}}{\Longleftrightarrow}\ & 6 - 6(1-p)^{4-n} + (n-4)\{(n-2)(n-3)p^3 + 3(n-3)p^2 + 6p\} \leq 0. \end{aligned} \tag{12}$$

To this end, we make a key (and nontrivial) observation that the function $h_n(p) := 6(1-p)^{4-n}$ has a Taylor expansion around 0 as

$$\begin{aligned} & h_n(p) \\ = \ & 6 + 6(n-4)p + 3(n-4)(n-3)p^2 + (n-4)(n-3)(n-2)p^3 + \underbrace{\frac{(n-4)(n-3)(n-2)(n-1)}{(1-\xi)^n}p^4}_{\text{remainder}}, \end{aligned}$$

for some $\xi \in (0, 1/n)$. Importantly, the remainder term is strictly positive, which concludes that the inequality (12) holds as well as $r'_n(p) \leq 0$ for all $p \in (0, 1/n]$. This means that $r_n(p)$ is a decreasing function and $r_n(n^{-1}) \leq r_n(p)$ for $p \in (0, 1/n]$. From the previous calculation in case (i), we know that $r_n(n^{-1}) \geq e^{-1} \times \frac{7}{3}$ for all $n \geq 4$ (when $p = 1/n$, we have $np = (np)^4$) and thereby prove the claim.

C.6 Proof of Lemma 2

We divide the proof into two parts depending on whether (i) $np > 1$ or (ii) $np \leq 1$. Throughout, we assume that $p > 0$ since the claim is trivial when $p = 0$.

Case (i) $np > 1$**.** The statement is relatively easier to prove when $np > 1$. Again, the intuition is that when np is large, $X\mathbb{1}(X \geq 4)$ behaves like X and thus $\mathrm{Var}[X\mathbb{1}(X \geq 4)] \approx \mathrm{Var}[X] = np(1-p)$ and $\mathbb{E}[X\mathbb{1}(X \geq 4)] \approx \mathbb{E}[X] = np$. Thus the expectation is larger than the variance. To make this statement more precise, note that the variance can be bounded by

$$\begin{aligned}
\mathrm{Var}[X\mathbb{1}(X \geq 4)] &= \mathbb{E}[X^2] - \mathbb{E}[X^2\mathbb{1}(X \leq 3)] - \{\mathbb{E}[X] - \mathbb{E}[X\mathbb{1}(X \geq 3)]\}^2 \\
&\leq \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 + 2\mathbb{E}[X]\mathbb{E}[X\mathbb{1}(X \geq 3)] \\
&\leq np(1-p) + 6np\mathbb{P}(X \geq 3) \\
&\leq 7np.
\end{aligned}$$

From Lemma 1, we know that $\mathbb{E}[X\mathbb{1}(X \geq 4)] \geq Cnp$ for some $C > 0$ when $np > 1$. Therefore, by taking γ larger than, for example, $7/0.85$, the desired result follows.

Case (ii) $np \leq 1$**.** Now we assume that $np \leq 1$, i.e. $p \leq 1/n$. In this case, we cannot ignore the effect of the indicator function $\mathbb{1}(X \geq 4)$ as in the case of $np > 1$ and thus it requires a more delicate analysis. To simplify the problem a bit, we first upper bound the ratio between the variance and the expectation as

$$\frac{\mathrm{Var}[X\mathbb{1}(X \geq 4)]}{\mathbb{E}[X\mathbb{1}(X \geq 4)]} \lesssim \frac{\mathbb{E}[X^2\mathbb{1}(X \geq 4)]}{(np)^4},$$

where we use the fact that $\mathrm{Var}[X\mathbb{1}(X \geq 4)] \leq \mathbb{E}[X^2\mathbb{1}(X \geq 4)]$ and $\mathbb{E}[X\mathbb{1}(X \geq 4)] \gtrsim (np)^4$ when $np \leq 1$ from Lemma 1. Our goal is to show that

$$r_n(p) := \frac{\mathbb{E}[X^2\mathbb{1}(X \geq 4)]}{(np)^4}$$

is upper bounded by some positive constant for all $n \geq 4$ and $p \leq 1/n$. In particular, we will show that $r_n(p)$ is a decreasing function in p for all $n \geq 4$ and

$$\lim_{p \to 0} r_n(p) = \frac{2(n^3 - 6n^2 + 11n - 6)}{3n^3} < 0.67 \quad \text{for all } n \geq 4.$$

To verify these claims, note that the second moment of $X\mathbb{1}(X \geq 4)$:

$$\begin{aligned}
&\mathbb{E}[X^2\mathbb{1}(X \geq 4)] = \mathbb{E}[X^2] - \mathbb{E}[X^2\mathbb{1}(X \leq 3)] \\
&= np(1-p) + (np)^2 - np(1-p)^{n-1} - 2n(n-1)p^2(1-p)^{n-2} - \frac{3}{2}n(n-1)(n-2)p^3(1-p)^{n-3},
\end{aligned}$$

and therefore $r_n(p)$ can be written as

$$r_n(p) = \frac{1-p}{(np)^3} + \frac{1}{(np)^2} - \frac{(1-p)^{n-1}}{(np)^3} - \frac{2(n-1)(1-p)^{n-2}}{n^3p^2} - \frac{3(n-1)(n-2)(1-p)^{n-3}}{2n^3p}.$$

To show that $r_n(p)$ is a decreasing function in p, consider its derivative with respect to p:

$$\begin{aligned} r_n'(p) \;=\; & \frac{3\,(n-3)\,(n-2)\,(n-1)\,(1-p)^{n-4}}{2n^3 p} + \frac{7\,(n-2)\,(n-1)\,(1-p)^{n-3}}{2n^3p^2} + \frac{5\,(n-1)\,(1-p)^{n-2}}{n^3p^3} \\ & + \frac{3\,(1-p)^{n-1}}{n^3p^4} - \frac{2}{n^2p^3} - \frac{1}{n^3p^3} - \frac{3(1-p)}{n^3p^4}. \end{aligned}$$

We would like to show that $r_n'(p)$ is less than or equal to 0 for all $0 < p \leq 1/n$ and $n \geq 4$, which is equivalent to verifying the following slightly simplified conditions:

$$\begin{aligned} & r_n'(p) \leq 0 \\ \overset{\text{iff}}{\Longleftrightarrow}\quad & \frac{3}{2}\,(n-3)\,(n-2)\,(n-1)\,np^3\,(1-p)^{n-4} + \frac{7}{2}\,(n-2)\,(n-1)\,np^2\,(1-p)^{n-3} \\ & + 5\,(n-1)\,np\,(1-p)^{n-2} + 3n\,(1-p)^{n-1} - 2n^2p - np - 3n(1-p) \;\leq\; 0 \\ \overset{\text{iff}}{\Longleftrightarrow}\quad & 3\,(n-3)\,(n-2)\,(n-1)\,np^3\,(1-p)^{n-4} + 7\,(n-2)\,(n-1)\,np^2\,(1-p)^{n-3} \\ & + 10\,(n-1)\,np\,(1-p)^{n-2} + 6n\,(1-p)^{n-1} - 4n^2p - 2np - 6n(1-p) \;\leq\; 0 \\ \overset{\text{iff}}{\Longleftrightarrow}\quad & 3\,(n-3)\,(n-2)\,(n-1)\,np^4\,(1-p)^{n-4} + 7\,(n-2)\,(n-1)\,np^3\,(1-p)^{n-3} \\ & + 10\,(n-1)\,np^2\,(1-p)^{n-2} + 6np\,(1-p)^{n-1} - 4n^2p^2 - 2np^2 - 6np(1-p) \;\leq\; 0. \end{aligned}$$

Focusing on the last equivalent condition, we first define

$$\begin{aligned} f_n(p) \;:=\; & 3\,(n-3)\,(n-2)\,(n-1)\,np^4\,(1-p)^{n-4} + 7\,(n-2)\,(n-1)\,np^3\,(1-p)^{n-3} \\ & + 10\,(n-1)\,np^2\,(1-p)^{n-2} + 6np\,(1-p)^{n-1}. \end{aligned}$$

Then the last condition simply becomes

$$f_n(p) \leq 6np + 4n(n-1)p^2. \tag{13}$$

Somewhat surprisingly, it can be seen that $6np + 4n(n-1)p^2$ is the second-order Taylor expansion of $f_n(p)$ near $p = 0$. In particular, we have

$$\begin{aligned} f_n(p) \;&=\; f_n(0) + f_n'(0)p + \frac{f_n''(0)}{2}p^2 + R_n \\ &=\; 6np + 4n(n-1)p^2 + R_n, \end{aligned}$$

where

$$R_n = \frac{f_n^{(3)}(\zeta)}{6} p^3 \quad \text{for some } \zeta \in (0, 1/n).$$

Therefore claim (13) follows once we show that $R_n \leq 0$ for all $n \geq 4$ and $p \leq 1/n$. Note that the third derivative of $f_n(p)$ is computed as

$$\begin{aligned} & f_n^{(3)}(p) \\ = \; & \frac{n(n-1)(n-2)(n-3)(n-4)(1-p)^n p^2 \{n(3n-4)p^2 + (35-29n)p + 55\}}{(p-1)^7} \end{aligned}$$

and since the denominator $(p-1)^7$ is negative, it holds that $f_n^{(3)}(p) \leq 0$ once

$$k_n(p) := n(3n-4)p^2 + (35-29n)p + 55$$

is positive for $p \in (0, 1/n]$. Note that $k_n(p)$ is a quadratic function of p and its global minimum is achieved at $p_n^* = \frac{29n-35}{2n(3n-4)}$. Moreover,

$$p_n^* - \frac{1}{n} = \frac{23n-27}{2n(3n-4)} > 0 \quad \text{for } n \geq 4,$$

which means that $k_n(p)$ is decreasing for all $p \leq \frac{1}{n}$. On the other hand, $k_n(n^{-1})$ is computed as

$$k_n(n^{-1}) = 29 + \frac{23}{n} > 0 \quad \text{for all } n \geq 4.$$

Therefore $k_n(p)$ is positive on $p \in (0, 1/n]$ for all $n \geq 4$, which in turn implies $R_n \leq 0$. In conclusion, the claim in (13) holds for all $n \geq 4$ and therefore we complete the proof.

C.7 Proof of Lemma 3

In contrast to a Poisson random variable that is characterized by one single parameter λ, a binomial random variable depends on two parameters n and p. As in the case of Lemma 2, this makes the proof significantly more challenging than the proof for a Poisson random variable. Indeed, we need to take a totally different approach than Canonne et al. (2018) and Kim (2022) to prove the given statement. The key ingredient of our proof is Efron–Stein inequality recalled in Lemma 7.

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Binomial}(1, p)$. Then by letting $X = \sum_{i=1}^n X_i$, we can represent the quantity of interest as a function of i.i.d. random variables as

$$f(X_1, \ldots, X_n) = X \sqrt{\min\{X, a\} \min\{X, b\}} \mathbb{1}(X \geq 4).$$

For simplicity, we denote $f(X_1,\ldots,X_n)$ by f. Then Efron–Stein inequality (Lemma 7) yields that

$$\mathrm{Var}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X\geq 4)\big]\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\big[\big(f-f^i\big)^2\big],$$

where we define $f^i = f(X_1,\ldots,X_{i-1},X_i',X_{i+1},\ldots,X_n)$ and $X_i'\sim \mathrm{Binomial}(1,p)$ independent of $X_1,\ldots,X_n$. By symmetry, it holds that

$$\mathrm{Var}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X\geq 4)\big]\leq \frac{n}{2}\mathbb{E}\big[\big(f-f^1\big)^2\big],$$

and we let $Y=\sum_{i=2}^n X_i$ for simplicity so that $X = X_1 + Y$. By the law of total expectation,

$$\begin{aligned}
n\mathbb{E}\big[\big(f-f^1\big)^2\big] &= n\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\big]\\
&= \underbrace{n\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\mathbb{1}(Y<4)\big]}_{\text{(I)}}+\underbrace{n\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\mathbb{1}(Y\geq 4)\big]}_{\text{(II)}}.
\end{aligned}$$

We analyze two terms (I) and (II) separately.

Analysis of (I). We first claim that

$$\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\mathbb{1}(Y<4)\big]=\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\mathbb{1}(Y=3)\big].$$

To see this, we unpack the expression of f and f^1 as

$$\begin{aligned}
&(f-f^1)^2\mathbb{1}(Y<4)\\
=\;&\Big\{(X_1+Y)\sqrt{\min\{(X_1+Y),a\}\min\{(X_1+Y),b\}}\mathbb{1}(X_1+Y\geq 4)\\
&-(X_1'+Y)\sqrt{\min\{(X_1'+Y),a\}\min\{(X_1'+Y),b\}}\mathbb{1}(X_1'+Y\geq 4)\Big\}^2\mathbb{1}(Y<4).
\end{aligned}$$

When $Y\in\{0,1,2\}$, the indicator functions $\mathbb{1}(X_1+Y\geq 4)$ and $\mathbb{1}(X_1+Y\geq 4)$ are equal to zero since X_1 and X_1' are either zero or one. Therefore, we only need to focus on the case of $Y=3$. Conditional on $Y=3$, we make use of the fact that X_1 and X_1' are independent and show that the expectation of $(f-f^1)^2\mathbb{1}(Y=3)$ is

$$2p(1-p)(Y+1)\sqrt{\min\{(Y+1),a\}\min\{(Y+1),b\}}\mathbb{1}(Y=3).$$

Now by taking the expectation over $Y\sim\mathrm{Binomial}(n-1,p)$,

$$\mathbb{E}_Y\big[\mathbb{E}_{X_1,X_1'}\big[(f-f^1)^2\,|\,Y\big]\mathbb{1}(Y<4)\big] = 8p(1-p)\sqrt{\min\{4,a\}\min\{4,b\}}\binom{n-1}{3}p^3(1-p)^{n-4}.$$

Therefore by using the condition that $n \geq 4$ and the inequality $(1+x) \leq e^x$ for $x \in \mathbb{R}$,

$$n(\mathrm{I}) \lesssim (np)^4(1-p)^{n-3} \lesssim \lambda^4 \exp(-\lambda/4).$$

Moreover, from Lemma 4, we know that

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

for constant $C > 0$. Hence, once we show

$$\lambda^4 \exp(-\lambda/4) \lesssim \min\big\{\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4\big\}, \tag{14}$$

then it implies that

$$n(\mathrm{I}) \lesssim \mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big]. \tag{15}$$

For $\lambda < 1$, the dominating term in the upper bound on inequality (14) is λ^4 and thus inequality (14) follows easily. For $\lambda \geq 1$, by assuming $a \leq b$ without loss of generality, it is enough to show

$$\begin{aligned}
&\lambda^4 \exp(-\lambda/4) \lesssim \lambda^2, \\
&\lambda^4 \exp(-\lambda/4) \lesssim \lambda\sqrt{\lambda}\sqrt{a}, \quad \text{for } \lambda \leq a, \\
&\lambda^4 \exp(-\lambda/4) \lesssim \lambda\sqrt{ab}, \quad \text{for } \lambda \lesssim \min\{a,b\}.
\end{aligned}$$

Starting with the first inequality, the Taylor expansion of $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ shows that $\exp(\lambda/4) \gtrsim \lambda^2$ for all $\lambda \geq 1$. For the second inequality, again, the Taylor expansion of e^x shows that $\exp(\lambda/4) \gtrsim \lambda^2 \gtrsim \lambda^{5/2}/\sqrt{a}$ for $\lambda \leq a$. The last inequality follows similarly. Hence inequality (15) follows as desired.

Analysis of (II). Switching to the second term (II), note that when $Y \geq 4$, it is always the case that $X_1 + Y \geq 4$ and $X_1' + Y \geq 4$. Therefore we can write

$$\begin{aligned}
&(f - f^1)^2\mathbb{1}(Y \geq 4) \\
= \ &\big\{(X_1+Y)\sqrt{\min\{(X_1+Y),a\}\min\{(X_1+Y),b\}} \\
&-(X_1'+Y)\sqrt{\min\{(X_1'+Y),a\}\min\{(X_1'+Y),b\}}\big\}^2\mathbb{1}(Y \geq 4).
\end{aligned}$$

Based on the condition that $X_1, X_1' \overset{\text{i.i.d.}}{\sim} \text{Binomial}(1,p)$, the expectation of $(f - f^1)^2\mathbb{1}(Y \geq 4)$ conditional on Y is computed as

$$\mathbb{E}_{X_1,X_1'}[(f-f^1)^2\mathbb{1}(Y \geq 4)\,|\,Y] = 2p(1-p)\big\{(Y+1)\sqrt{\min\{(Y+1),a\}\min\{(Y+1),b\}}$$

$$
\begin{aligned}
&\qquad - Y\sqrt{\min\{Y,a\}\min\{Y,b\}}\Big\}^2 \mathbb{1}(Y \geq 4) \\
&:= \; 2p(1-p)h(Y,a,b)\mathbb{1}(Y \geq 4).
\end{aligned}
$$

Therefore by taking the expectation over Y, the second term is

$$
n(\mathrm{II}) \;=\; 2np(1-p)\mathbb{E}_Y[h(Y,a,b)\mathbb{1}(Y \geq 4)].
$$

For simplicity, we denote the probability mass function of Binomial$(n-1,p)$ by $q_{n-1,p}(x)$ for $x = 0,1,\ldots,n-1$. Then

$$
np\mathbb{E}_Y[h(Y,a,b)\mathbb{1}(Y \geq 4)] \;=\; np\sum_{k=4}^{n-1} h(k,a,b)q_{n-1,p}(k).
$$

Recall that

$$
h(y,a,b) = \left[(y+1)\sqrt{\min\{y+1,a\}\min\{y+1,b\}} - y\sqrt{\min\{y,a\}\min\{y,b\}}\right]^2.
$$

and, by assuming $a \leq b$ without loss of generality and $y \geq 4$, bound this function under different cases carefully. In the following analysis, we use the fact that for $x, y \geq 4$ and some $\xi \geq 4$,

$$
(y+x)^{3/2} = y^{3/2} + \frac{3\sqrt{y}}{2}x + \frac{3}{8\sqrt{\xi+y}}x^2.
$$

Therefore, it holds that $[(y+1)^{3/2} - y^{3/2}]^2 \lesssim y$ for all $y \geq 4$. In fact, it can be numerically verified that

$$
[(y+1)^{3/2} - y^{3/2}]^2 \leq \underbrace{\frac{(5^{3/2} - 4^{3/2})^2}{4}}_{:=c_0} y \quad \text{for all } y \geq 4.
$$

Based on this observation, we upper bound $h(y,a,b)$ as follows:

- If $a \leq b \leq y$, then

$$
\begin{aligned}
h(y,a,b) = ab \;&\leq\; \min\{y,a\}\min\{y,b\} \\
&\leq\; y\sqrt{\min\{y,a\}\min\{y,b\}}.
\end{aligned}
$$

- If $a \leq y < y+1 = b$, then

$$
\begin{aligned}
h(y,a,b) \;&=\; [(y+1)\sqrt{a(y+1)} - y\sqrt{ay}]^2 = a[(y+1)^{3/2} - y^{3/2}]^2 \leq c_0 ay \\
&\leq\; c_0\min\{y,a\}\min\{y,b\}
\end{aligned}
$$

$$\le c_0 y\sqrt{\min\{y,a\}\min\{y,b\}}.$$

- If $a \le y < y+1 < b$, then

$$\begin{aligned} h(y,a,b) &= [(y+1)\sqrt{a(y+1)} - y\sqrt{ay}]^2 = a[(y+1)^{3/2} - y\sqrt{y}]^2 \le c_0 a y \\ &\le c_0 \min\{y,a\}\min\{y,b\} \\ &\le c_0 y\sqrt{\min\{y,a\}\min\{y,b\}}. \end{aligned}$$

- If $y = a < y+1 = b$,

$$\begin{aligned} h(y,a,b) &= [(y+1)\sqrt{ab} - y\sqrt{ay}]^2 \\ &= [(y+1)\sqrt{y(y+1)} - y^2]^2 \\ &< [(y+1)^2 - y^2]^2 = (2y+1)^2 \le \frac{81}{16}y^2 \quad \text{since } y \ge 4 \\ &= \frac{81}{16}\min\{y,a\}\min\{y,b\} \\ &\le \frac{81}{16}y\sqrt{\min\{y,a\}\min\{y,b\}}. \end{aligned}$$

- If $y = a < y+1 < b$,

$$\begin{aligned} h(y,a,b) &= [(y+1)\sqrt{a(y+1)} - y\sqrt{ay}]^2 \\ &= [(y+1)\sqrt{y(y+1)} - y^2]^2 \\ &< [(y+1)^2 - y^2]^2 = (2y+1)^2 \le \frac{81}{16}y^2 \quad \text{since } y \ge 4. \\ &= \frac{81}{16}\min\{y,a\}\min\{y,b\} \\ &\le \frac{81}{16}y\sqrt{\min\{y,a\}\min\{y,b\}}. \end{aligned}$$

- If $y < a = y+1 = b$,

$$\begin{aligned} h(y,a,b) &= [(y+1)^2 - y^2]^2 = (2y+1)^2 < \frac{81}{16}y^2 \quad \text{since } y \ge 4. \\ &= \frac{81}{16}\min\{y,a\}\min\{y,b\} \\ &\le \frac{81}{16}y\sqrt{\min\{y,a\}\min\{y,b\}}. \end{aligned}$$

- If $y < y+1 < a \le b$,

$$\begin{aligned} h(y,a,b) &= [(y+1)^2 - y^2]^2 = (2y+1)^2 < \frac{81}{16}y^2 \le 9\min\{b^2, y^2\} \quad \text{since } y \ge 4. \\ &= \frac{81}{16}\min\{y,a\}\min\{y,b\} \\ &\le \frac{81}{16}y\sqrt{\min\{y,a\}\min\{y,b\}}. \end{aligned}$$

In summary, for $y \ge 4$, there exists an absolute constant $C > 0$ such that

$$h(y,a,b) \le C\min\{y,a\}\min\{y,b\} \le Cy\sqrt{\min\{y,a\}\min\{y,b\}}.$$

Using this upper bound, we have

$$\begin{aligned} np\mathbb{E}_Y[h(Y,a,b)\mathbb{1}(Y \ge 4)] &= np\sum_{k=4}^{n-1} h(k,a,b)q_{n-1,p}(k) \\ &\overset{\text{(i)}}{\lesssim} np\sum_{k=4}^{n-1}\min\{y,a\}\min\{y,b\}q_{n-1,p}(k) \\ &\overset{\text{(ii)}}{\lesssim} np\sum_{k=4}^{n-1} y\sqrt{\min\{y,a\}\min\{y,b\}}q_{n-1,p}(k). \end{aligned}$$

The last inequality (ii) yields that

$$\begin{aligned} n(\mathrm{II}) &\lesssim np\mathbb{E}[Y\sqrt{\min\{Y,a\}\min\{Y,b\}}\mathbb{1}(Y \ge 4)] \\ &\lesssim np\mathbb{E}[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \ge 4)], \end{aligned} \tag{16}$$

where the second inequality above uses the fact that $Y\sqrt{\min\{Y,a\}\min\{Y,b\}}\mathbb{1}(Y \ge 4) \le (X_1 + Y)\sqrt{\min\{(X_1+Y),a\}\min\{(X_1+Y),b\}}\mathbb{1}(X_1 + Y \ge 4)$. On the other hand, using inequality (i), we can obtain

$$n(\mathrm{II}) \lesssim \lambda\min\{ab, a\lambda, \lambda^2 + \lambda\}. \tag{17}$$

Combining pieces. By the Efron–Stein inequality (Lemma 7), we obtained that

$$\mathrm{Var}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \ge 4)\big] \lesssim n(\mathrm{I}) + n(\mathrm{II}).$$

By focusing on $n(\mathrm{I})$, the bound given in (15) together with Lemma 4 shows that

$$n(\mathrm{I}) \lesssim \min\{np, \sqrt{ab}\}\mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big].$$

Therefore, it suffices to prove

$$n(\mathrm{II}) \lesssim \min\{np, \sqrt{ab}\}\mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big]. \tag{18}$$

Due to the inequality established in (16), we know that

$$n(\mathrm{II}) \lesssim np\mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big] \quad \text{for all } n \geq 9 \text{ and } p \in [0,1].$$

Hence, by assuming $\lambda = np > 1$ (when $\lambda \leq 1$, $\min\{\lambda, \sqrt{ab}\} = \lambda$), we only need to prove that

$$n(\mathrm{II}) \lesssim \sqrt{ab}\mathbb{E}\big[X\sqrt{\min\{X,a\}\min\{X,b\}}\mathbb{1}(X \geq 4)\big],$$

which is implied by

$$\lambda \min\{ab, a\lambda, \lambda^2 + \lambda\} \lesssim \sqrt{ab}\min\big\{\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4\big\},$$

due to Lemma 4 and the previous result (17). Moreover, since we assume $\lambda > 1$, the above condition is satisfied when

$$\min\{(ab)^2, a^2\lambda^2\} \lesssim ab\min(\lambda,a)\min(\lambda,b). \tag{19}$$

Now we see that

- if $\lambda \leq a \leq b$, the sufficient condition (19) becomes

$$a^2\lambda^2 \lesssim ab\lambda^2 \iff a \lesssim b,$$

- if $a \leq \lambda \leq b$, the sufficient condition (19) becomes

$$a^2\lambda^2 \lesssim a^2 b\lambda \iff \lambda \lesssim b,$$

- if $a \leq b \leq \lambda$, the sufficient condition (19) becomes

$$(ab)^2 \lesssim (ab)^2.$$

Therefore we verify the sufficient condition (19) and thus complete the proof.

C.8 Proof of Lemma 4

The proof of this lemma essentially follows the same lines of the proof of Claim 2.3 in Canonne et al. (2018). There are only two conditions that we need to check for a binomial random variable.

Condition 1. Suppose that $\lambda < 8$. For a Poisson random variable $Y \sim \text{Poisson}(\lambda)$, Canonne et al. (2018) shows that

$$\mathbb{E}\big[Y\sqrt{\min\{Y,a\}\min\{Y,b\}}\mathbb{1}(Y \geq 4)\big] \geq \frac{1}{768}\min\big\{\lambda\sqrt{\min(\lambda,a)\min(\lambda,b)}, \lambda^4\big\}. \tag{20}$$

The main step for the above result is $\mathbb{P}(Y = 4) \geq \lambda^4/4!$. We claim that the same bound holds for $X \sim \text{Binomial}(n,p)$ where $n \geq 9$ up to some constant factor. Indeed, we can write

$$\begin{aligned}
\mathbb{P}(X = 4) &= \binom{n}{4}p^4(1-p)^{n-4} \\
&\geq \frac{14}{729}\lambda^4\left(1 - \frac{\lambda}{n}\right)^{n-4} \\
&\geq \frac{14}{729}\left(1 - \frac{8}{9}\right)^5\lambda^4,
\end{aligned}$$

where the inequalities use the fact that $n \geq 9$ and $\lambda < 8$. We then follow the same steps in Canonne et al. (2018) and conclude that inequality (20) holds when Y is replaced by X up to some constant factor.

Condition 2. The second condition Canonne et al. (2018) makes use of is that for $Y \sim \text{Poisson}(\lambda)$ and $\lambda \geq 8$, $\mathbb{P}(Y \geq \lfloor \lambda/2 \rfloor) \geq 1/2$. We now prove that the same inequality holds for $X \sim \text{Binomial}(n,p)$. First assume that $\lfloor \lambda/2 \rfloor \neq 0$. Otherwise, the claim is trivial. By Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}(X < \lfloor \lambda/2 \rfloor) &= \mathbb{P}(-X + \lambda > \lambda - \lfloor \lambda/2 \rfloor) \\
&\leq \frac{\text{Var}[X]}{(\lambda - \lfloor \lambda/2 \rfloor)^2} \leq \frac{4\lambda}{\lambda^2} = \frac{4}{\lambda}.
\end{aligned}$$

Now $\mathbb{P}(X \geq \lfloor \lambda/2 \rfloor) \geq 1/2$ holds since we assume $\lambda \geq 8$. Given this probability bound, we can follow the same lines of the proof in Canonne et al. (2018) and show that

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

for some constant $C > 0$ and $\lambda \geq 8$. This completes the proof.

C.9 Proof of Lemma 5

It is well-known that $\boldsymbol{\sigma}$ is negatively associated (e.g. Joag-Dev and Proschan, 1983). Therefore, by the definition of negative association, for every pair of disjoint subsets A_1, A_2 of $\{1, \ldots, d\}$,

$$\mathrm{Cov}\{f_1(\sigma_i, i \in A_1), f_2(\sigma_j, j \in A_2)\} \leq 0,$$

whenever f_1 and f_2 are increasing. Now the result follows by taking $A_1 = \{i\}$, $A_2 = \{j\}$, and $f_1(x) = f_2(x) = x\mathbb{1}(x \geq 4)$. The second statement follows by the same reasoning.

C.10 Proof of Lemma 6

By the definition of the quantile function and letting $V_0 := V$, $q_{1-\alpha}$ can be written as

$$\begin{aligned} q_{1-\alpha} &= \inf\left\{x \in \mathbb{R} : 1 - \alpha \leq \frac{1}{B+1} \sum_{i=0}^{B} \mathbb{1}\big(V_i \leq x\big)\right\} \\ &= V_{(k)}, \end{aligned}$$

where $V_{(k)}$ is the $k = \lceil (1-\alpha)(B+1) \rceil$th order statistic of $\{V_0, V_1, \ldots, V_B\}$. Given this representation, we see that $V \leq q_{1-\alpha}$ holds if and only if

$$\frac{1}{B+1} \sum_{i=0}^{B} \mathbb{1}\big(V_i < V\big) < 1 - \alpha,$$

which is equivalent to

$$\frac{1}{B+1} \left[\sum_{i=1}^{B} \mathbb{1}(V_i \geq V) + 1\right] > \alpha.$$

This completes the proof.