arXiv:2112.11666v2 [math.ST] 6 Jan 2022

Local permutation tests for conditional independence

Ilmun Kim[‡] Matey Neykov[†] Sivaraman Balakrishnan[†] Larry Wasserman[†]

[†]Department of Statistics and Data Science, Carnegie Mellon University
[‡]Department of Statistics and Data Science, Yonsei University

January 7, 2022

Abstract

In this paper, we investigate local permutation tests for testing conditional independence between two random vectors X and Y given Z. The local permutation test determines the significance of a test statistic by locally shuffling samples which share similar values of the conditioning variables Z, and it forms a natural extension of the usual permutation approach for unconditional independence testing. Despite its simplicity and empirical support, the theoretical underpinnings of the local permutation test remain unclear. Motivated by this gap, this paper aims to establish theoretical foundations of local permutation tests with a particular focus on binning-based statistics. We start by revisiting the hardness of conditional independence testing and provide an upper bound for the power of any valid conditional independence test, which holds when the probability of observing "collisions" in Z is small. This negative result naturally motivates us to impose additional restrictions on the possible distributions under the null and alternate. To this end, we focus our attention on certain classes of smooth distributions and identify provably tight conditions under which the local permutation method is universally valid, i.e. it is valid when applied to any (binning-based) test statistic. To complement this result on type I error control, we also show that in some cases, a binning-based statistic calibrated via the local permutation method can achieve minimax optimal power. We also introduce a double-binning permutation strategy, which yields a valid test over less smooth null distributions than the typical single-binning method without compromising much power. Finally, we present simulation results to support our theoretical findings.

1 Introduction

Conditional independence (CI) is an important concept in a variety of statistical applications including graphical models (De Campos and Huete, 2000; Koller and Friedman, 2009) and causal inference (Imbens and Rubin, 2015; Pearl, 2014; Spohn, 1994). In these applications, the assumption of conditional independence offers significant representational and computational benefits, and helps disentangle causal relationships among variables in an efficient and tractable way. In a related vein, a problem of essential importance in statistical practice is that of variable selection (Dai et al., 2021; Williamson et al., 2021), which is concerned with selecting a parsimonious subset of features that are predictive of a response variable. In each of these settings, conditional independence tests are an essential tool to validate (or invalidate) critical modeling assumptions, and can lend additional credibility to the conclusions of our data analysis.

The performance of a statistical hypothesis test relies not only on the form of the test statistic but also heavily on the method used to ensure type I error control. Indeed, one might argue that

a huge part of the practical success and ubiquity of two-sample and (unconditional) independence tests is the fact that these tests can be tightly calibrated in a black-box fashion using a permutation method. This in turn frees the practitioner to focus on designing powerful test statistics, without having to further ensure that the distribution of their test statistics are analytically tractable under the null. For two-sample and (unconditional) independence testing, the permutation method is universal without any additional restrictions, i.e. it controls the type I error in a non-trivial sense for *any* underlying test statistic. As noted by Shah and Peters (2020), part of the hardness of conditional independence testing with a continuous conditioning variable Z stems from the fact that it is impossible to control the type I error, via for instance a permutation method, in any non-trivial sense without additional restrictions. Our broad goal in this paper is to propose and study natural extensions of the permutation method, namely the local permutation procedure, which are applicable to CI testing. In particular, we aim to investigate restrictions under which these extensions tightly control the type I error for a broad class of test statistics, and further to explore the power of tests calibrated via these methods.

The *local* permutation procedure calibrates a test statistic by locally shuffling samples based on the proximity of their conditioning variables Z. When the conditional variable is discrete, the resulting local permutation test has a universal guarantee on type I error control under relatively weak assumptions on the data-generating process. When the conditional variable is continuous, on the other hand, the validity of the local permutation test is far from obvious. While there is a line of work providing some empirical support (Doran et al., 2014; Fukumizu et al., 2008; Neykov et al., 2021; Sen et al., 2018), a rigorous theoretical foundation of the local permutation test has not been fully established in the continuous case. Motivated by this gap, the first aim of this paper is to identify provably tight conditions under which the type I error of the local permutation test is uniformly controlled at least for sufficiently large sample-sizes. To this end, we focus primarily on a binning-based local permutation procedure and determine the size of bins for which the resulting test is asymptotically valid under various smoothness assumptions.

Once the type I error is under control, our subsequent focus is on power. In contrast to type I error control, which requires the size of bins to be small, the use of bins that are too fine causes a loss of power due to the small sample size in each bin. Our next goal is to balance this trade-off and show that there is a choice of bin-widths which ensures that the local permutation method controls the type I error but still retains minimax optimal power. We achieve this goal by building on the recent work of Canonne et al. (2018) and Neykov et al. (2021) which study minimax-optimal CI tests and the work of Kim et al. (2020) which studies the power of the classical permutation method for two-sample and independence testing.

An interesting aspect of our results is that they elucidate a tension in conditional independence testing between ensuring tight control of the type I error, and ensuring high power of the resulting test. In many well-studied examples, permutation and other simulation methods represent an apparent free lunch, ensuring tight control of the type I error without sacrificing power (see for instance, Kim et al. (2020) for precise results on the minimax power of the permutation test in these settings). In conditional independence testing, the permutation method is no longer exact and we show that there is a trade-off when using the local permutation method for calibration. In certain cases, ensuring type I error control requires selecting bin-widths which are too small to guarantee high power. In some settings, we are able to mitigate this trade-off by designing a careful double-binning strategy where two resolutions are combined in the permutation method:

a finer resolution for permutations which ensures type I error control, and a coarser resolution for computing the test statistic which ensures high power (see Section 6). Before we state our contributions in more detail, we briefly review related work.

1.1 Related work

There is an extensive body of literature on CI measures and CI tests. Here, we give a selective review of existing methods, which can be categorized into several groups.

The first category of methods is based on kernel mean embeddings (see Muandet et al., 2017, for a review). The idea of kernel mean embeddings is to represent probability distributions as elements of a reproducing kernel Hilbert space (RKHS), which enables us to understand properties of these distributions using Hilbert space operations. One of the initial attempts to use kernel mean embeddings for CI testing was made by Fukumizu et al. (2008). In particular, Fukumizu et al. (2008) propose a test based on the empirical Hilbert–Schmidt norm of the conditional cross-covariance operator. Zhang et al. (2012) introduce another kernel-based test attempting to measure partial correlations, which in turn characterize CI (Daudin, 1980). Strobl et al. (2019) use random Fourier features to approximate kernel computations, and propose a more computationally efficient version of the test of Zhang et al. (2012). Other CI measures proposed by Doran et al. (2014) and Huang et al. (2020) are motivated by the kernel maximum mean discrepancy for two-sample testing (Gretton et al., 2012). In particular, the CI measure introduced by Huang et al. (2020) compares whether $Y|X, Z$ and $Y|X$ have the same distribution, and their measure can be viewed as a kernelized version of the CI measure of Azadkia and Chatterjee (2019). Recently, Sheng and Sriperumbudur (2019) and Park and Muandet (2020) propose kernel CI measures that are closely connected to the Hillbert–Schmidt independence criterion (Gretton et al., 2005). Sheng and Sriperumbudur (2019) also discuss the connection between their CI measure to the conditional distance correlation proposed by Wang et al. (2015).

Another category of methods relies on estimating regression functions. Consider random variables X and Y, and their regression residuals on Z, denoted by $\epsilon_{X,Z} := X - \mathbb{E}[X|Z]$ and $\epsilon_{Y,Z} := Y - \mathbb{E}[Y|Z]$. The underlying idea of regression-based methods is that the expected value of $\epsilon_{X,Z}\epsilon_{Y,Z}$ is zero if $X \perp\!\!\!\perp Y|Z$, and not necessarily zero if $X \not\perp\!\!\!\perp Y|Z$. Thus, one can use an empirical estimate of the expected value of $\epsilon_{X,Z}\epsilon_{Y,Z}$ as a test statistic for CI. We refer to Zhang et al. (2018) for a discussion of the relationship between $\epsilon_{X,Z} \perp\!\!\!\perp \epsilon_{Y,Z}$ and $X \perp\!\!\!\perp Y|Z$. Given that there exist a variety of successful regression algorithms to estimate $\mathbb{E}[X|Z]$ and $\mathbb{E}[Y|Z]$, the expected value of $\epsilon_{X,Z}\epsilon_{Y,Z}$ can be accurately estimated as well. This regression-based idea has been exploited by several authors to tackle CI testing. For instance, Shah and Peters (2020) propose the generalized covariance measure, which has been extended to functional linear models by Lundborg et al. (2021). The methods proposed by Zhang et al. (2012) and Strobl et al. (2019) rest on the regression of a function in a RKHS, thereby belonging to this category as well. We also note that there has been a growing interest in estimating the expected conditional covariance in semi-parametric statistics (Li et al., 2011; Newey and Robins, 2018; Robins et al., 2008) often employing non-parametric regression methods followed by adjustments to reduce bias, and this work in turn has implications for the design of CI tests.

Apart from the above two categories, there are many other novel approaches for CI testing developed in recent years. For example, Bellot and van der Schaar (2019); Shi et al. (2020) design nonparametric tests by leveraging the success of generative adversarial networks. Sen et al.

(2018, 2017) convert the CI testing problem into a binary classification problem, which allows one to leverage existing classification algorithms. Approaches based on the partial copula have been examined by Bergsma (2010, 2004); Patra et al. (2016); Petersen and Hansen (2021); Song (2009). A metric-based approach is also common in the literature, including tests based on the conditional Hellinger distance (Su and White, 2008) and conditional mutual information (Runge, 2018). The above methods are mainly for continuous data, whereas there are numerous CI tests available for discrete data as well (Agresti, 1992; Balakrishnan and Wasserman, 2018; Canonne et al., 2018; Kim and Agresti, 1997; Marx and Vreeken, 2019; Neykov et al., 2021; Yao and Tritchler, 1993). A more extensive review of CI tests can be found in Li and Fan (2020).

So far we have mainly reviewed various ways of measuring CI and constructing test statistics. For testing problems, it is also important to determine a reasonable critical value, that results in small type I and type II errors. The current literature usually considers one of the following three approaches for setting critical values.

- **Asymptotic method.** The first common approach is based on the limiting null distribution of a test statistic. Once the limiting null distribution is known, the critical value is determined by using a quantile of this limiting distribution or a bootstrap procedure. In order to obtain a tractable asymptotic distribution, the test statistic typically has an asymptotically linear or quadratic form. Examples of CI tests based on the asymptotic approach include Huang (2010); Shah and Peters (2020); Strobl et al. (2019); Su and White (2008); Wang et al. (2015); Zhang et al. (2012); Zhou et al. (2020). Due to technical hurdles, this line of work often focuses on a pointwise (rather than uniform) type I error guarantee with a few exceptions (Lundborg et al., 2021; Shah and Peters, 2020).

- **Model-X framework.** Formalized by Candés et al. (2018), the model-X framework builds on the assumption that the conditional distribution $P_{X|Z}$ is (approximately) known. In this case, one can compute a set of test statistics, which are exchangeable under the null, by exploiting the knowledge of $P_{X|Z}$ either through direct resampling as in Candés et al. (2018) or via a permutation method as in Berrett et al. (2020b). The critical value is then set to be an empirical quantile of these test statistics, and the resulting test has finite-sample validity. Berrett et al. (2020b) rigorously characterize the excess type I error when an estimate of $P_{X|Z}$ is considered, and also demonstrate situations where this excess error is asymptotically negligible. Nevertheless, this methodology may not be appropriate for applications where $P_{X|Z}$ is hard to estimate.

- **Local permutation method.** The third approach is based on local permutations. This method generates a reference distribution by randomly permuting Y within subclasses, which are defined in terms of the proximity of the conditional variable Z. Then the critical value is determined as a quantile of this reference distribution. The work of Doran et al. (2014); Fukumizu et al. (2008); Margaritis (2005); Sen et al. (2017) fall into this category. When observing multiple samples with the same value of Z is possible, this method can yield an exact CI test with reasonable power against certain alternatives. However, its validity has not been fully explored beyond discrete settings.

As mentioned before, our work heavily builds upon the recent work of Canonne et al. (2018), Neykov et al. (2021) and Kim et al. (2020). Canonne et al. (2018) construct tests for CI when

(X, Y, Z) are discrete random variables, but with a possibly large number of categories, and establish the optimality of their tests from a minimax perspective, in certain regimes. Neykov et al. (2021) extend the work of Canonne et al. (2018) to the case where Z is a continuous and bounded random variable. However, both tests considered in Canonne et al. (2018) and Neykov et al. (2021) rely on critical values that depend on unspecified constants. In this sense, it has been unknown whether there exists a minimax optimal CI test, which is easily implementable in practice. To address this issue, our work considers the local permutation method, which leads to an explicit critical value. In order to analyze the power of the resulting test, we build on results of Kim et al. (2020) who provide a sufficient condition under which the permutation test has non-trivial power for (unconditional) independence testing. To verify this sufficient condition, we build on the analysis of U-statistic-based tests from the previous work of Canonne et al. (2018) and Neykov et al. (2021).

1.2 Our contributions

We now outline our contributions.

- **Hardness result of CI testing (Section 3).** By leveraging the recent work of Barber et al. (2019); Barber (2020), our first contribution (Theorem 1) is to provide a new hardness result for CI testing. For two-sample and unconditional independence testing, one can use the permutation procedure to develop tests that can keep the type I error under control, while having non-trivial power against interesting alternatives (e.g. Kim et al., 2020). However, this is not the case for the continuous CI testing problem. As pointed out by Shah and Peters (2020), any valid CI test should have no power against any alternative when Z is a continuous random variable. In Theorem 1, we formalize that the impossibility of CI testing is more fundamentally determined by the probability of observing collisions in Z, rather than the type of Z. Therefore, even in the discrete or mixture setting, CI testing is difficult or even impossible when the probability of observing the same Z is extremely small.

- **Validity of local permutation tests (Section 4.2).** In continuous settings, one typical way to address the replication problem of Z in the design of tests is to hypothesize some notion of smoothness, i.e. the conditional distribution does not vary too much as a function of Z. Under this hypothesis, we have approximate replicates, which we can use to construct our test statistics, and can use to design an (approximate) permutation method. A basic question is to address the validity of the local permutation method. Our preliminary results (Lemma 1 and Lemma 2) show that one can control the type I error of binning based CI statistics when the binned distribution is indistinguishable from its CI projection (the distribution we obtain from permutations). See Figure 1 for a pictorial illustration.

- **Tightness of our conditions (Section 4.3).** We note that, counterintuitively, increasing the sample size n can make type I error control harder to achieve, because ensuring the indistinguishability of the product measures is more challenging as n increases. This forces us to use finer bins as n increases for type I error control. On the other hand, using bins that are too fine may result in a loss of power, which raises the question on a choice of the size of bins. In Theorem 2 and Theorem 3, we first present concrete upper bounds for the type I error in terms of the size of bins and the sample size n, under certain smoothness conditions.

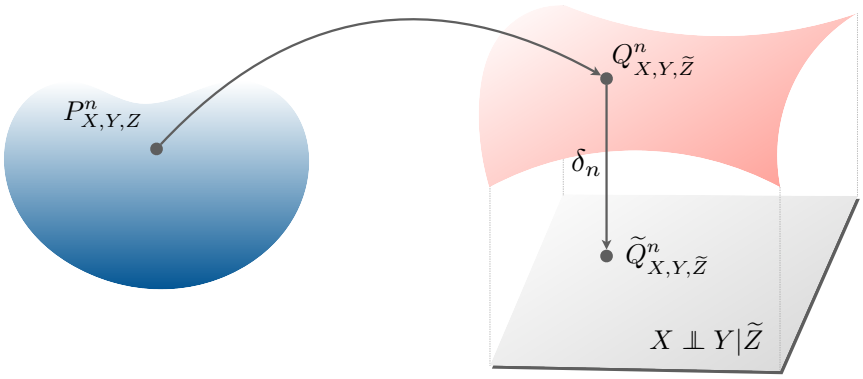
These results guide us on the size of bins, which yields rigorous type I error control. As a complementary result, Theorem 4 proves that the upper bounds in the previous results are asymptotically tight. In particular, we show that there exists a local permutation test whose type I error is arbitrarily close to one when the given upper bounds are sufficiently far from the significance level.

- **Power analysis (Section 5).** The next question we address is that of power. We start by revisiting the test statistics for discrete CI testing in Canonne et al. (2018). Theorem 5 then shows that the corresponding local permutation tests have the same power guarantee as the tests in Canonne et al. (2018). Unlike the discrete CI setting, the replication problem affects both type I and type II error control in the continuous CI setting. As mentioned earlier, taking finer bins helps us for type I error control but not for power. Our next result shows that in some cases, we are able to navigate this trade-off, i.e. there is a choice of binning which ensures that the local permutation method controls the type I error but still retains minimax power. In particular, we show in Theorem 6 and Theorem 7 that the local permutation tests using the same test statistics in Neykov et al. (2021) achieve the same minimax power in the total variation (TV) distance. However, this guarantee comes at a cost. Namely, the local permutation test with the optimal choice of binning is valid over a set of null distributions much smoother than those considered in Neykov et al. (2021).

- **Double-binning strategy (Section 6).** Finally, we develop and analyze a new double-binning based permutation test, which partly addresses the aforementioned drawback. More specifically, we consider bins of two distinct resolutions where finer bins are used for permutations and coarser bins are used to compute a test statistic. By permuting over the finer bins, our theory in Proposition 1 guarantees that a double-binning based test has type I error control over a larger class of null distributions than the single-binning counterpart. On the other hand, by computing the test statistic over the coarser bins, Theorem 8 proves that the power of the resulting test remains the same as the single-binning test, up to a constant factor, under certain regularity conditions. We further demonstrate our theoretical findings in Section 7.3 through simulations.

1.3 Organization

The rest of this paper is organized as follows. We start by explaining the local permutation procedure in Section 2 along with a basic background on probability metrics. Section 3 provides a new hardness result of CI testing that covers the discrete case of Z. We then move on to discussing the validity of the local permutation test in Section 4. In particular, we provide upper bounds for the type I error of the local permutation test under certain smoothness conditions. We further prove that these upper bounds are asymptotically tight in some cases. Focusing on the test statistics proposed by Canonne et al. (2018) and Neykov et al. (2021), we investigate the power property of the local permutation tests in Section 5. Section 6 introduces a double-binning strategy that allows us to choose a smaller binning size without sacrificing power up to a constant factor. Section 7 includes several illustrative simulation results. Finally, we end the paper with a discussion and future work in Section 8. All technical proofs are relegated to the Appendix.



Step 1. Discretization of Z Step 2. CI projection

Figure 1: A visualization of our analysis of the local permutation test. To proceed, we first discretize Z into several bins and denote the binned conditional variable by $\widetilde{Z}$. Given a product distribution $P^n_{X,Y,Z}$, the corresponding distribution smoothed over the bins is denoted by $Q^n_{X,Y,\widetilde{Z}}$ (see Section 4.2 for a more precise description). Next, we consider another distribution $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$, which is the CI projection of $Q^n_{X,Y,\widetilde{Z}}$ onto the space where $X \perp\!\!\!\perp Y|\widetilde{Z}$. The validity of the location permutation test is essentially determined by the TV distance δ_n (or its upper bound) between $Q^n_{X,Y,\widetilde{Z}}$ and $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$. In particular, the local permutation test is asymptotically valid when $\delta_n \to 0$ as $n \to \infty$. In Section 4, we present sharp and tractable conditions under which δ_n converges to zero, depending on the smoothness of underlying distributions.

2 Preliminaries

In this section, we set up the notation and introduce preliminaries including the local permutation procedure and probability metrics.

2.1 Notation

Throughout this paper, we mostly follow the notation used in Neykov et al. (2021). Let the triplet (X, Y, Z) have a distribution $P_{X,Y,Z}$ on a measurable space. We denote the conditional distribution of $X, Y|Z = z$ as $P_{X,Y|Z=z}$. We denote the (marginal) conditional distributions of $X|Z = z$ and $Y|Z = z$ by $P_{X|Z=z}$ and $P_{Y|Z=z}$, respectively. In addition, the marginal distributions of X, Y, Z are denoted by P_X, P_Y, P_Z and similarly the joint marginal distributions are denoted by $P_{X,Y}, P_{X,Z}, P_{Y,Z}$. Moreover, we will use the lowercase p to denote density functions with respect to a base measure. For example, $p_{X,Y|Z}(x, y|z)$, denotes the conditional density (or probability mass) function of $X, Y|Z = z$, evaluated at a point (x, y, z). We denote the set of all distributions for which $X \perp\!\!\!\perp Y|Z$ by $\mathcal{P}_0$.

2.2 Local permutation procedure

We formalize the local permutation procedure based on i.i.d. observations $\{(X_i, Y_i, Z_i)\}_{i=1}^n := (X^n, Y^n, Z^n)$ from $P_{X,Y,Z}$. Throughout this paper, we assume the conditional random variable

Z has compact support $\mathcal{Z}$ and briefly discuss an extension to unbounded support in Appendix B.4. Let $\{B_1, \ldots, B_M\}$ denote a partition of $\mathcal{Z}$ such that $\mathcal{Z} = \cup_{m=1}^{M} B_m$ and $\boldsymbol{\sigma} := \{\sigma_1, \ldots, \sigma_M\}$ denote the sample sizes within bins $\{B_1, \ldots, B_M\}$. Furthermore, let $\boldsymbol{W}_m$ denote the set of the pairs of (X_i, Y_i) that belong to the mth bin. More formally, by letting $(X_{i,m}, Y_{i,m})$ be the ith pair in the mth bin, we write $\boldsymbol{W}_m := \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\}$ when $\sigma_m \geq 1$ and otherwise $\boldsymbol{W}_m = \emptyset$.

Given this binned data, we consider a generic test statistic for CI testing, which maps from $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M$ to $\mathbb{R}$, i.e. for some function $f : \boldsymbol{W}_1, \ldots, \boldsymbol{W}_M \mapsto \mathbb{R}$ we compute our test statistic as:

$$T_{\mathrm{CI}} = f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M). \tag{1}$$

As a concrete example with real-valued data, one can take f to be the average function of arbitrary (unconditional) independence test statistics computed based on $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M$, respectively.

In order to determine significance of the statistic T_{CI}, we rely on the local permutation procedure summarized in Algorithm 1. To describe the algorithm, consider a permutation $\pi_m = \{\pi_m(1), \ldots, \pi_m(\sigma_m)\}$ of $\{1, \ldots, \sigma_m\}$ and denote $\boldsymbol{W}_m^{\pi_m} = \{(X_{1,m}, Y_{\pi_m(1),m}), \ldots, (X_{\sigma_m,m}, Y_{\pi_m(\sigma_m),m})\}$ for $m = 1, \ldots, M$. Notice that when $\sigma_m = 0$, there is nothing to permute and we set $\boldsymbol{W}_m^{\pi_m} = \emptyset$. The test statistic computed using the locally permuted data set is denoted by

$$T_{\mathrm{CI}}^{\boldsymbol{\pi}} = f(\boldsymbol{W}_1^{\pi_1}, \ldots, \boldsymbol{W}_M^{\pi_M}). \tag{2}$$

Let us further denote the set of all possible such local permutations $\boldsymbol{\pi} := \{\pi_1, \ldots, \pi_M\}$ by $\boldsymbol{\Pi}$ whose cardinality is $K = \prod_{m=1}^{M} \sigma_m!$. Given this notation, we describe the local permutation procedure in Algorithm 1.

Algorithm 1 Local permutation procedure

Input: data $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$, a partition of $\mathcal{Z}$: $\{B_1, \ldots, B_M\}$, a test statistic T_{CI}, a nominal level α

1. For each $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, compute $T_{\mathrm{CI}}^{\boldsymbol{\pi}}$ as in (2) and denote the resulting statistics by $T_{\mathrm{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\mathrm{CI}}^{\boldsymbol{\pi}_K}$.
2. By comparing the statistic T_{CI} in (1) with the permuted ones, calculate the p-value as

$$p_{\mathrm{perm}} = \frac{1}{K} \sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}} \mathbb{1}\{T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \geq T_{\mathrm{CI}}\}. \tag{3}$$

3. Given the nominal level $\alpha \in (0, 1)$, define the test function $\phi_{\mathrm{perm},n} = \mathbb{1}(p_{\mathrm{perm}} \leq \alpha)$ and reject the null when $\phi_{\mathrm{perm},n} = 1$.

The local permutation procedure, like other randomized or permutation procedures (e.g. Chapter 15 of Lehmann and Romano, 2006), can be used with any binning-based test statistic for CI testing. For simplicity, our theoretical results are based on Algorithm 1 but they can be easily extended to a more practical permutation procedure via Monte Carlo simulations as remarked below.

Remark 1.

- **Monte Carlo approximation.** The permutation p-value (3) may be practically unappealing as its computational cost is prohibitively expensive for large n. To alleviate this computational issue, it is a common practice to approximate p_{perm} using Monte Carlo simulations as in (22). As noted in Lehmann and Romano (2006), the difference between p_{perm} and its Monte Carlo approximation can be made arbitrarily small by taking a sufficiently large number of Monte Carlo samples. This can be formally stated using Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956) and we refer to Corollary 6.1 of Kim (2021) or Proposition 4 of Schrab et al. (2021) for such argument.

- **Randomization.** It is well-known that the permutation test can be made exact by introducing randomization. We state the randomized permutation test (Hoeffding, 1952) for completeness. For a nominal level α, we denote $k = K - [K\alpha]$ where $[K\alpha]$ is the largest integer less than or equal to $K\alpha$. In addition let K^+ and K^0 be the numbers of $T_{\mathrm{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\mathrm{CI}}^{\boldsymbol{\pi}_K}$, which are greater than or equal to T_{CI}, respectively. Given $a = (K\alpha - K^+)/K^0$ and the kth order statistic $T_{\mathrm{CI}}^{(k)}$ of $T_{\mathrm{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\mathrm{CI}}^{\boldsymbol{\pi}_K}$, we define $\phi_{\mathrm{perm},n,a} = 1, a$ or 0 depending on whether $T_{\mathrm{CI}} > T_{\mathrm{CI}}^{(k)}$, $T_{\mathrm{CI}} = T_{\mathrm{CI}}^{(k)}$ or $T_{\mathrm{CI}} < T_{\mathrm{CI}}^{(k)}$, respectively. Then under the exchangeability assumption of the permuted statistics, it holds that $\mathbb{E}[\phi_{\mathrm{perm},n,a}] = \alpha$, whereas $\phi_{\mathrm{perm},n}$ from Algorithm 1 has a weaker guarantee that $\mathbb{E}[\phi_{\mathrm{perm},n}] \leq \alpha$ in general.

In the next subsection, we present several statistical distances between probability measures that we make use of throughout this paper.

2.3 Probability metrics

Let P and Q be two probability measures over a measurable space $(\Omega, \mathcal{F})$ and denote the densities of P and Q with respect to a common dominating measure μ by p and q, respectively. There are two classes of probability metrics that will be considered in this paper. The first class, we call the generalized Hellinger distance (e.g. Kamps, 1989), includes the TV distance and the Hellinger distance as special cases.

Definition 1 (Generalized Hellinger distances)**.** Given $\gamma \geq 1$, the generalized Hellinger distance with parameter γ between P and Q is defined as

$$\mathcal{D}_{\gamma,\mathrm{H}}(P, Q) = \left(\frac{1}{2} \int \left| p^{1/\gamma} - q^{1/\gamma} \right|^{\gamma} d\mu \right)^{1/\gamma}.$$

From the definition, it is clear that the above distance becomes the TV distance when $\gamma = 1$ and the Hellinger distance when $\gamma = 2$. Since these two values deserve special attention, we denote the corresponding TV distance and Hellinger distance by $\mathcal{D}_{\mathrm{TV}}(P, Q)$ and $\mathcal{D}_{\mathrm{H}}(P, Q)$, respectively. The generalized Hellinger distance has the monotonicity property that $\mathcal{D}_{\gamma_2,\mathrm{H}}^{\gamma_2}(P, Q) \leq \mathcal{D}_{\gamma_1,\mathrm{H}}^{\gamma_1}(P, Q)$ for $1 \leq \gamma_1 \leq \gamma_2$ (Corollary 3 of Kamps, 1989). This monotonic relationship generalizes the well-known inequality between the TV and Hellinger distances, namely $\mathcal{D}_{\mathrm{H}}^2(P, Q) \leq \mathcal{D}_{\mathrm{TV}}(P, Q)$ (e.g. Chapter 4 of Le Cam, 2012).

Another class of probability metrics that we consider is Rényi divergence defined as follows.

Definition 2 (Rényi divergences)**.** For $\gamma \in (0, \infty)$, Rényi divergence of order γ of P from Q is defined as

$$\mathcal{D}_{\gamma,\mathrm{R}}(P\|Q) = \begin{cases} \dfrac{1}{\gamma-1}\log\left\{\displaystyle\int \left(\frac{p}{q}\right)^{\gamma} q d\mu\right\}, & \text{if } \gamma \neq 1, \\ \displaystyle\int p \log\left(\frac{p}{q}\right) d\mu, & \text{if } \gamma = 1. \end{cases}$$

In the above definition, some notable values of γ include $\gamma \in \{1/2, 1, 2\}$ and the corresponding Rényi divergence is directly or indirectly associated with the Hellinger distance ($\gamma = 1/2$), Kullback–Leibler (KL) divergence ($\gamma = 1$) and χ^2 divergence ($\gamma = 2$) as stated in Appendix B. We refer the reader to Van Erven and Harremos (2014) and Sason and Verdú (2016) for more information on Rényi divergences.

3 Fundamental limits of CI testing

Before we start analyzing local permutation tests, we provide a new hardness result for CI testing. In view of the recent hardness result of Shah and Peters (2020), further revisited by Neykov et al. (2021), CI testing is intrinsically difficult in the following sense. Let $\mathcal{P}_{\mathbb{R}^d}$ be the set of all distributions for (X, Y, Z) on $\mathbb{R}^{d_x+d_y+d_z}$, and let $\mathcal{P}_K \subset \mathcal{P}_{\mathbb{R}^d}$ be the subset of $\mathcal{P}_{\mathbb{R}^d}$ whose support is defined within a L_∞ ball of radius K. We also assume that the distributions in $\mathcal{P}_K$ are absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{P}_{0,K} \subset \mathcal{P}_K$ be the subset of distributions such that $X \perp\!\!\!\perp Y|Z$ and denote its complement by $\mathcal{P}_{1,K} = \mathcal{P}_K \setminus \mathcal{P}_{0,K}$. By denoting the joint distribution of n i.i.d. random vectors from $P_{X,Y,Z}$ by $P^n_{X,Y,Z}$, the result of Shah and Peters (2020) states that any valid CI test ϕ for the class of null distributions $\mathcal{P}_{0,K}$ should satisfy $\mathbb{E}_{P^n_{X,Y,Z}}[\phi] \leq \alpha$ for any $P_{X,Y,Z} \in \mathcal{P}_{1,K}$. In words, no CI test, which has valid type I error control for all absolutely continuous conditionally independent distributions, can have meaningful power against any single alternative distribution in $\mathcal{P}_{1,K}$. It therefore emphasizes that one should consider smaller sets of null and alternative distributions in order to make the CI testing problem feasible.

The story, on the other hand, is different when Z has a discrete or a mixture distribution where one can observe the same value of Z. In this case, by permuting the samples within groups having the same value of Z, the local permutation test can be valid, while possessing non-trivial power against certain alternatives. However, even in this case, there exists an intrinsic difficulty of CI testing when the probability of observing the same Z is extremely small. We precisely characterize this challenge in the following theorem. To describe the result, let $\mathcal{P}_0 \subset \mathcal{P}_{\mathbb{R}^d}$ be the subset of distributions such that $X \perp\!\!\!\perp Y|Z$, and define $\mathcal{P}_1 = \mathcal{P}_{\mathbb{R}^d} \setminus \mathcal{P}_0$. Further, let $\mathcal{P}_{0,\mathrm{disc}} \subset \mathcal{P}_0$ be the subset of null distributions where Z is supported on a countable set. Then our result is stated as follows.

Theorem 1 (Hardness of CI testing)**.** *For an arbitrary integer $J \geq n(n-1)$, let us define $\rho_{J,P} := \mathbb{P}\{Z_1, \ldots, Z_J$ are distinct$\}$, where $Z_1, \ldots, Z_J$ are i.i.d. samples from the marginal distribution of Z. Suppose that a test ϕ satisfies $\sup_{P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{disc}}} \mathbb{E}_{P^n_{X,Y,Z}}[\phi] \leq \alpha$ for $\alpha \in (0,1)$. Then for any $P_{X,Y,Z} \in \mathcal{P}_1$, the power of ϕ is bounded above by*

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi] \leq \alpha \times \rho_{J,P} + (1 - \rho_{J,P}) + \frac{n(n-1)}{J}. \tag{4}$$

A few remarks on this result are given below.

Remark 2.

- Theorem 1 states that what makes the CI problem hard is not just whether Z is discrete or continuous, but whether one can observe the same value of Z repeatedly with high probability. This difficulty is precisely captured by the quantity $\rho_{J,P}$. As an illustration, suppose that Z has a multinomial distribution with equal probabilities over bins. Intuitively, when the number of bins is much larger than the sample size, one cannot expect to see the same Z even twice with high probability, and thus the bound (4) becomes close to α. We make this intuition more precise in Remark 10 where we demonstrate that any valid CI test becomes (asymptotically) powerless if Z is uniformly distributed over bins and the number of bins increases much faster than n^4. We also refer to Section 7.1 for a numerical illustration.

- Our result is not restricted to the case of discrete Z. Suppose that ϕ is valid over a subset of $\mathcal{P}_0$ that contains $\mathcal{P}_{0,\text{disc}}$. Then the same result trivially follows. Let $\mathcal{P}_{1,\text{no-atom}} \subset \mathcal{P}_1$ be the subset of alternative distributions where the marginal distribution of Z has no atoms, i.e. $\rho_{J,P} = 1$. Then, as a corollary of Theorem 1, it holds that $\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,\text{no-atom}}} \mathbb{E}_{P^n_{X,Y,Z}}[\phi] \leq \alpha$ for any ϕ such that $\sup_{P_{X,Y,Z} \in \mathcal{P}_0} \mathbb{E}_{P^n_{X,Y,Z}}[\phi] \leq \alpha$. However, in this argument, it is crucial to assume that $\mathcal{P}_{0,\text{disc}} \subset \mathcal{P}_0$. In contrast, the hardness result of Shah and Peters (2020) does not require $\mathcal{P}_{0,\text{disc}} \subset \mathcal{P}_0$. In this sense, Theorem 1 is a weaker result than Theorem 2 of Shah and Peters (2020).

- On the other hand, the proof of Theorem 1 is much simpler than that of the hardness result of Shah and Peters (2020), being highly motivated by the recent impossibility results in distribution-free conditional predictive inference e.g., Lemma A.1 of Barber et al. (2019) and Lemma 1 of Barber (2020). The key idea of the proof is to introduce ghost samples $\{X_i, Y_i, Z_i\}_{i=1}^J$ and express the type II error of ϕ as the iterative expectations associated with sampling without replacement from these ghost samples. When $Z_1, \ldots, Z_J$ are distinct, random draws from these ghost samples with replacement can be viewed as random draws under the null. We note that such argument via sampling with replacement can be traced back to Gretton et al. (2012) (see Example 1) who provide a negative result for two-sample testing. Given this key observation, we can connect the type II error with the significance level α, and the result follows by a union bound along with the total variation distance between sampling with and without replacement. The details can be found in Appendix A.1.

In summary, the hardness result of Shah and Peters (2020) and Theorem 1 indicate that CI testing is a difficult task without further assumptions. This negative result naturally motivates us to explore reasonable conditions under which CI testing, especially based on the local permutation procedure, is feasible. This is the main topic of the next section.

4 Validity under smoothness conditions

The goal of this section is two-fold: one is to identify universal conditions under which any local permutation test based on a binned statistic is asymptotically valid; the other is to show that these

conditions are tight under certain smoothness conditions in the sense that there exists a permutation test whose type I error is not controlled even asymptotically when our conditions are violated. We start by introducing our smoothness assumptions (Section 4.1) and then state the main results (Section 4.2 and Section 4.3).

4.1 Smoothness conditions

The validity of the local permutation test crucially relies on the smoothness assumption on conditional distributions. For instance, suppose that $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$ are constant with respect to z for all $z \in \mathcal{Z}$. In this case, X and Y are independent of Z and, consequently, CI testing is the same as unconditional independence testing for which the (local) permutation test has finite-sample validity. Motivated by this observation, we consider similar (but more general) smoothness classes to those in Neykov et al. (2021) defined as follows.

Definition 3 (γ-Hellinger Lipschitzness)**.** *Let $\mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L) \subset \mathcal{P}_0$ be the collection of distributions $P_{x,y,z}$ such that for all $z, z' \in \mathcal{Z}$,*

$$\mathcal{D}_{\gamma,\mathrm{H}}(P_{X|Z=z}, P_{X|Z=z'}) \le L\delta(z,z') \quad \textit{and} \quad \mathcal{D}_{\gamma,\mathrm{H}}(P_{Y|Z=z}, P_{Y|Z=z'}) \le L\delta(z,z'),$$

where $\delta(z,z')$ is a distance between z and z' in $\mathcal{Z}$.

Another smoothness assumption is made based on Rényi divergence.

Definition 4 (γ-Rényi Lipschitzness)**.** *Let $\mathcal{P}_{0,\mathrm{R},\gamma,\delta}(L) \subset \mathcal{P}_0$ be the collection of distributions $P_{x,y,z}$ such that for all $z, z' \in \mathcal{Z}$,*

$$\mathcal{D}^{1/2}_{\gamma,\mathrm{R}}(P_{X|Z=z} \| P_{X|Z=z'}) \le L\delta(z,z') \quad \textit{and} \quad \mathcal{D}^{1/2}_{\gamma,\mathrm{R}}(P_{Y|Z=z} \| P_{Y|Z=z'}) \le L\delta(z,z'),$$

where $\delta(z,z')$ is a distance between z and z' in $\mathcal{Z}$.

For both Lipschitz conditions, we let $\delta(\cdot,\cdot)$ be the Euclidean distance by default when $Z \in \mathbb{R}^d$. With these smoothness conditions in place, the next section studies the asymptotic validity of the local permutation test.

4.2 Validity

To state the validity result, we begin with additional notation. Let $(X^n, Y^n, \widetilde{Z}^n)$ be the binned version of (X^n, Y^n, Z^n) where $\widetilde{Z} \in \{1, \dots, M\}$ is a discrete random variable with probability $\mathbb{P}(\widetilde{Z} = m) = \mathbb{P}(Z \in B_m)$ for $m = 1, \dots, M$. For simplicity, let us write $q_{\widetilde{Z}}(m) = \mathbb{P}(\widetilde{Z} = m)$ and denote the conditional distribution of $Z|Z \in B_m$ by $\widetilde{P}_{Z|Z\in B_m}$. Then $(X, Y, \widetilde{Z})$ has its density function $q_{X,Y,\widetilde{Z}}(x,y,m) = q_{XY|\widetilde{Z}}(x,y|m) q_{\widetilde{Z}}(m)$ where

$$q_{XY|\widetilde{Z}}(x,y|m) = \int_{B_m} p_{X,Y|Z}(x,y|z) d\widetilde{P}_{Z|Z\in B_m}(z),$$

and we denote the corresponding joint distribution by $Q_{X,Y,\widetilde{Z}}$.

Under the considered binning scheme, we make a key observation that the test statistic T_{CI} is defined only through the binned data, which means that T_{CI} based on (X^n, Y^n, Z^n) is equal in distribution to that based on $(X^n, Y^n, \widetilde{Z}^n)$. Furthermore, since our test function $\phi_{\mathrm{perm},n}$ is computed only through the binned data, we observe that

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] = \mathbb{E}_{Q^n_{X,Y,\widetilde{Z}}}[\phi_{\mathrm{perm},n}]. \tag{5}$$

To proceed further, we consider a product density $\widetilde{q}_{X,Y,\widetilde{Z}}(x,y,m) = q_{X\cdot|\widetilde{Z}}(x|m)q_{\cdot Y|\widetilde{Z}}(y|m)q_{\widetilde{Z}}(m)$ where $q_{X\cdot|\widetilde{Z}}(x|m)$ and $q_{\cdot Y|\widetilde{Z}}(y|m)$ are the marginals of $q_{XY|\widetilde{Z}}(x,y|m)$, and denote the corresponding joint distribution by $\widetilde{Q}_{X,Y,\widetilde{Z}}$. By construction, $(X^n, Y^n, \widetilde{Z}^n)$ from $\widetilde{Q}_{X,Y,\widetilde{Z}}$ satisfies $X \perp\!\!\!\perp Y|\widetilde{Z}$ and therefore exchangeability of Y within each bin yields

$$\mathbb{E}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}[\phi_{\mathrm{perm},n}] \leq \alpha.$$

Combining the above inequality with the identity (5) yields a generic type I error bound for the local permutation test in terms of the total variation distance.

Lemma 1 (Type I error bound in terms of the TV distance)**.** *Suppose that the distribution $P_{X,Y,Z}$ belongs to $\mathcal{P}_0$. Then for any $\alpha \in (0,1)$, the type I error of $\phi_{\mathrm{perm},n}$ is bounded above by*

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] \leq \alpha + \mathcal{D}_{\mathrm{TV}}\big(Q^n_{X,Y,\widetilde{Z}}, \widetilde{Q}^n_{X,Y,\widetilde{Z}}\big).$$

The given bound implies that the local permutation test is valid when the binned null distribution is indistinguishable from its CI projection. The proof of this result follows by the definition of the TV distance. It is worth pointing out that since the randomized permutation test $\phi_{\mathrm{perm},n,a}$ (Remark 1) is exact under the law of $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$, one can establish a stronger result that

$$\big|\mathbb{E}_{P^n_{X,Y,Z},R}[\phi_{\mathrm{perm},n,a}] - \alpha\big| \leq \mathcal{D}_{\mathrm{TV}}\big(Q^n_{X,Y,\widetilde{Z}}, \widetilde{Q}^n_{X,Y,\widetilde{Z}}\big),$$

where $R \sim \mathrm{Uniform}[0,1]$ is for randomization and independent of the data. In either randomized or non-randomized test, our main task boils down to identifying reasonable conditions under which the TV distance between $Q^n_{X,Y,\widetilde{Z}}$ and $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$ tends to zero asymptotically. In general, however, it is challenging to directly work with the TV distance between two product measures. Instead we upper bound the TV distance by the Hellinger distance as follows.

Lemma 2 (Type I error bound in terms of the Hellinger distance)**.** *Suppose that the distribution $P_{X,Y,Z}$ belongs to $\mathcal{P}_0$. Then for any $\alpha \in (0,1)$, the type I error of $\phi_{\mathrm{perm},n}$ is bounded above by*

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] \leq \alpha + \left\{2n \sum_{m=1}^{M} q_{\widetilde{Z}}(m) \times \mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big)\right\}^{1/2}, \tag{6}$$

where the second term on the right-hand side is simply $\sqrt{2n}\mathcal{D}_{\mathrm{H}}\big(Q_{X,Y,\widetilde{Z}}, \widetilde{Q}_{X,Y,\widetilde{Z}}\big)$.

We mention that, while the Hellinger bound (6) is provably looser than that based on the TV distance, it has the same characterization as the TV bound in terms of when the local permutation is asymptotically valid. More specifically, using well-known bounds relating the TV distance and the Hellinger distance, it can be verified that

$$\mathcal{D}_{\mathrm{TV}}\big(Q^n_{X,Y,\widetilde{Z}}, \widetilde{Q}^n_{X,Y,\widetilde{Z}}\big) \to 0 \text{ if and only if } n\mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y,\widetilde{Z}}, \widetilde{Q}_{X,Y,\widetilde{Z}}\big) \to 0.$$

We are now ready to state the main results of this section. To this end, we denote the maximum diameter of bins $B_1, \ldots, B_M$ by h, that is

$$h := \max_{1\le m\le M} \sup_{z,z'\in B_m} \delta(z, z'). \tag{7}$$

We first state the result under γ-Hellinger Lipschitzness in the next theorem and then consider γ-Rényi Lipschitzness in Theorem 3.

Theorem 2 (Validity of $\phi_{\mathrm{perm},n}$ under γ-Hellinger Lipschitzness)**.** *For any $\alpha \in (0,1)$, the type I error of $\phi_{\mathrm{perm},n}$ under γ-Hellinger Lipschitzness is bounded above by*

$$\sup_{P_{X,Y,Z}\in\mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)} \mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] \le \begin{cases} \alpha + C_\gamma n^{1/2} L^\gamma h^\gamma, & \text{if } \gamma \in [1,2], \\ \alpha + C_\gamma n^{1/2} L^2 h^2, & \text{if } \gamma > 2, \end{cases} \tag{8}$$

where C_γ is a constant that only depends on γ.

A few remarks are in order.

Remark 3.

- The result of Theorem 2 shows that once we assume that L and γ are fixed in the sample size, a sufficient condition for the validity of $\phi_{\mathrm{perm},n}$ is $h = o(n^{-1/2\gamma})$ if $\gamma \in [1,2]$ and $h = o(n^{-1/4})$ if $\gamma > 2$ under γ-Hellinger Lipschitzness. Since users have control over h, Theorem 2 provides a guideline for the choice of binning that ensures type I error control of the local permutation test. We also note that when h is too small, most of the bins are empty, which adversely affects the power performance. In other words, there is an intriguing trade-off between the type I error and power in terms of the choice of h. We discuss this trade-off more in Section 5.

- In most practical applications, the smoothness parameter γ is unknown. In such case, one can choose h such that $n^{1/2}h \to 0$ as $n \to \infty$, which leads to an asymptotically valid permutation test for all $\gamma \ge 1$, assuming other parameters are fixed. However, as mentioned before, choosing small h comes at a price of low power under the alternative. It would be interesting to see whether an adaptive way of choosing h is possible without much loss of power. We leave this direction to future work.

- We observe an interesting phenomenon that there exists a sharp transition at $\gamma = 2$, which corresponds to the Hellinger distance. In particular, the result illustrates that the smoothness condition beyond $\gamma > 2$ does not really help to improve the convergence rate of the type I error. Importantly, the given upper bound (8) is tight in terms of n and h in some cases. More specifically, we show in Section 4.3 that there exists a local permutation test whose type I error rate can be made arbitrarily large unless the upper bound (8) converges to α.

- The proof of Theorem 2 builds on Lemma 2 and the monotonicity property of the generalized Hellinger distance. We note that Lemma 2 has a bound in terms of smoothed distributions over partitions, whereas γ-Hellinger Lipschitzness is stated in terms of original distributions. The bulk of the effort in proving Theorem 2 lies in connecting the Hellinger distance between $Q_{X,Y|\widetilde{Z}=m}$ and $\widetilde{Q}_{X,Y|\widetilde{Z}=m}$ to the γ-Hellinger distance between $P_{X|Z=z}$ and $P_{X|Z=z'}$ (and also between $P_{Y|Z=z}$ and $P_{Y|Z=z'}$). The details can be found in Appendix A.3.

Next we present a similar result under γ-Rényi Lipschitzness.

Theorem 3 (Validity of $\phi_{\mathrm{perm},n}$ under γ-Rényi Lipschitzness)**.** *For any $\alpha \in (0,1)$ and $\gamma > 0$, the type I error of $\phi_{\mathrm{perm},n}$ under γ-Rényi Lipschitzness is bounded above by*

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{R},\gamma,\delta}(L)} \mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] \ \leq \alpha + C_\gamma n^{1/2} L^2 h^2, \tag{9}$$

where C_γ is a constant that only depends on γ.

In contrast to Theorem 2, the above result indicates that the smoothness parameter γ in Rényi Lipschitzness does not affect the type I error of $\phi_{\mathrm{perm},n}$ by more than a constant factor. At a high-level, we observe this phenomenon because Rényi divergence is lower bounded by the squared Hellinger distance, up to a constant factor, for any $\gamma > 0$ (see Lemma 5). In other words, the conditional distribution in $\mathcal{P}_{0,\mathrm{R},\gamma,\delta}(L)$ is at least as smooth as the one in $\mathcal{P}_{0,\mathrm{H},\gamma=2,\delta}(L)$, which means that we are essentially in the second regime of Theorem 2 for $\gamma \geq 2$. Indeed, it should be clear from the proof that the same upper bound in Theorem 3 holds for any Lipschitzness condition whose underlying divergence is lower bounded by the squared Hellinger distance such as KL divergence and χ^2 divergence.

Remark 4 (Poissonization)**.** To analyze the power of binning-based tests, it is often convenient to assume that the sample size has a Poisson distribution (e.g. Balakrishnan and Wasserman, 2019; Canonne et al., 2018; Neykov et al., 2021). This, so-called, Poissonization trick allows us to bypass the difficulty in dealing with the dependence between different bins. In fact, as we proved in Proposition 2 in Appendix B.3, the local permutation test under Poissonization has the same validity as before in Theorem 2 and Theorem 3. To explain it briefly, we shall use the convenient notation

$$\mathbb{E}_{P^N_{X,Y,Z},N}[\cdot] := \sum_{k=0}^{\infty} \mathbb{P}(N=k) \mathbb{E}_{P^k_{X,Y,Z}}[\cdot]$$

to denote the expectation operator with respect to $P^N_{X,Y,Z}$ where N is a random sample size. We similarly write $\mathbb{P}_{P^N_{X,Y,Z},N}[\cdot]$ to denote $\mathbb{E}_{P^N_{X,Y,Z},N}[\mathbb{1}(\cdot)]$. Suppose now that $\sup_{P_{X,Y,Z} \in \mathcal{P}'_0} \mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n}] \leq \alpha + Cn^{-\epsilon}$ for $\mathcal{P}'_0 \subset \mathcal{P}_0$ and some constants $C, \epsilon > 0$. Then under the same condition, the permutation test under Poissonization (i.e. $N \sim \mathrm{Pois}(n)$) satisfies $\sup_{P_{X,Y,Z} \in \mathcal{P}'_0} \mathbb{E}_{P^N_{X,Y,Z},N}[\phi_{\mathrm{perm},N}] \leq \alpha + C_\epsilon n^{-\epsilon}$ where C_ϵ is a constant that only depends on ϵ. See Proposition 2 for a more rigorous statement.

We now move to the next section where we provide complementary results of this section.

4.3 Lower bounds

In this section, we demonstrate that the upper bounds for the type I error established in Section 4.2 cannot be improved further in some cases. In particular, we prove that there exists a local permutation test whose type I error cannot be controlled if one chooses h in such a way that the upper bounds (8) or (9) diverge. In order to simplify our presentation, we focus on the case where X and Y are discrete random variables whereas Z is continuous and bounded between $[0, 1]$. Other cases such as multivariate continuous (X, Y, Z) will be discussed in Remark 5.

Suppose that X and Y are discrete random variables supported on $[\ell_1] \times [\ell_2]$ for some positive integers ℓ_1 and ℓ_2. By convention, $[\ell_1]$ denotes the set of integers $\{1, \ldots, \ell_1\}$ and $[\ell_2]$ is similarly defined. Let $\{B_1, \ldots, B_M\}$ be an equi-partition of $[0, 1]$ so that the length of each bin is $h = M^{-1}$. The given partition yields the binned data sets $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M$ defined in Section 2.2. To study lower bounds, we work with the weighted sum of U-statistics proposed by Canonne et al. (2018) and Neykov et al. (2021). Let

$$\psi^m_{ij}(x, y) = \mathbb{1}(X_{i,m} = x, Y_{i,m} = y) - \mathbb{1}(X_{i,m} = x)\mathbb{1}(Y_{j,m} = y),$$

and define a kernel function as

$$h^m_{i_1,i_2,i_3,i_4} = \frac{1}{4!} \sum_{\pi \in \boldsymbol{\Pi}_4} \sum_{x \in [\ell_1], y \in [\ell_2]} \psi^m_{\pi(1)\pi(2)}(x, y)\psi^m_{\pi(3)\pi(4)}(x, y),$$

where $\boldsymbol{\Pi}_4$ is the set of all permutations of $\{i_1, i_2, i_3, i_4\}$. By linearity of expectations, it is seen that $h^m_{i_1,i_2,i_3,i_4}$ is an unbiased estimator of the squared L_2 norm between $Q_{X,Y|\widetilde{Z}=m}$ and $\widetilde{Q}_{X,Y|\widetilde{Z}=m}$. Given this kernel and by recalling $\boldsymbol{W}_m = \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\}$, the resulting U-statistic is calculated as

$$U(\boldsymbol{W}_m) := \frac{1}{\binom{\sigma_m}{4}} \sum_{i_1<i_2<i_3<i_4 : (i_1,i_2,i_3,i_4) \in [\sigma_m]} h^m_{i_1,i_2,i_3,i_4}. \tag{10}$$

The final statistic is a weighted sum of $U(\boldsymbol{W}_1), \ldots, U(\boldsymbol{W}_M)$ given by

$$T_{\mathrm{CI}} := \sum_{m \in [M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U(\boldsymbol{W}_m). \tag{11}$$

Several properties of T_{CI}, such as minimax power optimality, have been studied under Poissonization by Canonne et al. (2018) and Neykov et al. (2021). To fully benefit from their results, we work with a modified local permutation test: First draw $N \sim \mathrm{Pois}(n/2)$ and accept the null when $N > n$. If $N \leq n$, we carry out a local permutation test with N samples randomly chosen from (X^n, Y^n, Z^n). Formally, we define the modified local permutation test as

$$\phi^\dagger_{\mathrm{perm},n} := \phi_{\mathrm{perm},N} \times \mathbb{1}(N \leq n), \tag{12}$$

where $\phi_{\mathrm{perm},N}$ denotes the local permutation test using T_{CI} (11) computed based on the N samples chosen before. By Proposition 2 along with the inequality $\phi^\dagger_{\mathrm{perm},n} \leq \phi_{\mathrm{perm},N}$, it is clear that $\phi^\dagger_{\mathrm{perm},n}$ is asymptotically valid whenever the upper bounds in (8) and (9) converge to α. The next theorem provides complementary results establishing lower bounds for the type I error of $\phi^\dagger_{\mathrm{perm},n}$.

Theorem 4 (Lower bounds). *For arbitrarily small but fixed $\alpha, \beta > 0$, choose $h_n \to 0$ such that $\sqrt{n}h_n^\gamma \to \infty$ for $\gamma \in [1, 2]$ and $\sqrt{n}h_n^2 \to \infty$ for $\gamma > 2$ under γ-Hellinger Lipschitzness. On the other hand, choose $h_n \to 0$ such that $\sqrt{n}h_n^2 \to \infty$ for $\gamma > 0$ under γ-Rényi Lipschitzness. Consider $\phi^\dagger_{\mathrm{perm},n}$ based on T_{CI} in (11). Then there exist constants $n_0, L > 0$, where L depends only on γ, such that for all $n \geq n_0$, the following two inequalities hold:*

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)} \mathbb{E}_{P^N_{X,Y,Z},N}[\phi^\dagger_{\mathrm{perm},n}] \geq 1 - \beta \quad \textit{and}$$

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{R},\gamma,\delta}(L)} \mathbb{E}_{P^N_{X,Y,Z},N}[\phi^\dagger_{\mathrm{perm},n}] \geq 1 - \beta.$$

Let us provide several comments on Theorem 4.

Remark 5.

- A crucial observation is that the type I error of $\phi^\dagger_{\mathrm{perm},n}$ for CI testing corresponds to its power for testing

$$\widetilde{H}_0 : Q_{X,Y,\widetilde{Z}} = \widetilde{Q}_{X,Y,\widetilde{Z}} \quad \text{versus} \quad \widetilde{H}_1 : Q_{X,Y,\widetilde{Z}} \neq \widetilde{Q}_{X,Y,\widetilde{Z}}. \tag{13}$$

  This means that, in order to verify that the type I error of $\phi^\dagger_{\mathrm{perm},n}$ is inflated, it suffices to show that $\phi^\dagger_{\mathrm{perm},n}$ is asymptotically powerful against the above alternative (13). With this observation in place, our main task is to construct a distributional setting where $\phi^\dagger_{\mathrm{perm},n}$ is able to distinguish the binned distribution and its CI projection with high probability. In fact, the conditions of Theorem 4 guarantee that these two distributions are far enough for $\phi^\dagger_{\mathrm{perm},n}$ to be asymptotically powerful.

- In order to ease our analysis, we carefully design the distribution of Z such that most samples are observed in one of the partitions with high probability. In this case, the test statistic T_{CI} approximates $\sigma_1 U(\boldsymbol{W}_1)$, which is much easier to handle. It is then sufficient to study the permutation test based on $\sigma_1 U(\boldsymbol{W}_1)$ and prove that it is asymptotically powerful under the given conditions. We show that this is indeed the case by building on the results of Kim et al. (2020) where the authors investigate the permutation test based on $U(\boldsymbol{W}_1)$ for unconditional independence testing.

- We expect that $\phi_{\mathrm{perm},n}$ (i.e. without Poissionization) also achieves the same error bounds in Theorem 4 as it always uses more samples than $\phi^\dagger_{\mathrm{perm},n}$ by definition. See empirical evidence in Figure 4. However, we found it challenging to analyze T_{CI} without Poissonization, especially its variance, due to a non-trivial dependence between the summands of T_{CI}. Due to this technical difficulty, we focus on the Poisson-sampling scheme as in Canonne et al. (2018) and Neykov et al. (2021) and leave the detailed analysis of $\phi_{\mathrm{perm},n}$ to future work. Nevertheless, the concentration property of a Poisson random variable allows us to say a certain negative result on $\phi_{\mathrm{perm},n}$ without Poissonization. Specifically, note that a Poisson random variable N with parameter $n/2$ is bound between $cn \leq N \leq Cn$ with high probability where $c, C > 0$ some positive constants (e.g. Canonne, 2017). Thus it is guaranteed that one can find a fixed sample size $\widetilde{n}$ such that $cn \leq \widetilde{n} \leq Cn$ and Theorem 4 holds for $\phi_{\mathrm{perm},\widetilde{n}}$ without Poissonization.

- For simplicity, we prove Theorem 4 using the example where (X, Y) are discrete and Z is a univariate continuous random variable. Nevertheless, our proof can be extended to the case of multivariate continuous (X, Y, Z) as follows. Consider piecewise constant densities for (X, Y) and assume that the components $Z^{(2)}, \ldots, Z^{(d_Z)}$ of $Z \in \mathbb{R}^{d_Z}$ are independent of the rest of variables. In this case, we are essentially in the setting where X, Y are discrete and Z is univariate. Therefore, the same proof carries through except now that the maximum diameter h depends on d_Z. When an equi-partition is considered, we note that d_Z only affects the scaling factor in h and hence the statement of Theorem 4 remains true, provided that d_Z is fixed. We also note that the upper bound results (Theorem 2 and Theorem 3) are stated in terms of the maximum diameter h; thereby the upper bounds remain the same for both univariate and multivariate cases of Z.

So far we have explored type I error control of the local permutation test. Next we turn our attention to the power and study its optimality in certain regimes.

5 Power analysis

This section considers both discrete and continuous cases of the conditional variable Z and investigates the power property of local permutation tests. In order to achieve meaningful power, we focus on a subset of alternatives, which are at least ε far away from the null in terms of the TV distance. Our main interest is then to characterize ε for which local permutation tests can be powerful.

5.1 Discrete (X, Y, Z)

To start with the discrete case where $(X, Y, Z) \in [\ell_1] \times [\ell_2] \times [M]$, we revisit the test statistics proposed by Canonne et al. (2018) and demonstrate the power property of the local permutation procedure based on the same test statistics. Canonne et al. (2018) propose two test statistics for CI testing. The first one is T_{CI} given in (11), which is defined as the sum of unweighted U-statistics. While T_{CI} is simple and performs well in certain regimes, it may suffer from a large variance especially when the dimensions ℓ_1 and ℓ_2 of X and Y are large. To mitigate this issue, Canonne et al. (2018) propose another statistic building on the flattening idea of Diakonikolas and Kane (2016). The latter statistic can be viewed as the sum of weighted U-statistics as observed by Neykov et al. (2021).

Weighted U-statistic. To proceed, let us formally write down the weighted U-statistic. First recall that $\boldsymbol{W}_m = \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\}$ is the set of pairs of (X_i, Y_i) with $Z_i = m$. Suppose that the sample size of $\boldsymbol{W}_m$ is $\sigma_m \geq 4$ and $\sigma_m = 4 + 4t_m$ for some $t_m \in \mathbb{N}$. Following the notation in Neykov et al. (2021), let $t_{1,m} := \min\{t_m, \ell_1\}$ and $t_{2,m} := \min\{t_m, \ell_2\}$. We then randomly split the data $\boldsymbol{W}_m$ into three sets $\boldsymbol{W}_{X,m}$, $\boldsymbol{W}_{Y,m}$ and $\boldsymbol{W}_{XY,m}$ of size $t_{1,m}$, $t_{2,m}$ and $2t_m + 4$, respectively, where $\boldsymbol{W}_{X,m} := \{X_{i,m} : i \in [t_{1,m}]\}$, $\boldsymbol{W}_{Y,m} := \{Y_{i,m} : t_{1,m} + 1 \leq i \leq t_{1,m} + t_{2,m}\}$ and $\boldsymbol{W}_{XY,m} := \{(X_{i,m}, Y_{i,m} : 2t_m + 1 \leq i \leq \sigma_m)\}$. The purposes of these splits are different: the first two will be used to compute weights and the last one will be used to compute the U-statistic. In particular, as a weight function, consider a positive integer $1 + a_{xy,m} = (1 + a_{x,m})(1 + a_{y,m})$ where

$a_{x,m}$ is the number of occurrences x in $\boldsymbol{W}_{X,m}$ and, similarly, $a_{y,m}$ is the number of occurrences y in $\boldsymbol{W}_{Y,m}$. Next let $h^{m,\boldsymbol{a}}_{i_1,i_2,i_3,i_4}$ denote a weighted kernel function defined as

$$h^{m,\boldsymbol{a}}_{i_1,i_2,i_3,i_4} := \frac{1}{4!}\sum_{\pi\in\boldsymbol{\Pi}_4}\sum_{x\in[\ell_1],y\in[\ell_2]}\frac{\psi^m_{\pi(1)\pi(2)}(x,y)\psi^m_{\pi(3)\pi(4)}(x,y)}{1+a_{xy,m}}.$$

Given this kernel, we compute the weighted U-statistic for each $1\le m\le M$ as

$$U_W(\boldsymbol{W}_m) := \frac{1}{\binom{2t_m+4}{2}}\sum_{i_1<i_2<i_3<i_4:(i_1,i_2,i_3,i_4)\in\boldsymbol{W}_{XY,m}} h^{m,\boldsymbol{a}}_{i_1,i_2,i_3,i_4},$$

where $(i_1,i_2,i_3,i_4)\in\boldsymbol{W}_{XY,m}$ stands for taking four observations from $\boldsymbol{W}_{XY,m}$. Now, by letting $\omega_m := \sqrt{\min(\sigma_m,\ell_1)\min(\sigma_m,\ell_2)}$, the final test statistic for CI is defined as a weighted sum of $U_W(\boldsymbol{W}_1),\ldots,U_W(\boldsymbol{W}_M)$ given by

$$T_{\mathrm{CI},W} := \sum_{m\in[M]}\mathbb{1}(\sigma_m\ge 4)\sigma_m\omega_m U_W(\boldsymbol{W}_m). \tag{14}$$

Tests of Canonne et al. (2018). For both test statistics T_{CI} in (11) and $T_{\mathrm{CI},W}$ in (14), Canonne et al. (2018) suggest that one rejects the null when the test statistic is larger than $\zeta\sqrt{\min(n,M)}$ where ζ is a sufficiently large (but unspecified) constant. This cutoff value can be roughly understood as an upper bound of the standard deviation of the test statistic under the null. For ease of reference, we let $\phi_{\mathrm{CDKS},1} := \mathbb{1}(T_{\mathrm{CI}}\ge\zeta\sqrt{\min(n,M)})$ denote the test based on the unweighted statistic and similarly let $\phi_{\mathrm{CDKS},2} := \mathbb{1}(T_{\mathrm{CI},W}\ge\zeta\sqrt{\min(n,M)})$ denote the test based on the weighted test statistic computed based on $N\sim\mathrm{Pois}(n)$ samples. To describe their power results, let $\mathcal{P}_{[M]}$ be the set of discrete distributions defined on the support $[\ell_1]\times[\ell_2]\times[M]$. Moreover, let $\mathcal{P}_{0,[M]}\subset\mathcal{P}_{[M]}$ where $X\perp\!\!\!\perp Y|Z$ and $\mathcal{P}_{1,[M]} := \mathcal{P}_{[M]}\backslash\mathcal{P}_{0,[M]}$. Canonne et al. (2018) first consider the regime where ℓ_1 and ℓ_2 are fixed and ε satisfies a certain condition recalled in (56) of the Appendix, and then show that

$$\sup_{P_{X,Y,Z}\in\mathcal{P}_{1,[M]}:\inf_{Q\in\mathcal{P}_{0,[M]}}\mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z},Q)\ge\varepsilon}\mathbb{E}_{P^N_{X,Y,Z},N}[1-\phi_{\mathrm{CDKS},1}]\le\frac{1}{100}. \tag{15}$$

For example, the condition for ε is fulfilled when $\varepsilon\ge c\max\{M^{1/4}/n^{1/2},M^{7/8}/n,M^{3/4}/n^{7/8}\}$ for some large constant $c>0$. In the second regime where ℓ_1 and ℓ_2 can vary, the authors consider a more involved condition for ε depending on (ℓ_1,ℓ_2,M,n), and show that

$$\sup_{P_{X,Y,Z}\in\mathcal{P}_{1,[M]}:\inf_{Q\in\mathcal{P}_{0,[M]}}\mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z},Q)\ge\varepsilon}\mathbb{E}_{P^N_{X,Y,Z},N}[1-\phi_{\mathrm{CDKS},2}]\le\frac{1}{100}. \tag{16}$$

The condition for ε in this second regime is recalled in (52) of the Appendix for completeness.

Main results for the discrete case. In contrast to Canonne et al. (2018), we consider relatively more practical tests calibrated by the local permutation procedure, which do not rely on unspecified constants. We then argue that the permutation-based tests have the same theoretical guarantee as the tests of Canonne et al. (2018). As mentioned earlier, the local permutation test can control the type I error rate in the discrete setting without any further assumptions. Therefore our focus is on the power of the test. In the next remark, we explain a modified local permutation procedure, which we refer to as the "half-permutation" procedure, that facilitates the power analysis of the test based on $T_{\mathrm{CI},W}$.

Remark 6 (Full- versus. half-permutation)**.** For the weighted test statistic $T_{\mathrm{CI},W}$, there are two possible ways of calibrating the test via the local permutation procedure. The first one, we call "*full-permutation*", computes the p-value by permuting all Y labels within $\boldsymbol{W}_m$, independently, for each m. This is equivalent to the procedure described in Algorithm 1. The second one, we call "*half-permutation*", only permutes the Y labels within $\boldsymbol{W}_{XY,m}$, independently, for each m. Both approaches have finite-sample validity but the power of the first approach is intrinsically more difficult to analyze since each permutation destroys the independence structure among $\boldsymbol{W}_{X,m}$, $\boldsymbol{W}_{Y,m}$ and $\boldsymbol{W}_{XY,m}$. On the other hand, the half-permutation approach preserves the independence between $\boldsymbol{W}_{XY,m}$ and $\{\boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}\}$ even after permutations. Moreover, it has computational advantage over the full-permutation test since we do not need to recompute weights $1 + a_{xy,m}$ for each permutation. A similar strategy has been used in Kim et al. (2020, 2021) to analyze two-sample and (unconditional) independence tests.

We are now ready to state the main results of this subsection. As in Canonne et al. (2018), suppose that we draw N i.i.d. samples from $P_{X,Y,Z} \in \mathcal{P}_{[M]}$ where $N \sim \mathrm{Pois}(n)$. Given these samples, let $\phi_{\mathrm{perm},1}$ be the local permutation test based on the unweighted test statistic T_{CI} (11) through the full-permutation procedure described in Remark 6. Similarly, we let $\phi_{\mathrm{perm},2}$ be the local permutation test base on the weighted test statistic $T_{\mathrm{CI},W}$ (14) through the half-permutation procedure described in Remark 6. For both tests, we set the significance level $\alpha = 0.01$ for simplicity. These tests have the following guarantee on the type II error rate.

Theorem 5 (Type II error for discrete X, Y, Z)**.** *Consider the local permutation tests $\phi_{\mathrm{perm},1}$ and $\phi_{\mathrm{perm},2}$, described above. In the setting of discrete (X, Y, Z), $\phi_{\mathrm{perm},1}$ and $\phi_{\mathrm{perm},2}$ have the same type II error guarantee as in (15) and (16), respectively.*

The implications of Theorem 5 are as follows.

Remark 7.

- In Theorem 5, we set the type I error and the type II error by 1/100 for simplicity. In fact, 1/100 can be replaced with an arbitrarily small number by adjusting the constant factor in the condition for ε given in (52) of the Appendix.
- To make the given tests feasible for a fixed sample size, one can apply the truncation trick as in (12) and consider $\phi^{\dagger}_{\mathrm{perm},1} = \phi_{\mathrm{perm},1} \times \mathbb{1}(N \le n)$ and $\phi^{\dagger}_{\mathrm{perm},2} = \phi_{\mathrm{perm},2} \times \mathbb{1}(N \le n)$ where $N \sim \mathrm{Pois}(n/2)$. As discussed before, these modified tests have smaller type I errors than $\phi_{\mathrm{perm},1}$ and $\phi_{\mathrm{perm},2}$ based on $N \sim \mathrm{Pois}(n/2)$, respectively, and have the same power guarantee up to $e^{-n/8}$ factor. See equation (26) in the Appendix for more details.

- Canonne et al. (2018) further prove that the condition (52) in terms of n cannot be improved in certain regimes (depending on $\ell_1, \ell_2, M, \varepsilon$) by providing matching lower bounds. This together with Theorem 5 implies that the corresponding permutation test shares the same rate optimality as Canonne et al. (2018) whenever the tests of Canonne et al. (2018) are rate optimal in terms of the sample complexity. Despite the same optimality property, the permutation test may be more attractive than the corresponding test of Canonne et al. (2018) as it does not depend on an unspecified constant and it tightly controls the type I error rate in finite-sample settings.

- A major difficulty of proving Theorem 5 is in controlling randomness arising from the permutation procedure. We tackle this difficulty by building on the recent work of Kim et al. (2020). In particular, we derive an upper bound for the $1-\alpha$ quantile of the permutation distribution of the test statistic, which holds with high probability. More details can be found in Appendix A.6.

Next we switch gear to the continuous case of Z and develop similar results as in the discrete case.

5.2 Continuous Z

In this subsection, we build on the recent work of Neykov et al. (2021) and investigate the power of local permutation tests for continuous data. The idea of Neykov et al. (2021) is to carefully discretize Z into several bins and apply the tests based on T_{CI} and $T_{\mathrm{CI},W}$ as if the original data were discrete. Neykov et al. (2021) investigate the type I and II errors of these tests and prove that they are minimax optimal under certain smoothness conditions. However, their tests depend on unspecified constants and, in their simulations, the authors instead use the local permutation procedure to determine critical values. Hence there is a gap between theory and practice. The goal of this subsection is to close this gap by showing that the local permutation tests have the same power property as the tests considered in Neykov et al. (2021).

Discrete X, Y and continuous Z. First recall the setting described in Section 4.3 where X and Y are discrete random variables supported on $[\ell_1]\times[\ell_2]$ and Z is continuous and bounded between $[0,1]$. Denote by $\mathcal{P}_{[0,1]}$ the collection of distributions $P_{X,Y,Z}$ of such random variables. Let $\mathcal{P}_{0,[0,1]} \subset \mathcal{P}_{[0,1]}$ where $X \perp\!\!\!\perp Y | Z$ and $\mathcal{P}_{1,[0,1]} := \mathcal{P}_{[0,1]} \setminus \mathcal{P}_{0,[0,1]}$. Furthermore, let $\mathcal{P}_{1,[0,1],\mathrm{TV}}(L) \subset \mathcal{P}_{1,[0,1]}$ be the subset of alternative distributions that satisfies the TV smoothness:

$$\mathcal{D}_{\mathrm{TV}}(P_{X,Y|Z=z}, P_{X,Y|Z=z'}) \le L|z-z'| \quad \text{for all } z, z' \in [0,1]. \tag{17}$$

To describe the result of Neykov et al. (2021), draw $N \sim \mathrm{Pois}(n/2)$. If $N \le n$, we take a random subset of size N from $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and otherwise accept the null. Given that $N \le n$, compute the unweighted test statistic T_{CI} (11) using the binned data set $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M$ where $M = \lceil n^{2/5} \rceil$ and $\sum_{i=1}^M \sigma_i = N$. For a sufficiently large constant ζ depending on L, the corresponding test of Neykov et al. (2021) is defined as

$$\phi_{\mathrm{NBW},1} := \mathbb{1}(T_{\mathrm{CI}} \ge \zeta n^{1/5}) \times \mathbb{1}(N \le n).$$

In terms of the type II error, the authors show that there exists a sufficiently large constant c depending on $(\zeta, L, \ell_1, \ell_2)$, and for $\varepsilon \geq cn^{-2/5}$,

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1],\mathrm{TV}}(L): \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \geq \varepsilon} \mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi_{\mathrm{NBW},1}] \leq \frac{1}{100} + e^{-n/8}. \tag{18}$$

The type I error of $\phi_{\mathrm{NBW},1}$ is also guaranteed over a class of null distributions determined by certain smoothness conditions.

As mentioned before, a test based on the unweighted U-statistic may not perform well when ℓ_1 and ℓ_2 potentially increase with n. To avoid this issue, Neykov et al. (2021) follow the idea of Canonne et al. (2018) and propose another test based on the weighted test statistic $T_{\mathrm{CI},W}$ (14) where $M = \lceil \frac{n^{2/5}}{(\ell_1 \ell_2)^{1/5}} \rceil$. More formally, for a sufficiently large ζ depending on L, the second test is defined as

$$\phi_{\mathrm{NBW},2} := \mathbb{1}(T_{\mathrm{CI},W} \geq \sqrt{\zeta M}) \times \mathbb{1}(N \leq n). \tag{19}$$

Again, let c be a sufficiently large constant depending on (ζ, L) such that $\varepsilon \geq c\frac{(\ell_1 \ell_2)^{1/5}}{n^{2/5}}$. Under this condition for ε and an extra condition that $M\ell_1 \lesssim n$ for $\ell_1 \geq \ell_2$, Theorem 5.5 of Neykov et al. (2021) guarantees that

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1],\mathrm{TV}}(L): \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \geq \varepsilon} \mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi_{\mathrm{NBW},2}] \leq \frac{1}{100} + e^{-n/8}. \tag{20}$$

Furthermore, the authors prove that no CI test can be uniformly powerful in the TV distance when ε is much less than $\frac{(\ell_1 \ell_2)^{1/5}}{n^{2/5}}$. That means, $\phi_{\mathrm{NBW},1}$ and $\phi_{\mathrm{NBW},2}$ achieve minimax optimal rate, while the optimality of $\phi_{\mathrm{NBW},1}$ is only guaranteed when ℓ_1 and ℓ_2 are bounded. However the optimal power of $\phi_{\mathrm{NBW},2}$ over a broader regime comes at the cost of decreasing the size of null distributions. In fact, the type I error of $\phi_{\mathrm{NBW},2}$ is guaranteed over a χ^2-smooth class of null distributions, which is smaller than the class of null distributions considered for $\phi_{\mathrm{NBW},1}$. See Neykov et al. (2021) for more details.

Having described the results of Neykov et al. (2021), our aim is to reproduce the type II error guarantees (18) and (20) based on the same test statistics but with (explicit) cutoff values determined by the local permutation procedure. Given the data set of size $N \leq n$ where $N \sim \mathrm{Pois}(n/2)$, let us denote the local permutation tests based on T_{CI} and $T_{\mathrm{CI},W}$ by $\phi_{\mathrm{perm},1}$ and $\phi_{\mathrm{perm},2}$, respectively. As described in Remark 6, it greatly simplifies the power analysis for $T_{\mathrm{CI},W}$ when we consider the half-permutation method. Hence, unlike $\phi_{\mathrm{perm},1}$ that builds on Algorithm 1, the p-value of $\phi_{\mathrm{perm},2}$ is determined by the half-permutation method. Now in order to take into account the random sample size N, the final test functions are defined as $\phi^\dagger_{\mathrm{perm},1} := \phi_{\mathrm{perm},1} \times \mathbb{1}(N \leq n)$ and $\phi^\dagger_{\mathrm{perm},2} := \phi_{\mathrm{perm},2} \times \mathbb{1}(N \leq n)$. These tests have the following guarantee on the type II error rate.

Theorem 6 (Type II error for discrete X, Y and continuous Z)**.** *Consider the local permutation tests $\phi^\dagger_{\mathrm{perm},1}$ and $\phi^\dagger_{\mathrm{perm},2}$, described above. In the setting of discrete (X, Y) and continuous Z, $\phi^\dagger_{\mathrm{perm},1}$ and $\phi^\dagger_{\mathrm{perm},2}$ have the same type II error guarantee as in (18) and (20), respectively.*

Several remarks are provided below.

Remark 8.

- We highlight once again that the local permutation tests $\phi^{\dagger}_{\mathrm{perm},1}$ and $\phi^{\dagger}_{\mathrm{perm},2}$ do not require the knowledge on unspecified constant ζ in $\phi_{\mathrm{NBW},1}$ and $\phi_{\mathrm{NBW},2}$, while achieving the same power (in terms of rate) as stated in Theorem 6.
- However, such a nice property does not come for free. In general, the local permutation test requires a stronger condition on the class of null distributions than the corresponding theoretical test for type I error control. For instance, under the class of null distributions $\mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)$ in Definition 3, we require that $n^{1/2}h^{\gamma} \to 0$ for $\gamma \in [1,2]$ and $n^{1/2}h^2 \to 0$ for $\gamma \geq 2$ for the local permutation test to be valid (Theorem 2). With the choice of $M = \lceil n^{2/5} \rceil$ (recall that $M = h^{-1}$), for instance, we have only shown that $\phi^{\dagger}_{\mathrm{perm},1}$ is valid when $\gamma > 5/4$. On the other hand, $\phi_{\mathrm{NBW},1}$ can control the type I error even when $\gamma = 1$ (see Theorem 5.2 of Neykov et al., 2021). Similarly, in order for $\phi^{\dagger}_{\mathrm{perm},2}$ to be valid under the same χ^2-smoothness condition in Theorem 5.5 of Neykov et al. (2021), we require that $n^{1/2}h^2 = \frac{(\ell_1\ell_2)^{2/5}}{n^{3/10}} \to 0$, which was not needed for $\phi_{\mathrm{NBW},2}$. In Section 6, we attempt to partly address this drawback by introducing a novel double-binning strategy, which requires less stringent conditions for type I error control.
- The proof of Theorem 6 is similar to that of Theorem 5, that is, we show that the $1-\alpha$ quantile of the permutation distribution of T_{CI} is upper bounded by $\zeta n^{1/5}$ with high probability under the alternative. This result yields that the type II error of $\phi^{\dagger}_{\mathrm{perm},1}$ is upper bounded by that of $\phi_{\mathrm{NBW},1}$, up to a small error term. From here, we can directly benefit the previous bounds (18) and (20) and show that the type II error of $\phi^{\dagger}_{\mathrm{perm},1}$ is small. The proof for $\phi^{\dagger}_{\mathrm{perm},2}$ follows similarly. The details can be found in Appendix A.7.

Continuous X, Y, Z. Next we develop a similar result for the case where (X, Y, Z) is supported on $[0,1]^3$ with a joint distribution absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{P}_{[0,1]^3}$ be the set of distributions of such random variables and $\mathcal{P}_{0,[0,1]^3} \subset \mathcal{P}_{[0,1]^3}$ for which $X \perp\!\!\!\perp Y | Z$. Let $\mathcal{P}_{1,[0,1]^3,\mathrm{TV}}$ be the subset of $\mathcal{P}_{1,[0,1]^3} := \mathcal{P}_{[0,1]^3} \setminus \mathcal{P}_{0,[0,1]^3}$, which satisfies the TV smoothness condition in (17). In addition, we assume that, for any $P_{X,Y,Z} \in \mathcal{P}_{1,[0,1]^3,\mathrm{TV}}$, the corresponding conditional density function $p_{X,Y|Z}$ given any $z \in [0,1]$ belongs to $\mathcal{H}^{2,s}(L)$, where $\mathcal{H}^{2,s}(L)$ is the class of Hölder smooth functions $[0,1]^2 \mapsto \mathbb{R}$ with parameter s defined in Definition 5 of the Appendix.

In order to apply $T_{\mathrm{CI},W}$ to continuous data, we need to further discretize X, Y into several bins. For this purpose, for a given $M > 0$ (specified in the sequel), consider a partition of $[0,1]$ into $M' = \lceil M^{1/s} \rceil$ bins of equal size denoted by $\{B'_m\}_{m=1}^{M'}$. We then transform X (and similarly Y) through the map $g : [0,1] \mapsto \{1, \ldots, M'\}$ by defining $g(x) = m$ if and only if $x \in B_m$. On the other hand, we partition the conditional variable Z into M bins of equal size. Given this binned data set, we implement $\phi_{\mathrm{NBW},2}$ in (19), but with a different choice of $M = \lceil n^{2s/(5s+2)} \rceil$. According to Theorem 5.6 of Neykov et al. (2021), the resulting test achieves minimax optimal power. In particular, they show that there exists a sufficiently large constant c depending on ζ, L, and if $\varepsilon \geq cn^{-2s/(5s+2)}$, then

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1]^3,\mathrm{TV}}(L) : \inf_{Q \in \mathcal{P}_{0,[0,1]^3}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \geq \varepsilon} \mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi_{\mathrm{NBW},2}] \leq \frac{1}{100} + e^{-n/8}. \tag{21}$$

We now show that the corresponding local permutation test $\phi^{\dagger}_{\mathrm{perm},2}$ has the same power property.

Theorem 7 (Type II error for continuous X, Y, Z)**.** *Consider the local permutation test $\phi^{\dagger}_{\mathrm{perm},2}$ applied to the discretized data set described above. In the setting of continuous (X, Y, Z), $\phi^{\dagger}_{\mathrm{perm},2}$ has the same type II error guarantee as in (21) and thereby shares the same optimal power as $\phi_{\mathrm{NBW},2}$.*

The same points in Remark 8 apply to Theorem 7. While $\phi^{\dagger}_{\mathrm{perm},2}$ has the same optimality as $\phi_{\mathrm{NBW},2}$ in terms of power, we need to restrict the class of null distributions further to rigorously control the type I error. In particular, under γ-Hellinger Lipschitzness with $\gamma \geq 2$, the underlying conditional density function should be smooth enough to ensure that $n^{1/2}h^2 = n^{-4s/(5s+2)+1/2} \to 0$, equivalently $s > 2/3$, for type I error control. As we discussed in more detail in the introduction, the tension in CI testing between tightly controlling the type I error (requiring narrow bins) and ensuring high power (requiring, in some cases, wider bins) is a unique feature of CI testing. In the next section, we introduce a novel double-binning permutation test that allows us to consider less smooth null distributions, while maintaining the power (up to a constant factor).

6 Double-binning strategy

As we discussed above, the type I error of the local permutation test is guaranteed to be small over a smaller set of null distributions than the conservatively calibrated U-statistic test used by Neykov et al. (2021). The reason for this gap is that the permutation approach relies on an additional condition for its validity. In particular, it requires that the binned distribution $Q^n_{X,Y,\widetilde{Z}}$ and its CI projection $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$ be close in the TV distance. The goal of this section is to mitigate this issue via double-binning. The idea is to consider bins of two distinct resolutions where a test statistic is computed over coarser bins, whereas the permutation procedure is implemented over finer bins. This double-binning strategy allows us to keep the TV distance smaller than the single-binning approach, while maintaining similar power under certain conditions.

To elaborate on the idea, recall that $\{B_1, \ldots, B_M\}$ is a partition of $\mathcal{Z}$. For $m = 1, \ldots, M$, let us further partition B_m into b bins, which results in smaller bins $\{B_{m,1}, \ldots, B_{m,b}\}$. For $(m, k) \in [M] \times [b]$, let $\sigma_{m,k}$ be the sample size within $B_{m,k}$, and $\boldsymbol{W}_{m,k}$ denote the set of the pairs of (X_i, Y_i) that belong to $B_{m,k}$. More formally, by letting $(X_{i,m,k}, Y_{i,m,k})$ be the ith pair in $B_{m,k}$, we write $\boldsymbol{W}_{m,k} := \{(X_{1,m,k}, Y_{1,m,k}), \ldots, (X_{\sigma_{m,k},m,k}, Y_{\sigma_{m,k},m,k})\}$ when $\sigma_{m,k} \geq 1$ and otherwise $\boldsymbol{W}_{m,k} = \emptyset$. Under this setting, we compute the test statistic T_{CI} as in (1) based on the larger bins $\{B_1, \ldots, B_M\}$. The permuted test statistic $T^{\boldsymbol{\pi}}_{\mathrm{CI}}$ is computed similarly as before except now that Y values are permuted within the smaller bins. We then compute the permutation p-value as in (3) by counting how many permuted statistics are larger than or equal to T_{CI}.

Here, to simplify our theoretical analysis, we focus our attention on a subset of all possible local permutations. In particular, within each small bin $B_{m,k}$ for $(m, k) \in [M] \times [b]$, we only consider a set of cyclic permutations of $\{1, \ldots, \sigma_{m,k}\}$. As an illustration, when $\sigma_{m,k} = 4$, we have four distinct cyclic permutations of $\{1, 2, 3, 4\}$ as $\{1, 2, 3, 4\}, \{2, 3, 4, 1\}, \{3, 4, 1, 2\}, \{4, 1, 2, 3\}$. A formal definition of a cyclic permutation can be found in Definition 6. This cyclic restriction results in $K_* := \prod_{(m,k)\in[M]\times[b]} \max\{\sigma_{m,k}, 1\}$ number of local permutations, and the set of all possible such cyclic

local permutations is denoted by $\mathbf{\Pi}_{\text{cyclic}}$. Given this notation, we describe the local permutation procedure via double-binning in Algorithm 2. We also refer to Figure 2 for an illustration of the procedure.

Algorithm 2 Local permutation procedure via double-binning

Input: data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, a super-partition of $\mathcal{Z}$: $\{B_1, \ldots, B_M\}$, sub-partitions of each B_m : $\{B_{m,1}, \ldots, B_{m,b}\}$ for $m = 1, \ldots, M$, a test statistic T_{CI}, a nominal level α

1. For each $\boldsymbol{\pi} \in \mathbf{\Pi}_{\text{cyclic}}$, compute $T_{\text{CI}}^{\boldsymbol{\pi}}$ as in (2) and denote the resulting statistics by $T_{\text{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\text{CI}}^{\boldsymbol{\pi}_{K_*}}$.
2. By comparing the statistic T_{CI} in (1) with the permuted ones, calculate the p-value as
$$p_{\text{perm}} = \frac{1}{K_*} \sum_{\boldsymbol{\pi}_i \in \mathbf{\Pi}_{\text{cyclic}}} \mathbb{1}\{T_{\text{CI}}^{\boldsymbol{\pi}_i} \geq T_{\text{CI}}\}.$$
3. Given the nominal level $\alpha \in (0, 1)$, define the test function $\phi_{\text{perm},n} = \mathbb{1}(p_{\text{perm}} \leq \alpha)$ and reject the null when $\phi_{\text{perm},n} = 1$.

Next, we discuss the type I and type II errors of the local permutation test via double-binning. The analysis of the type I error is relatively straightforward. Indeed, all of the results in Section 4.2 continue to hold for this new approach but now the maximum diameter (7) becomes smaller as it is defined over the finer bins. Therefore, in view of Theorem 2 and Theorem 3, an upper bound for the type I error can be much tighter than the single-binning approach based on $\{B_1, \ldots, B_M\}$. The only caveat, here, is that we consider the set of cyclic local permutations $\mathbf{\Pi}_{\text{cyclic}}$, rather than all possible local permutations. However, this change does not affect the validity.

Proposition 1 (Type I error of the double-binning test)**.** *Let $\phi_{\text{perm},n}$ be the local permutation test via double-binning in Algorithm 2. Then the same bounds (8) and (9) hold for $\phi_{\text{perm},n}$ with the maximum diameter h defined as*
$$h = \max_{(m,k) \in [M] \times [b]} \sup_{z, z' \in B_{m,k}} \delta(z, z').$$

We note that the above result holds under Poissonization in a similar fashion to Proposition 2 in the Appendix. Indeed, once the bounds (8) and (9) are given, the validity of the double-binning test under Poissonization can be proved along the same lines of the proof of Proposition 2. Given the above proposition, our main concern is the type II error. Here, to illustrate ideas, we only focus on the case where $(X, Y, Z) \in [\ell_1] \times [\ell_2] \times [0, 1]$ and ℓ_1, ℓ_2 are fixed. In addition, we recall the definition of $\mathcal{P}_{1,[0,1],\text{TV}}(L)$ from Section 5.2 and consider a subset of $\mathcal{P}_{1,[0,1],\text{TV}}(L)$, denoted by $\mathcal{P}'_{1,[0,1],\text{TV}}(L)$, such that the marginal density of Z is bounded below by some fixed constant $c_{\text{low}} > 0$, i.e. $p_Z(z) \geq c_{\text{low}}$ for all $z \in [0, 1]$. Furthermore, we assume that any distribution in $\mathcal{P}'_{1,[0,1],\text{TV}}(L)$ satisfies
$$\mathcal{D}_{\text{TV}}(P_{X|Z=z}, P_{X|Z=z'}) \leq L|z - z'| \quad \text{and} \quad \mathcal{D}_{\text{TV}}(P_{Y|Z=z}, P_{Y|Z=z'}) \leq L|z - z'|,$$

for all $z, z' \in [0,1]$.

To describe type II error results, we consider the test statistic T_{CI} used in $\phi_{\mathrm{NBW},1}$. Moreover, assume that each of the coarser bins and the finer bins has the same length of an interval $1/M$ and $1/(Mb)$, respectively. Now, given a sample $\{(X_i, Y_i, Z_i)\}_{i=1}^N$ where $N \sim \mathrm{Pois}(n/2)$, we compute $\phi_{\mathrm{perm},N}$ based on T_{CI} using Algorithm 2 at significance level $\alpha = 0.01$ (for simplicity) and define $\phi^{\dagger}_{\mathrm{perm},n} := \phi_{\mathrm{perm},N} \times \mathbb{1}\{N \le n\}$ as in (12). Under this setting, the resulting test has the following type II error guarantee.

Theorem 8 (Type II error of the double-binning test)**.** *Consider the test $\phi^{\dagger}_{\mathrm{perm},n}$ defined above with the number of larger bins $M = \lceil n^{2/5} \rceil$. Suppose that we choose b such that $n/(Mb)^2 > 400 c_{\mathrm{low}}^{-1}$ and $Mb \to \infty$ as $n \to \infty$. Suppose further that $\varepsilon \ge c n^{-2/5}$ for a sufficiently large c depending on (L, ℓ_1, ℓ_2). Then*

$$\sup_{P_{X,Y,Z} \in \mathcal{P}'_{1,[0,1],\mathrm{TV}}(L): \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \ge \varepsilon} \mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi^{\dagger}_{\mathrm{perm},n}] \le \frac{1}{100} + e^{-n/8} + \rho_n,$$

where ρ_n is a positive sequence converging to zero as $n \to \infty$.

We first note that an explicit form of ρ_n in the above result can be found in (70) given in the Appendix. Next, observe that the above double-binning test achieves the same minimax separation rate as $\phi_{\mathrm{NBW},1}$ without being dependent on an unspecified constant in its critical value. Its guarantee holds over a smaller set of alternative distributions, namely $\mathcal{P}'_{1,[0,1],\mathrm{TV}}(L)$ (in contrast, the result of Neykov et al. (2021) does not require any lower bound on the density of Z).

On the other hand, compared to the corresponding permutation test via single-binning, the double-binning method controls the type I error rate over a larger class of null distributions without sacrificing power up to a constant factor. In particular, with an optimal choice of $M = \lceil n^{2/5} \rceil$, the single-binning test requires $\gamma > 5/4$ in order to control the type I error under γ-Hellinger smoothness. In contrast, the double-binning test is valid for $\gamma > 1$ as long as b is chosen appropriately such that $n/(Mb)^{2\gamma} \to 0$ but $n/(Mb)^2$ is bounded below by a large constant.

Remark 9 (Extension to other settings)**.** It is worth highlighting that the validity result of the double-binning test in Proposition 1 holds universally for any binning-based test statistic. Theorem 8 regarding the power, on the other hand, is based specifically on the test statistic T_{CI} used in $\phi_{\mathrm{NBW},1}$. While we believe that a similar power result can be developed based on other test statistics including $T_{\mathrm{CI},W}$ used in $\phi_{\mathrm{NBW},2}$, a detailed treatment of this direction is beyond the scope of this paper. Another important direction one can pursue is to see whether the lower bound restriction on the density of Z can be removed in Theorem 8. We leave these topics to future work.

7 Simulations

In this section, we illustrate the numerical performance of the local permutation test through Monte Carlo simulations. Throughout our experiments, we use Monte Carlo simulations to compute the permutation p-value, defined as p_{perm} in (3). In particular, we draw B permutations, denoted by

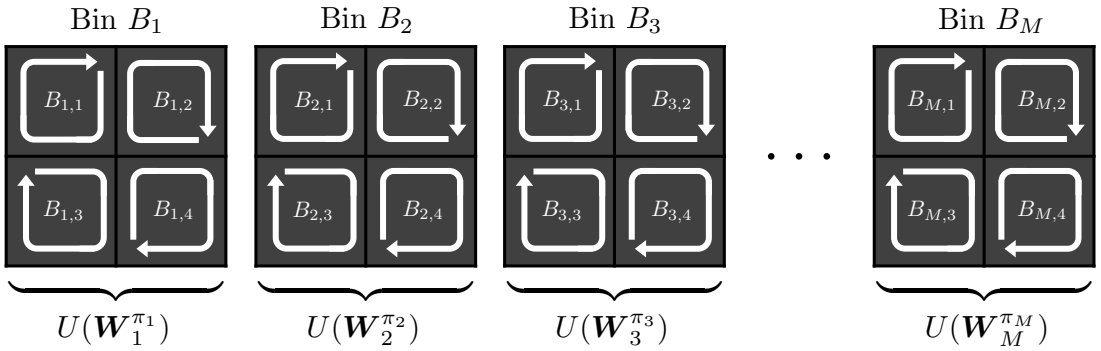


Figure 2: Illustration of the double-binning strategy where we consider bins of two distinct resolutions. The purpose of coarser bins $\{B_1, \ldots, B_M\}$ is to compute a test statistic, whereas the purpose of finer bins $\{B_{1,1}, \ldots, B_{M,4}\}$ is to perform the permutation procedure. In particular, we only permute the observations within the finer bins. This strategy allows us to keep the type I error under control over a larger class of null distributions, while maintaining the competitive power under certain conditions.

$\boldsymbol{\pi}'_1, \ldots, \boldsymbol{\pi}'_B$, from $\boldsymbol{\Pi}$ with replacement. Then the permutation p-value using the test statistic T_{CI} is defined as

$$\widehat{p}_{\mathrm{perm}} := \frac{1}{B+1}\left[\sum_{i=1}^{B} \mathbb{1}\{T_{\mathrm{CI}}^{\boldsymbol{\pi}'_i} \geq T_{\mathrm{CI}}\} + 1\right]. \tag{22}$$

It is well-known that $\mathbb{1}(\widehat{p}_{\mathrm{perm}} \leq \alpha)$ is a valid test in finite-sample scenarios, whenever $T_{\mathrm{CI}}, T_{\mathrm{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\mathrm{CI}}^{\boldsymbol{\pi}_K}$ are exchangeable under the null. Furthermore, $\widehat{p}_{\mathrm{perm}}$ can be arbitrarily close to p_{perm} for a sufficiently large B, and we take $B = 100$ in our simulations. We also note that the type I error and the power presented in this section are approximated by repeating simulations 1000 times at significance level $\alpha = 0.05$.

7.1 Experiment 1

In our first experiment, we demonstrate Theorem 1 in a discrete setting of (X, Y, Z) where (X, Y, Z) is distributed over $[\ell_1] \times [\ell_2] \times [M]$ and $\ell_1 = \ell_2 = 2$ and $M \in \{10, 20, \ldots, 110, 120\}$. In particular, we let Z have a multinomial distribution with equal probabilities over the bins. Similarly, we let X have a multinomial distribution with equal probabilities, and set $X = Y$ and $(X, Y) \perp\!\!\!\perp Z$. We are under the alternative hypothesis where X and Y are perfectly correlated conditional on Z. In this setting, we compute the empirical power of the local permutation test based on the test statistic T_{CI} considered in Theorem 5 by varying (M, n). The result can be found in the left panel of Figure 3. As predicted by Theorem 1, we see that the power of the test degrades quickly as M increases for any given n. This in turn illustrates that CI testing is impossible unless the probability observing the same value of Z is properly controlled.

7.2 Experiment 2

In our second experiment, we demonstrate the validity result of the local permutation test in Section 4. To generate the data, we consider a distributional setup used in construction of our lower bound result (Theorem 4). In particular, we consider the marginal density of Z in (25) with $\epsilon = n^{-1}$, and the conditional density of $X|Z = z$ in (33). We further let $Y|Z = z$ have the same conditional density as $X|Z = z$ for all $z \in [0, 1]$, while satisfying $X \perp\!\!\!\perp Y|Z$. As proved in Appendix A.5.1 and Appendix A.5.3, the considered distribution satisfies γ-Hellinger Lipschitzness with any fixed $\gamma \geq 1$ as well as γ-Rényi Lipschitzness with any fixed $\gamma > 0$. Therefore, by Theorem 2 and Theorem 3, the type I error of the local permutation test based on any test statistic is approximately α as long as $nM^{-4} \to 0$. We also note from Theorem 4 that there exists a test statistic such that the corresponding local permutation test fails to control the type I error rate when $nM^{-4} \to \infty$. To demonstrate both results, we use the test statistic in (11) by varying the number of bins $M \in \{n, \lceil n^{1/2} \rceil, \lceil n^{1/4} \rceil, \lceil n^{1/10} \rceil\}$. The result is given in the right panel of Figure 3. As we can see from the result, the type I error is well controlled when M is chosen such that $nM^{-4} \to 0$. On the other hand, the error tends to increase when $nM^{-4} \to \infty$, which coincides with our theory.

7.3 Experiment 3

In our third experiment, we illustrate type I and II error control of the double-binning test in Theorem 8 by setting $M = b = \lceil n^{2/5} \rceil$. The permutation p-value of the double-binning procedure is approximated similarly as $\widehat{p}_{\text{perm}}$ in (22) but by drawing B cyclic permutations from $\mathbf{\Pi}_{\text{cyclic}}$ without replacement. As before, we choose $B = 100$ for our third simulation as well. To demonstrate the performance, we let Z have a uniform distribution over the interval $[0, 1]$ and X, Y be Bernoulli random variables with the following conditional probability mass functions:

$$p_{X|Z}(X = 1|z) = p_{Y|Z}(Y = 1|z) = e^{\sin(\theta z)}/4. \tag{23}$$

The considered distribution depends on the parameter θ, which controls the smoothness of the conditional probability mass function. In particular, the conditional marginals (23) become more wiggly as θ increases, which makes it more difficult to control the type I error under the null.

(a). **Type I error.** To illustrate type I error control, we consider the null distribution with the conditional marginals (23). We then draw $n = 100$ samples from the null distribution, and compute the test statistic T_{CI} as well as the p-value. The finite-sample type I error is approximated by Monte Carlo simulations and the result is provided in the left panel of Figure 4. As a reference point, we also consider the single-binning test based on the same test statistic with $M = \lceil n^{2/5} \rceil$ and its type I error rate is also provided in the left panel of Figure 4. From the result, we see that the type I error of the single-binning test increases with θ much faster than that of the double-binning test. This empirical result supports Proposition 1 that claims that the double-binning test is valid over a larger class of null distributions than the corresponding single-binning test.

(b). **Power.** To illustrate the power performance, we consider an alternative distribution with the

same marginals (23) with $\theta = 1$. In particular, by writing $f(z) = e^{\sin(z)}/4$, we set

$$p_{XY|Z}(X = 1, Y = 1|z) = f(z)^2 + f(z)/5,$$
$$p_{XY|Z}(X = 0, Y = 0|z) = \{1 - f(z)\}^2 + f(z)/5 \quad \text{and}$$
$$p_{XY|Z}(X = 1, Y = 0|z) = p_{XY|Z}(X = 0, Y = 1|z) = 4f(z)/5 - f(z)^2.$$

One can check that the above joint distribution is a valid alternative distribution where the conditional joint distribution differs from the product of the conditional marginal distributions. With n draws from the above distribution, we compute the same test statistic and p-value as before and approximate the power of the test by changing $n \in \{30, 50, 100, 150, \ldots, 350, 400\}$. The right-panel of Figure 4 collects the power approximates for both single-binning and double-binning tests. Overall, the power of the single-binning test is higher than that of the double-binning test, but the difference seems marginal, especially when the power is close to one. This may be viewed as empirical evidence of Theorem 8, which shows that both tests have the same power up to a constant factor in certain regimes.

In both the panels of Figure 4, we also present the type I error and power of the corresponding tests under Poisson sampling. Specifically, for $N \sim \text{Poisson}(n)$, we draw i.i.d. samples (X^N, Y^N, Z^N) from the joint distribution of (X, Y, Z) and compute each permutation test using (X^N, Y^N, Z^N). As we can see, the results under Poisson sampling are not significantly different from the previous results with the fixed sample size n, and we anticipate that our theoretical results will continue to hold with a fixed sample-size.

8 Discussion

In this paper, we investigated several statistical properties of the local permutation method for CI testing. We started by presenting a new hardness result of CI testing, which, along with the recent work of Shah and Peters (2020), motivates us to consider reasonable assumptions under which CI testing becomes possible. Under certain smoothness assumptions, we provided upper bounds for the type I error of the local permutation test and further showed that these bounds are tight in some cases. Turning to the power, we demonstrated that the local permutation test can retain minimax power, while rigorously controlling the type I error, under certain circumstances. In particular, we showed that the local permutation tests using the same test statistics in Canonne et al. (2018); Neykov et al. (2021) have the same power guarantee. However, compared to the previous tests, the type I error of the local permutation test is guaranteed over a smaller set of null distributions in the continuous case of Z. To this end, we introduced and analyzed a double-binning strategy, which mitigates this drawback.

Future directions. We close by discussing several interesting directions for future work.

- **Adaptive binning strategy.** Throughout this paper, we have been mainly concerned with equal-sized bins. This strategy, as we saw earlier, can lead to optimal CI tests from a minimax perspective. However, when there exists a local structure on the distribution of Z, many of bins would be empty. In this case, it would be more desirable to consider an adaptive binning

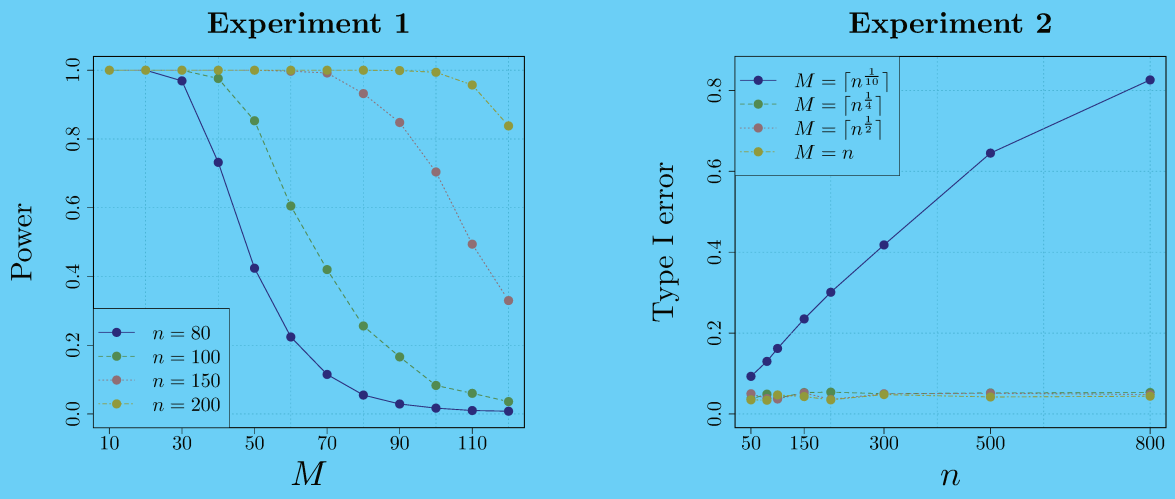


Figure 3: Illustration of the power and type I error of the local permutation test using T_{CI}. The left panel demonstrates that the power of the permutation test keeps decreasing as M increases even though X and Y are perfectly correlated conditional on Z; thereby, confirming Theorem 1. The right panel demonstrates that the type I error of the local permutation test is well-controlled under the smoothness condition, described in Section 7.2, unless M is chosen such that $nM^{-4} \to \infty$.

scheme that uses different sizes of bins over different regions of Z. This strategy, potentially data-dependent, requires a more delicate analysis for both type I and type II error control, which we leave to future work. As mentioned before, it would also be interesting to see whether it is possible to develop an adaptive test to the unknown smoothness parameters without sacrificing power much.

- **Different metrics.** In this work, we have focused on the smoothness conditions based on the generalized Hellinger distance and Rényi divergence. It may be possible to obtain different and potentially sharper validity conditions by considering other metrics. In addition, one can impose a smoothness assumption on higher order derivatives of a conditional distribution and see whether its improves the validity result. It is also worth investigating the minimax power of the local permutation test in different metrics other than the TV distance.

- **Other test statistics.** While our validity result can be applied to any binning-based statistic, the power analysis was specifically based on U-statistics with discrete-type kernels. We believe that it is also possible to obtain similar minimax power results using other test statistics. In particular, exploring the power of RKHS-based test statistics (e.g. Fukumizu et al., 2008) is a promising direction for future work. This can be potentially explored by building on the recent work of Kim et al. (2020); Meynaoui et al. (2019) that investigate the minimax power of unconditional independence tests based on Gaussian kernels.

- **Depoissonization.** As in the previous work (Balakrishnan and Wasserman, 2019; Canonne et al., 2018; Neykov et al., 2021), it was crucial to use Poissonization technique for our power

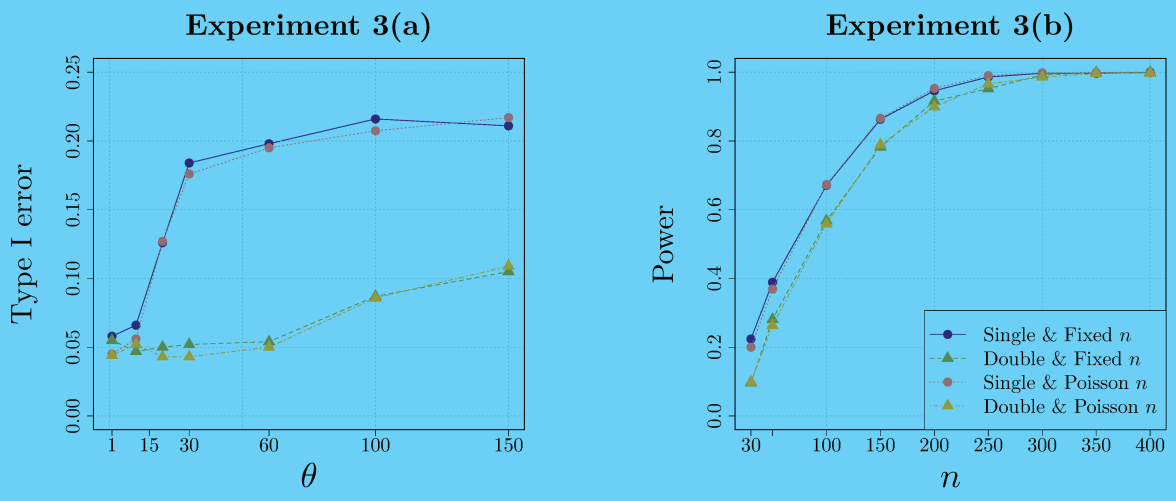


Figure 4: Illustration of the type I error and power of the single- and double-binning tests using T_{CI}. The left-panel considers the null distribution with conditional marginals (23) and shows that the double-binning test has better type I error control than the single-binning test over different values of smoothness parameter θ. The right-panel is concerned with the power as n varies, demonstrating that the single-binning performs better than the double-binning test in terms of power. We also include simulation results under Poissonization, which illustrates that the Possionization version performs similarly as the fixed-n counterpart. A more detailed explanation can be found in Section 7.3.

analysis. Since a Poisson random variable is tightly concentrated around its mean, it sounds plausible that the same power guarantee can be achieved by the local permutation test without Poissonization as empirically demonstrated in Section 7.3. However, a formal proof is not available at the current stage, which we leave as future work.

- **Multivariate Z.** Our results on type I error control show that the validity of a local permutation test crucially relies on the maximum diameter of bins, which is well-defined even when Z is a multivariate random vector. Our results on minimax power, on the other hand, only deal with the univariate case of Z. Indeed, tight minimax separation rates for CI testing are only known for the case when the dimension of Z is either one or two (Neykov et al., 2021). As the purpose of our work is to demonstrate that local permutation tests can achieve the same optimal power as their theoretical counterparts, we focus only on the univariate Z case. It therefore remains for future work to establish minimax separation rates for a general multivariate case and see whether local permutation tests can achieve these rates.

- **Theoretical vs. practical calibration.** While the double-binning strategy allows us to consider less smooth null distributions than the corresponding single-binning test, there still remain settings where minimax-optimal tests are available (Neykov et al., 2021) but where we do not have practical methods to calibrate these optimal tests. Understanding these gaps and more clearly establishing the fundamental limits for *practical* conditional independence tests are an important direction for future work.

Acknowledgements

This work was partially supported by funding from the NSF grants DMS-1713003, DMS-2113684 and CIF-1763734, as well as Amazon AI and a Google Research Scholar Award to SB. IK was partially supported by EPSRC grant EP/N031938/1 and Yonsei University Research Fund of 2021-22-0332.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Azadkia, M. and Chatterjee, S. (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*.

Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749.

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.

Barber, R. F., , Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*.

Barber, R. F. (2020). Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487–3524.

Bellot, A. and van der Schaar, M. (2019). Conditional independence testing using generative adversarial networks. *arXiv preprint arXiv:1907.04068*.

Bergsma, W. (2010). Nonparametric testing of conditional independence by means of the partial copula. *Available at SSRN 1702981*.

Bergsma, W. P. (2004). *Testing conditional independence for continuous random variables*. Citeseer.

Berrett, T. B., Kontoyiannis, I., and Samworth, R. J. (2020a). Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics (to appear)*.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020b). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197.

Candés, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Canonne, C. L. (2017). A short note on poisson tail bounds. http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57. IEEE.

Dai, B., Shen, X., and Pan, W. (2021). Significance tests of feature relevance for a blackbox learner. *arXiv preprint arXiv:2103.04985*.

Daudin, J. (1980). Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590.

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.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A permutation-based kernel conditional independence test. In *UAI*, pages 132–141.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

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.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23(2):169–192.

Huang, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091.

Huang, Z., Deb, N., and Sen, B. (2020). Kernel Partial Correlation Coefficient–a Measure of Conditional Dependence. *arXiv preprint arXiv:2012.14804*.

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

Kamps, U. (1989). Hellinger distances and α-entropy in a one-parameter class of density functions. *Statistical Papers*, 30(1):263–269.

Kim, D. and Agresti, A. (1997). Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Computational Statistics & Data Analysis*, 24(1):89–104.

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli*, 27(1):419–441.

Kim, I., Balakrishnan, S., and Wasserman, L. (2020). Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*.

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434.

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

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

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

Li, L., Tchetgen, E. T., van der Vaart, A., and Robins, J. M. (2011). Higher order inference on a treatment effect under low regularity conditions. *Statistics & probability letters*, 81(7):821–828.

Lundborg, A. R., Shah, R. D., and Peters, J. (2021). Conditional Independence Testing in Hilbert Spaces with Applications to Functional Data Analysis. *arXiv preprint arXiv:2101.07108*.

Margaritis, D. (2005). Distribution-free learning of bayesian network structure in continuous domains. In *AAAI*, volume 5, pages 825–830.

Marx, A. and Vreeken, J. (2019). Testing conditional independence on discrete data using stochastic complexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 496–505. PMLR.

Meynaoui, A., Albert, M., Laurent, B., and Marrel, A. (2019). Adaptive test of independence based on hsic measures. *arXiv preprint arXiv:1902.06441*.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

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

Park, J. and Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33.

Patra, R. K., Sen, B., and Székely, G. J. (2016). On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Petersen, L. and Hansen, N. R. (2021). Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research*, 22(70):1–47.

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics.

Runge, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR.

Sason, I. and Verdú, S. (2016). f-divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2021). MMD Aggregated Two-Sample Test. *arXiv preprint arXiv:2110.15073*.

Sen, R., Shanmugam, K., Asnani, H., Rahimzamani, A., and Kannan, S. (2018). Mimic and classify: A meta-algorithm for conditional independence testing. *arXiv preprint arXiv:1806.09708*.

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-powered conditional independence test. In *Advances in neural information processing systems*, pages 2951–2961.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Sheng, T. and Sriperumbudur, B. K. (2019). On distance and kernel measures of conditional independence. *arXiv preprint arXiv:1912.01103*.

Shi, C., Xu, T., Bergsma, W., and Li, L. (2020). Double Generative Adversarial Networks for Conditional Independence Testing. *arXiv preprint arXiv:2006.02615*.

Song, K. (2009). Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics*, 37(6B):4011–4045.

Spohn, W. (1994). On the properties of conditional independence. In *Patrick suppes: Scientific philosopher*, pages 173–196. Springer.

Stam, A. J. (1978). Distance between sampling with and without replacement. *Statistica Neerlandica*, 32(2):81–91.

Strobl, E. V., Zhang, K., and Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1).

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

Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

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

Zhang, H., Zhou, S., and Guan, J. (2018). Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 804–813.

Zhou, Y., Liu, J., and Zhu, L. (2020). Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis*, 175:104557.

A Proofs of the results in the main text

In this section, we collect the proofs of the results in the main text. Some auxiliary lemmas required for the proofs can be found in Appendix B. We start by introducing additional notation that simplifies our presentation.

Additional notation. For sequences a_n and b_n, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists an absolute constant $C > 0$ such that $a_n \leq C b_n$ for all n. In addition, $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$. As convention, $\|\cdot\|_p$ represents the L_p norm. $C, C_1, C_2, \ldots$ denote some constants whose value may differ in different places.

A.1 Proof of Theorem 1

As mentioned in the main text, the proof of this result is highly motivated by the impossibility result of distribution-free conditional predictive inference (see e.g., Lemma A.1 of Barber et al. 2019 and Lemma 1 of Barber 2020). We also mention again that negative results for testing argued via sampling with replacement can be traced back to Example 1 of Gretton et al. (2012). Following the

construction of Lemma 1 in Barber (2020), let $\mathcal{L} = \{(X_1, Y_1, Z_1), \ldots, (X_J, Y_J, Z_J)\} \overset{\text{i.i.d.}}{\sim} P_{X,Y,Z} \in \mathcal{P}_1$ be ghost samples for some $J \geq n(n-1)$. Given $\mathcal{L}$, draw $\{(X'_1, Y'_1, Z'_1), \ldots, (X'_n, Y'_n, Z'_n)\}$ from $\mathcal{L}$ *without replacement*. We denote this conditional sampling distribution by P_{without}. Then by marginalizing over $\mathcal{L}$, we simply have the identity

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi] = \mathbb{E}_{\mathcal{L}}[\mathbb{E}_{P_{\text{without}}}[\phi|\mathcal{L}]]. \tag{24}$$

Next consider drawing $\{(X'_1, Y'_1, Z'_1), \ldots, (X'_n, Y'_n, Z'_n)\}$ from $\mathcal{L}$ *with replacement* and denote this conditional sampling distribution by P_{with}. By Stam (1978), the total variation distance between P_{with} and P_{without} can be bounded by $n(n-1)/J$. Therefore, using the identity (24), we have

$$\mathbb{E}_{P^n_{X,Y,Z}}[\phi] \;\leq\; \mathbb{E}_{\mathcal{L}}[\mathbb{E}_{P_{\text{with}}}[\phi|\mathcal{L}]] + \frac{n(n-1)}{J}.$$

Let $\mathcal{A}$ be the event that $Z_1, \ldots, Z_J$ are distinct. Under the event $\mathcal{A}$, we have $X \perp\!\!\!\perp Y|Z$ for $(X, Y, Z) \sim P_{\text{with}}$ since the conditional distribution of X, Y given any Z takes a single value (i.e. conditionally degenerate). Therefore we should have

$$\mathbb{E}_{P_{\text{with}}}[\phi|\mathcal{L}, \mathcal{A}] \leq \alpha \quad \text{and} \quad \mathbb{E}_{P_{\text{with}}}[\phi|\mathcal{L}, \mathcal{A}^c] \leq 1$$

as $\sup_{P \in \mathcal{P}_{0,\text{disc}}} \mathbb{E}_P[\phi] \leq \alpha$ from our condition and $\phi \leq 1$. Combining the results together with the law of total expectation, we have for any $P_{X,Y,Z} \in \mathcal{P}_1$

$$\begin{aligned}
\mathbb{E}_{P^n_{X,Y,Z}}[\phi] \;&\leq\; \mathbb{E}_{\mathcal{L}}[\mathbb{E}_{P_{\text{with}}}[\phi|\mathcal{L}, \mathcal{A}] \times \mathbb{P}(\mathcal{A}|\mathcal{L})] + \mathbb{E}_{\mathcal{L}}[\mathbb{E}_{P_{\text{with}}}[\phi|\mathcal{L}, \mathcal{A}^c] \times \mathbb{P}(\mathcal{A}^c|\mathcal{L})] + \frac{n(n-1)}{J} \\
&\leq\; \alpha \times \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{A}^c) + \frac{n(n-1)}{J}.
\end{aligned}$$

Therefore the desired result follows.

Remark 10.

- From Equation (2.1) of Stam (1978), the result still holds if we replace $\frac{n(n-1)}{J}$ in the bound (4) with $2 - 2\frac{J!}{(J-n)!J^n}$ where one can prove that $2 - 2\frac{J!}{(J-n)!J^n} \leq \frac{n(n-1)}{J}$ by induction.
- Suppose that Z has a uniform distribution supported on a set of size $M_n := M > J_n := J$. Then $\rho_{J,P} = \mathbb{P}(\mathcal{A})$ is the probability that there is no collision, which can be computed as

  $$\rho_{J,P} = \frac{M}{M} \times \frac{M-1}{M} \times \cdots \times \frac{M-(J-1)}{M} = \frac{M!}{M^J(M-J)!}.$$

  This probability is lower and upper bounded by

  $$1 - \frac{J(J-1)}{2M} \leq \rho_{J,P} \leq \exp\left(-\frac{J(J-1)}{2M}\right).$$

  See Equation (2.4) of Stam (1978) for details. Therefore, in this case, the test ϕ has asymptotically no power when

  $$\frac{J_n^2}{M_n} \to 0 \quad \text{and} \quad \frac{n^2}{J_n} \to 0.$$

Now let a_n be a positive sequence that goes to zero arbitrarily slowly as $n \to \infty$. Then by taking $J_n^2 = M_n \times a_n$, the above conditions become equivalent to

$$\frac{n^4}{M_n \times a_n} \to 0.$$

This essentially means that any valid test has asymptotically no power in this setup if M_n increases much faster than n^4.

- We also note that the same proof carries over even if we assume that (X, Y, Z) is supported on a compact set.

A.2 Proof of Lemma 2

Continuing from Lemma 1, we have the list of inequalities:

$$\begin{aligned}
\mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\text{perm},n}] \;&\leq\; \alpha + \mathcal{D}_{\text{TV}}\big(Q^n_{X,Y,\widetilde{Z}}, \widetilde{Q}^n_{X,Y,\widetilde{Z}}\big) \overset{\text{(i)}}{\leq} \alpha + \sqrt{2}\mathcal{D}_{\text{H}}\big(Q^n_{X,Y,\widetilde{Z}}, \widetilde{Q}^n_{X,Y,\widetilde{Z}}\big) \\
&\overset{\text{(ii)}}{\leq} \alpha + \sqrt{2n}\mathcal{D}_{\text{H}}\big(Q_{X,Y,\widetilde{Z}}, \widetilde{Q}_{X,Y,\widetilde{Z}}\big) \\
&\overset{\text{(iii)}}{=} \alpha + \left\{2n \sum_{m=1}^{M} q_{\widetilde{Z}}(m) \times \mathcal{D}^2_H\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big)\right\}^{1/2},
\end{aligned}$$

where step (i) uses the well-known inequality $\mathcal{D}_{\text{TV}}(P, Q) \leq \sqrt{2}\mathcal{D}_{\text{H}}(P, Q)$ and step (ii) use subadditivity of the squared Hellinger distance for product measures and step (iii) can be verified by the definition of the Hellinger distance. This completes the proof of Lemma 2.

A.3 Proof of Theorem 2

We start with the case of $\gamma \in [1, 2]$. In this case, the first inequality (71) of Lemma 4 shows that

$$\begin{aligned}
&\mathcal{D}^2_{\text{H}}\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \leq \mathcal{D}^{\gamma}_{\gamma,\text{H}}\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \quad \text{and} \\
&\mathcal{D}^2_{\text{H}}\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big) \leq \mathcal{D}^{\gamma}_{\gamma,\text{H}}\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big).
\end{aligned}$$

Therefore, from Lemma 3, it holds that

$$\begin{aligned}
&\mathcal{D}^2_{\text{H}}\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \\
\leq\; &6\mathbb{E}_{Z_m,Z'_m,Z''_m}\big[\mathcal{D}^2_{\text{H}}\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \times \mathcal{D}^2_{\text{H}}\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big)\big] \\
\leq\; &6\mathbb{E}_{Z_m,Z'_m,Z''_m}\big[\mathcal{D}^{\gamma}_{\gamma,\text{H}}\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \times \mathcal{D}^{\gamma}_{\gamma,\text{H}}\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big)\big] \\
\overset{\text{(i)}}{\leq}\; &6L^{2\gamma} \sup_{z,z' \in B_m} \delta^{2\gamma}(z, z') \overset{\text{(ii)}}{\leq} 6L^{2\gamma}h^{2\gamma} \quad \text{for any } m \in [M],
\end{aligned}$$

where step (i) uses γ-Hellinger Lipschitzness and step (ii) follows by the definition of h in (7). Applying this result to the type I error bound in terms of the Hellinger distance in Lemma 2 yields the result.

Next we turn to the case of $\gamma > 2$. In this case, Corollary 1 proves

$$\mathcal{D}_{\mathrm{H}}^2\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \le C_\gamma \mathcal{D}_{\gamma,\mathrm{H}}^2\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \quad \text{and}$$
$$\mathcal{D}_{\mathrm{H}}^2\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big) \le C'_\gamma \mathcal{D}_{\gamma,\mathrm{H}}^2\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big),$$

where C_γ and C'_γ are constants that only depend on γ. Then following the same steps as before, it holds that

$$\mathcal{D}_{\mathrm{H}}^2\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \le \; C''_\gamma L^2 h^2 \quad \text{for any } m \in [M],$$

where C''_γ is another constant that only depends on γ. Combining this inequality with Lemma 2 yields the result. This completes the proof of Theorem 2.

A.4 Proof of Theorem 3

The proof of Theorem 3 follows exactly the same lines of the proof of Theorem 2 by replacing Corollary 1 with Lemma 5. For this reason, we omit the details.

A.5 Proof of Theorem 4

We proceed by considering the case of γ-Hellinger Lipschitzness (Appendix A.5.1 for $2 \le \gamma$ and Appendix A.5.2 for $1 \le \gamma < 2$) and γ-Rényi Lipschitzness (Appendix A.5.3) in order. We start with a high-level sketch of the proof.

- As discussed in Remark 5, we would like to show that the permutation test $\phi^{\dagger}_{\mathrm{perm},n}$ has power greater than $1-\beta$ for the hypothesis (13). This in turn implies our desired claim that $\phi^{\dagger}_{\mathrm{perm},n}$ has the type I error rate greater than $1-\beta$ for $H_0 : X \perp\!\!\!\perp Y|Z$. In this regard, we need to consider a powerful test that can reliably distinguish between $Q_{X,Y,\widetilde{Z}}$ and $\widetilde{Q}_{X,Y,\widetilde{Z}}$ to prove the result. To this end, we analyze the test statistic (11):

$$T_{\mathrm{CI}} = \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m U(\boldsymbol{W}_m).$$

- However, it is quite complicated to study T_{CI} directly as it is a sum of dependent variables. Our strategy to overcome this complication is to construct a distributional setting where $T_{\mathrm{CI}} \approx \sigma_1 U(\boldsymbol{W}_1)$ with high probability. More formally, we let the marginal density of Z be

$$p_Z(z) = \begin{cases} M-\epsilon, & \text{if } z \in [0, M^{-1}], \\ \frac{\epsilon}{M-1}, & \text{if } z \in (M^{-1}, 1], \\ 0, & \text{otherwise}, \end{cases} \tag{25}$$

for some $\epsilon := \epsilon_n > 0$, which may vary depending on n. Intuitively, when ϵ is sufficiently small (this will become precise in the proof), we only observe samples from the first bin with high probability and thus our test statistic will be dominated by the first term of the test statistic, i.e.

$$\sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m U(\boldsymbol{W}_m) \approx \sigma_1 U(\boldsymbol{W}_1).$$

If this is the case, then we are essentially testing for unconditional independence between X and Y within the first bin, and we can leverage the results of Kim et al. (2020) to investigate the power of the permutation test based on $\sigma_1 U(\boldsymbol{W}_1)$.

- In particular, Kim et al. (2020) present sufficient conditions using the first two moments of a test statistic under which the permutation test based on $U(\boldsymbol{W}_1)$ has significant power. At a high-level, these sufficient conditions guarantee that the test statistic is greater than the critical value of the permutation test with high probability, which yields high power.

- Throughout this proof, we only deal with the type II error (or 1−power) of $\phi_{\mathrm{perm},N}$ under Poissonization since that of $\phi^{\dagger}_{\mathrm{perm},n}$ is upper bounded by

$$\begin{aligned} \mathbb{E}[1-\phi^{\dagger}_{\mathrm{perm},n}] &\leq \mathbb{E}[1-\phi_{\mathrm{perm},N}] + \mathbb{P}(N > n) \\ &\leq \mathbb{E}[1-\phi_{\mathrm{perm},N}] + e^{-n/8}, \end{aligned} \tag{26}$$

where the second inequality uses an exponential Poisson tail bound (e.g. Canonne, 2017). Therefore, once we establish that the type II error of $\phi_{\mathrm{perm},N}$ is small, it follows that the type II error of $\phi^{\dagger}_{\mathrm{perm},n}$ is also small by taking n large enough.

The next section formalizes it under γ-Hellinger Lipschitzness for $2 \leq \gamma$.

A.5.1 γ-Hellinger Lipschitzness for $2 \leq \gamma$

Due to the inequality (26), it suffices to show that the type II error of $\phi_{\mathrm{perm},N}$ for the hypothesis (13) is sufficiently small under the given conditions. To proceed, let us denote the entire sample by $\mathcal{W}_N = \{(X_i, Y_i, Z_i)\}_{i=1}^N$ and denote the expectation and the variance operator under the permutation law by $\mathbb{E}_{\boldsymbol{\pi}}[\cdot|\boldsymbol{\sigma}, \mathcal{W}_N]$ and $\mathrm{Var}_{\boldsymbol{\pi}}[\cdot|\boldsymbol{\sigma}, \mathcal{W}_N]$ where we recall $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_M\}$. First of all, we make an observation that

$$\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}|\boldsymbol{\sigma}, \mathcal{W}_N] = \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \mathbb{E}_{\boldsymbol{\pi}}[U(\boldsymbol{W}^{\pi_m}_m)|\boldsymbol{\sigma}, \mathcal{W}_N] = 0, \tag{27}$$

where we use the fact that $\mathbb{E}_{\boldsymbol{\pi}}[U(\boldsymbol{W}^{\pi_m}_m)|\boldsymbol{\sigma}, \mathcal{W}_N] = 0$ when $\sigma_m \geq 4$ (see Appendix I of Kim et al., 2020). Thus a modification of Lemma 3.1 of Kim et al. (2020) shows that if

$$\mathbb{E}_{P^N_{X,Y,Z},N}[T_{\mathrm{CI}}] \geq \sqrt{\frac{2\mathrm{Var}_{P^N_{X,Y,Z},N}[T_{\mathrm{CI}}]}{\beta}} + \sqrt{\frac{2\mathbb{E}_{P^N_{X,Y,Z},N}\{\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}|\boldsymbol{\sigma}, \mathcal{W}_N]\}}{\alpha\beta}} \tag{28}$$

holds then the power of $\phi_{\mathrm{perm},N}$ is lower bounded by $1-\beta$. Therefore our goal is to prove the inequality (28) under the given conditions.

In what follows, we suppress the dependence on $P_{X,Y,Z}^N, N$ in the expectation and the variance for notational convenience, and often write $U(\boldsymbol{W}_m)$ as U_m if there is no confusion. The inequality (28) is verified by showing that there exists $P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)$ such that

$$\mathbb{E}[T_{\mathrm{CI}}] \gtrsim nh_n^4, \tag{29a}$$

$$\mathrm{Var}[T_{\mathrm{CI}}] \lesssim \mathbb{E}[T_{\mathrm{CI}}] + 1, \tag{29b}$$

$$\mathbb{E}\{\mathrm{Var}_{\boldsymbol{\pi}}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}, \mathcal{W}_N]\} \lesssim 1. \tag{29c}$$

When these inequalities are fulfilled for large n, the condition (28) is automatically satisfied since it is assumed that $nh_n^4 \to \infty$. Therefore our task boils down to confirming (29a), (29b) and (29c).

Proof of the inequality (29a). Let us define $p_i = \mathbb{P}(Z \in B_i)$ for $i = 1, \ldots, M$. Since the expectation of U_m is non-negative, the expected value of T_{CI} can be lower bounded by

$$\begin{aligned}\mathbb{E}[T_{\mathrm{CI}}] &\geq \mathbb{E}[\mathbb{1}(\sigma_1 \geq 4)\sigma_1 U_1] \\ &\overset{\text{(i)}}{\geq} (1 - 5e^{-1}/2)\min\{np_1, (np_1)^4\}\|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2,\end{aligned} \tag{30}$$

where (i) makes use of Lemma 3.1 of Canonne et al. (2018) and the fact that $\sigma_1 \sim \mathrm{Pois}(np_1)$. We also use the observation that $\mathbb{E}[U_1|\sigma_1]\mathbb{1}(\sigma_1 \geq 4) = \|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2\mathbb{1}(\sigma_1 \geq 4)$. By our choice of the density of Z in (25), we have $p_1 = 1 - M_n^{-1}\epsilon_n$ and we set $\epsilon_n \leq 1/2$. In this case, $np_1 \geq n/2$ and thus

$$\mathbb{E}[T_{\mathrm{CI}}] \gtrsim n\|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2. \tag{31}$$

Therefore it is sufficient to construct $P_{X|Z}$ and $P_{Y|Z}$ that satisfy (i) $P_{X|Z}P_{Y|Z}P_Z \in \mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)$ for $\gamma \geq 2$ and (ii) the squared L_2 distance between the corresponding $Q_{XY|\widetilde{Z}=1}$ and $Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}$ is lower bounded by

$$\|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2 \gtrsim h_n^4. \tag{32}$$

We construct such example below. For notational convenience, we often suppress the dependence of n in h_n and M_n, and also write $h = M^{-1}$ as we work on the equi-partition.

Example of the conditional distribution in $\mathcal{P}_{0,H,\gamma,\delta}(L)$. Suppose X and Y are Bernoulli random variables. Then the Lipschitz condition in Definition 3 can be equivalently written as

$$\begin{aligned}&\left|p_{X|Z}^{1/\gamma}(x=1|z) - p_{X|Z}^{1/\gamma}(x=1|z')\right|^\gamma + \left|p_{X|Z}^{1/\gamma}(x=0|z) - p_{X|Z}^{1/\gamma}(x=0|z')\right|^\gamma \\ \leq\ & 2L^\gamma|z - z'|^\gamma.\end{aligned}$$

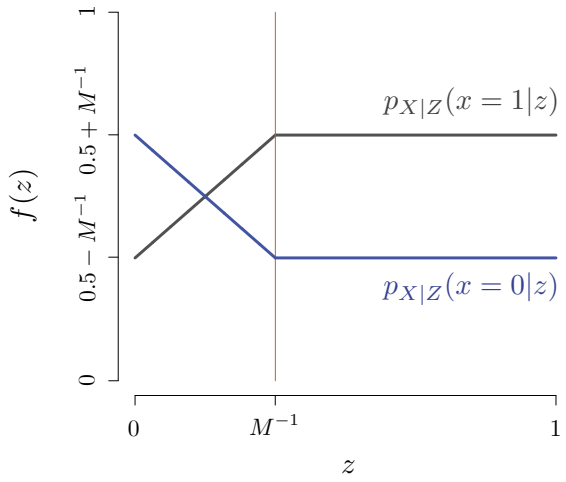


Figure 5: Illustration of $p_{X|Z}(x=1|z)$ and $p_{X|Z}(x=0|z)$ described in (33) as a function of z.

For any $M \geq 1$, the following probability mass functions are well-defined:

$$p_{X|Z}(x=1|z) = \begin{cases} \frac{1}{2} - \frac{1}{2M} + z, & \text{for } z \in [0, 1/M], \\ \frac{1}{2} + \frac{1}{2M}, & \text{for } z \in (1/M, 1], \\ 0, & \text{otherwise}, \end{cases}$$
$$p_{X|Z}(x=0|z) = \begin{cases} \frac{1}{2} + \frac{1}{2M} - z, & \text{for } z \in [0, 1/M], \\ \frac{1}{2} - \frac{1}{2M}, & \text{for } z \in (1/M, 1], \\ 0, & \text{otherwise}. \end{cases} \tag{33}$$

See Figure 5 for an illustration. In addition, by taking M large enough, we see that both $p_{X|Z}(x=1|z)$ and $p_{X|Z}(x=0|z)$ are bounded below and also above by, let's say, $[1/4, 3/4]$ for $z \in [0,1]$. Thus, using Taylor's theorem, one can verify that

$$\begin{aligned} & \left|p_{X|Z}^{1/\gamma}(x=1|z) - p_{X|Z}^{1/\gamma}(x=1|z')\right|^\gamma + \left|p_{X|Z}^{1/\gamma}(x=0|z) - p_{X|Z}^{1/\gamma}(x=0|z')\right|^\gamma \\ \leq\ & C_\gamma |z - z'|^\gamma, \end{aligned}$$

for some constant $C_\gamma > 0$. Thus it is seen that the distribution with the probability mass function (33) satisfies Definition 3 with $L^\gamma = C_\gamma/2$. Similarly we let $p_{Y|Z}(y=1|z) = p_{X|Z}(x=1|z)$ and thus $p_{Y|Z}(y|z)$ is also Lipschitz in γ-Hellinger distance.

Next we calculate $q_{XY|\widetilde{Z}}(x,y|m=1)$ and $q_{X\cdot|\widetilde{Z}}(x|m=1)q_{\cdot Y|\widetilde{Z}}(y|m=1)$. To start with $q_{XY|\widetilde{Z}}(x,y|m=1)$, recall that

$$d\widetilde{P}_{Z|Z\in B_1}(z) = \frac{dP_Z(z)}{\mathbb{P}(Z \in [0,h])}$$

and also recall that Z has the density in (25). Then since $h = M^{-1}$

$$\begin{aligned} q_{XY|\widetilde{Z}}(x=1,y=1|m=1) = & \int_0^h p_{X|Z}(x=1|z)p_{Y|Z}(y=1|z)d\widetilde{P}_{Z|Z\in B_1}(z) \\ = & \ h^{-1}\int_0^h p_{X|Z}(x=1|z)p_{Y|Z}(y=1|z)dz \\ = & \ h^{-1}\int_0^h \left(\frac{1}{2}-\frac{h}{2}+z\right)^2 dz \\ = & \ \frac{h^2}{12}+\frac{1}{4}. \end{aligned}$$

Similarly,

$$q_{X\cdot|\widetilde{Z}}(x=1|m=1) = q_{\cdot Y|\widetilde{Z}}(y=1|m=1) = \int_0^h p_{X|Z}(x=1|z)d\widetilde{P}_{Z|Z\in B_1}(z) = \frac{1}{2}.$$

For the binary case, the squared L_2 distance simply becomes

$$\begin{aligned} & \|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2 \\ = & \ 4|q_{XY|\widetilde{Z}}(x=1,y=1|m=1) - q_{X\cdot|\widetilde{Z}}(x=1|m=1)q_{\cdot Y|\widetilde{Z}}(y=1|m=1)|^2 \\ = & \ \frac{h^4}{144}. \end{aligned}$$

Therefore the condition (32) is fulfilled and the inequality (29a) follows.

Proof of the inequality (29b). By modifying the proof of Theorem 5.2 in Neykov et al. (2021), the variance of T_{CI} is upper bounded by

$$\begin{aligned} \mathrm{Var}[T_{\mathrm{CI}}] \ \lesssim \ & \mathbb{E}[T_{\mathrm{CI}}] + \mathbb{E}\sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4) \\ \lesssim \ & \mathbb{E}[T_{\mathrm{CI}}] + \sum_{m\in[M]} \mathbb{P}(\sigma_m \geq 1) \\ \lesssim \ & \mathbb{E}[T_{\mathrm{CI}}] + 1 - e^{-np_1} + \sum_{m=2}^{M}\left(1-e^{-np_m}\right), \end{aligned} \tag{34}$$

where we use the fact that $\sigma_m \sim \mathrm{Pois}(np_m)$, and C is some positive constant. Now by choosing $\epsilon_n > 0$ in the density function (25) such that

$$\epsilon_n \ \leq \ \frac{M(M-1)}{n}\log\left(\frac{M-1}{M-2}\right), \tag{35}$$

the summation in (34) can be controlled as $\sum_{m=2}^{M}(1-e^{-np_m}) \le 1$. In more detail, for any $m \ge 2$, the binned probability p_m is the same as $\frac{\epsilon_n}{M(M-1)}$. Thus we have a series of equivalent conditions as

$$\begin{aligned}
\sum_{m=2}^{M}\left(1-e^{-np_m}\right) \le 1 &\iff (M-1)(1-e^{-\frac{n\epsilon_n}{M(M-1)}}) \le 1 \\
&\iff \frac{M-2}{M-1} \le e^{-\frac{n\epsilon_n}{M(M-1)}} \\
&\iff \epsilon_n \le \frac{M(M-1)}{n}\log\left(\frac{M-1}{M-2}\right).
\end{aligned}$$

With the choice of such ϵ_n in the density of Z (25) (and also let $\epsilon_n \le 1/2$ from the previous step), the variance is upper bounded by

$$\mathrm{Var}[T_{\mathrm{CI}}] \lesssim \mathbb{E}[T_{\mathrm{CI}}] + 1.$$

Therefore the inequality (29b) follows.

Proof of the inequality (29c). Let us denote $U(\boldsymbol{W}_m^{\pi_m})$ by $U_m^{\pi_m}$ for simplicity. Then using the independence between $\pi_1, \ldots, \pi_M$, the conditional variance of $T_{\mathrm{CI}}^{\boldsymbol{\pi}}$ is

$$\mathrm{Var}_{\boldsymbol{\pi}}\left[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}, \mathcal{W}_N\right] = \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m^2 \mathrm{Var}_{\boldsymbol{\pi}}[U_m^{\pi_m}|\boldsymbol{\sigma}, \mathcal{W}_N].$$

Based on the result in Appendix J of Kim et al. (2020),

$$\begin{aligned}
\mathbb{E}\{\mathrm{Var}_{\boldsymbol{\pi}}[U_m^{\pi_m}|\boldsymbol{\sigma}, \mathcal{W}_N]|\boldsymbol{\sigma}\} &\lesssim \sigma_m^{-2}\max\left\{\|Q_{XY|\widetilde{Z}=m}\|_2^2, \|Q_{X\cdot|\widetilde{Z}=m}Q_{\cdot Y|\widetilde{Z}=m}\|_2^2\right\} \\
&\lesssim \sigma_m^{-2},
\end{aligned}$$

where the second inequality holds since $\|P\|_2^2 \le \|P\|_1^2 = 1$ for a discrete probability distribution P. Therefore, with the choice of ϵ_n as in (35),

$$\mathbb{E}\{\mathrm{Var}_{\boldsymbol{\pi}}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}, \mathcal{W}_N]\} \lesssim \sum_{m\in[M]} \mathbb{P}(\sigma_m \ge 4) \lesssim 1,$$

where the last inequality holds since $\sigma_m \sim \mathrm{Pois}(np_m)$ and

$$\sum_{m\in[M]} \mathbb{P}(\sigma_m \ge 4) \le \sum_{m\in[M]} \mathbb{P}(\sigma_m \ge 1) \le 1 + \sum_{m=2}^{M}\left(1-e^{-np_m}\right) \le 2.$$

This verifies the inequality (29c).

Indeed, our proof in this subsection goes through for the case where $1 \le \gamma < 2$ as well. However, the condition $\sqrt{n}h_n^2 \to \infty$ is not tight for this case, and we need a more careful analysis detailed in the next subsection.

A.5.2 γ-Hellinger Lipschitzness for $1 \le \gamma < 2$

The underlying idea of the proof of this result is the same as in the previous section. That is, we want to show that the inequality (28) holds when $\sqrt{n}h_n^\gamma \to \infty$ for some $P_{X,Y,Z} \in \mathcal{P}_{0,\mathrm{H},\gamma,\delta}(L)$ with $\gamma \in [1,2]$. However, we need to obtain a sharper bound for the variance of the test statistic in order to achieve the desired rate, which requires some nontrivial effort.

Suppose that X and Y are binary random variables taking a value between $\{0,1\}$. We then set

$$
\begin{aligned}
p_{X|Z}(x=1|z) &= \begin{cases} z^\gamma, & \text{for } z \in [0, 2^{-1/\gamma}], \\ 2^{-1}, & \text{for } z \in [2^{-1/\gamma}, 1], \\ 0, & \text{otherwise.} \end{cases} \\
p_{X|Z}(x=0|z) &= \begin{cases} 1-z^\gamma, & \text{for } z \in [0, 2^{-1/\gamma}], \\ 2^{-1}, & \text{for } z \in [2^{-1/\gamma}, 1], \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{36}
$$

Similarly, we let $p_{X|Z}(x=1|z) = p_{Y|Z}(y=1|z)$ and $p_{X|Z}(x=0|z) = p_{Y|Z}(y=0|z)$. It can be checked that $P_{X|Z}$ and $P_{Y|Z}$ with the above probability mass functions are γ-Hellinger smooth for $\gamma \ge 1$ using Taylor's theorem. Here we note that the Lipschitz constant depends on γ and is assumed to be fixed. Also note that this example is more sophisticated than the one in (33). For instance, when $\gamma = 1$, the distribution (36) is TV smooth but not smooth in the Hellinger distance (since $\sqrt{x}$ is not a Lipschitz function). In addition to the condition on the conditional distributions (36), we let Z have the marginal density as in (25) with a sufficiently small $0 < \epsilon_n \le 1/2$ (specified later).

Under this setting, we will show that the mean, the variance and the expected conditional variance of the test statistic T_{CI} are bounded by

$$
\mathbb{E}[T_{\mathrm{CI}}] \gtrsim n h_n^{4\gamma}, \tag{37a}
$$

$$
\mathrm{Var}[T_{\mathrm{CI}}] \lesssim \mathbb{E}[T_{\mathrm{CI}}] h_n^{2\gamma} + h_n^{4\gamma}, \tag{37b}
$$

$$
\mathbb{E}\big\{\mathrm{Var}_{\boldsymbol{\pi}}\big[T_{\mathrm{CI}}^{\boldsymbol{\pi}} | \boldsymbol{\sigma}, \mathcal{W}_N\big]\big\} \lesssim h_n^{4\gamma}. \tag{37c}
$$

If this is the case, then the inequality (28) holds as $\sqrt{n}h_n^\gamma \to \infty$, which completes the proof. In what follows, we prove these inequalities in order.

Proof of the inequality (37a). As we define the density of Z to be (25) with $\epsilon_n \le 1/2$, the inequality (31) follows in this case as well, that is

$$
\mathbb{E}[T_{\mathrm{CI}}] \gtrsim n \| Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1} Q_{\cdot Y|\widetilde{Z}=1} \|_2^2. \tag{38}
$$

Therefore it is enough to show that $\| Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1} Q_{\cdot Y|\widetilde{Z}=1} \|_2^2 \gtrsim h^{4\gamma}$. Then the inequality (37a) follows. In fact, when $h \le 1/2$,

$$
q_{XY|\widetilde{Z}}(x=1, y=1|m=1) = \int_0^h p_{X|Z}(x=1|z) p_{Y|Z}(y=1|z) d\widetilde{P}_{Z|Z\in B_1}(z)
$$

$$= h^{-1} \int_0^h p_{X|Z}(x=1|z) p_{Y|Z}(y=1|z) dz$$
$$= h^{-1} \int_0^h z^{2\gamma} dz = \frac{h^{2\gamma}}{2\gamma+1}.$$

Similarly,

$$q_{X\cdot|\widetilde{Z}}(x=1|m=1) = q_{\cdot Y|\widetilde{Z}}(y=1|m=1) = \int_0^h p_{X|Z}(x=1|z) d\widetilde{P}_{Z|Z\in B_1}(z) = \frac{h^\gamma}{\gamma+1}.$$

When X and Y are binary, the squared L_2 norm simply becomes

$$\begin{aligned} & \|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1} Q_{\cdot Y|\widetilde{Z}=1}\|_2^2 \\ = \; & 4\big\{q_{XY|\widetilde{Z}}(x=1,y=1|m=1) - q_{X\cdot|\widetilde{Z}}(x=1|m=1) q_{\cdot Y|\widetilde{Z}}(y=1|m=1)\big\}^2 \qquad (39) \\ = \; & 4h^{4\gamma} \left(\frac{1}{2\gamma+1} - \frac{1}{(\gamma+1)^2}\right)^2, \end{aligned}$$

which concludes the inequality (37a).

Proof of the inequality (37b). Bounding the variance requires a bit more work. First the law of total variance yields

$$\mathrm{Var}[T_{\mathrm{CI}}] = \mathbb{E}\{\mathrm{Var}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\} + \mathrm{Var}\{\mathbb{E}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\}. \qquad (40)$$

To start with the first term, we use the independence between different bins conditional on $\boldsymbol{\sigma}$ and see

$$\begin{aligned} \mathbb{E}\{\mathrm{Var}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\} \; &= \; \sum_{m\in[M]} \mathbb{E}\{\mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \mathrm{Var}[U_m|\boldsymbol{\sigma}]\} \\ &\leq \; \mathbb{E}\{\mathbb{1}(\sigma_1 \geq 4)\sigma_1^2 \mathrm{Var}[U_1|\sigma_1]\} + \sum_{m=2}^M \mathbb{E}[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2], \end{aligned}$$

where the second inequality uses $|U_m| \leq 1$. Since we have freedom to choose ϵ in the density of Z, we can make the second term smaller than

$$\sum_{m=2}^M \mathbb{E}[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2] \; \lesssim \; h^{4\gamma}. \qquad (41)$$

For example, since $\sigma_m \sim \mathrm{Pois}(np_m)$ where $p_m = \frac{\epsilon}{M(M-1)}$ for $m = 2, \ldots, M$, applying Cauchy–Schwarz inequality yields

$$\sum_{m=2}^M \mathbb{E}[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2] \; \lesssim \; M\sqrt{\mathbb{P}(\sigma_2 \geq 4)\mathbb{E}[\sigma_2^4]}$$

$$\lesssim\ M\sqrt{1-e^{-np_2}},$$

where the second inequality follows since $\mathbb{P}(\sigma_2 \geq 4) \leq \mathbb{P}(\sigma_2 \geq 1) = 1 - e^{-np_2}$ and $\mathbb{E}[\sigma_2^4] = np_2(1 + 7np_2 + 6n^2p_2^2 + n^3p_2^3) \leq 15$ by assuming $np_2 \leq 1$, which is satisfied when $\epsilon \lesssim 1/(nh^2)$. Therefore, the inequality (41) holds, for instance, when

$$\epsilon\ \lesssim\ \frac{1}{nh^2}\log\left(\frac{1}{1-h^{4\gamma+1}}\right).$$

Next we analyze $\mathbb{E}\{\mathbb{1}(\sigma_1 \geq 4)\sigma_1^2\mathrm{Var}[U_1|\sigma_1]\}$. Intuitively the conditional variance of $U_1|\sigma_1$ becomes smaller for large M (equivalently small h) under our distributional setting (36). In particular, observe that $q_{XY|\widetilde{Z}}(x=0,y=0|m=1) \to 1$ and thus $(X,Y)|Z \in B_1$ becomes degenerate at $X=0, Y=0$ as M increases. However the bound for $\mathrm{Var}[U_1|\sigma_1]$ given in Lemma 5.1 of Neykov et al. (2021) (and also see the proof of Proposition 5.3 in Kim et al., 2020):

$$\begin{aligned}\mathrm{Var}[U_1|\sigma_1]\ &\lesssim\ \frac{\mathbb{E}[U_1|\sigma_1]\max\left\{\|Q_{XY|\widetilde{Z}=1}\|_2, \|Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2\right\}}{\sigma_1}\\ &\quad+\frac{\max\left\{\|Q_{XY|\widetilde{Z}=1}\|_2^2, \|Q_{X\cdot|\widetilde{Z}=1}Q_{\cdot Y|\widetilde{Z}=1}\|_2^2\right\}}{\sigma_1^2}\quad\text{for } \sigma_1 \geq 4,\end{aligned}$$

does not capture this intuition. It turns out that this variance bound is not tight for the binary case. In fact, we can obtain a sharper bound as

$$\begin{aligned}\mathrm{Var}[U_1|\sigma_1]\ &\lesssim\ \frac{\mathbb{E}[U_1|\sigma_1]\max\left\{q_{XY|\widetilde{Z}}(1,1|1), q_{X\cdot|\widetilde{Z}}(1|1)q_{\cdot Y|\widetilde{Z}}(1|1)\right\}}{\sigma_1}\\ &\quad+\frac{\max\left\{q_{XY|\widetilde{Z}}^2(1,1|1), q_{X\cdot|\widetilde{Z}}^2(1|1)q_{\cdot Y|\widetilde{Z}}^2(1|1)\right\}}{\sigma_1^2}\quad\text{for } \sigma_1 \geq 4,\end{aligned}\tag{42}$$

where $q_{XY|\widetilde{Z}}(1,1|1) = q_{XY|\widetilde{Z}}(x=1,y=1|m=1)$ and the other quantities are similarly defined. We will prove this inequality at the end of this section. Therefore, using the previous calculations of $q_{XY|\widetilde{Z}}$, $q_{X\cdot|\widetilde{Z}}$ and $q_{\cdot Y|\widetilde{Z}}$, the (conditional) variance is bounded by

$$\mathrm{Var}[U_1|\sigma_1]\ \lesssim\ \frac{\mathbb{E}[U_1|\sigma_1]h^{2\gamma}}{\sigma_1}+\frac{h^{4\gamma}}{\sigma_1^2}\quad\text{for } \sigma_1 \geq 4.\tag{43}$$

In summary, we have

$$\begin{aligned}\mathbb{E}\{\mathrm{Var}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\}\ &\lesssim\ \mathbb{E}\{\mathbb{1}(\sigma_1 \geq 4)\sigma_1\mathbb{E}[U_1|\sigma_1]\}h^{2\gamma}+h^{4\gamma}\\ &\lesssim\ \mathbb{E}[T_{\mathrm{CI}}]h^{2\gamma}+h^{4\gamma}.\end{aligned}\tag{44}$$

Next we focus on the second term $\mathrm{Var}\{\mathbb{E}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\}$ in (40) and see

$$\mathrm{Var}\{\mathbb{E}[T_{\mathrm{CI}}|\boldsymbol{\sigma}]\}\ =\ \sum_{m\in[M]}\mathrm{Var}[\sigma_m\mathbb{1}(\sigma_m \geq 4)]\times\|Q_{XY|\widetilde{Z}=m}-Q_{X\cdot|\widetilde{Z}=m}Q_{\cdot Y|\widetilde{Z}=m}\|_2^4$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\lesssim} \sum_{m\in[M]} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \times \|Q_{XY|\widetilde{Z}=m} - Q_{X\cdot|\widetilde{Z}=m} Q_{\cdot Y|\widetilde{Z}=m}\|_2^4 \\
&\overset{\text{(ii)}}{\lesssim} np_1 \times \left|q_{X,Y|\widetilde{Z}}(1,1|1) - q_{X\cdot|\widetilde{Z}}(1|1) q_{\cdot Y|\widetilde{Z}}(1|1)\right|^4 + h^{4\gamma} \\
&\overset{\text{(iii)}}{\lesssim} \mathbb{E}[T_{\mathrm{CI}}] h^{2\gamma} + h^{4\gamma},
\end{aligned}
$$

where step (i) uses Claim 2.1 of Canonne et al. (2018), step (ii) can be deduced by the inequality (41) and the identity (39). To give a more detail on step (ii), consider a decomposition

$$
\begin{aligned}
&\sum_{m\in[M]} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \times \|Q_{XY|\widetilde{Z}=m} - Q_{X\cdot|\widetilde{Z}=m} Q_{\cdot Y|\widetilde{Z}=m}\|_2^4 \\
= \; &\underbrace{\mathbb{E}[\sigma_1 \mathbb{1}(\sigma_1 \geq 4)] \times \|Q_{XY|\widetilde{Z}=1} - Q_{X\cdot|\widetilde{Z}=1} Q_{\cdot Y|\widetilde{Z}=1}\|_2^4}_{\text{(I)}} \\
&+ \underbrace{\sum_{m=2}^{M} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \times \|Q_{XY|\widetilde{Z}=m} - Q_{X\cdot|\widetilde{Z}=m} Q_{\cdot Y|\widetilde{Z}=m}\|_2^4}_{\text{(II)}}.
\end{aligned}
$$

For the first term (I), we make use of the identity (39) and $\mathbb{E}[\sigma_1] = np_1$, which yields (I) $\lesssim np_1 \times \left|q_{X,Y|\widetilde{Z}}(1,1|1) - q_{X\cdot|\widetilde{Z}}(1|1) q_{\cdot Y|\widetilde{Z}}(1|1)\right|^4$. For the second term, note that the L_2 norm of $Q_{XY|\widetilde{Z}=m} - Q_{X\cdot|\widetilde{Z}=m} Q_{\cdot Y|\widetilde{Z}=m}$ is bounded by two for any m. Therefore, (II) $\lesssim \sum_{m=2}^{M} \mathbb{E}[\sigma_m^2 \mathbb{1}(\sigma_m \geq 4)] \lesssim h^{4\gamma}$ by the inequality (41). This explains step (ii). Step (iii) follows by using the lower bound for $\mathbb{E}[T_{\mathrm{CI}}]$ in (38) and the observation that $|q_{XY|\widetilde{Z}}(1,1|1) - q_{X\cdot|\widetilde{Z}}(1|1) q_{\cdot Y|\widetilde{Z}}(1|1)|^2 \lesssim h^{4\gamma} \leq h^{2\gamma}$. Combining this with (44) yields the claim (37b). The last step is to verify the sharper bound (42).

Verification of the bound (42). Recall that the kernel of U_1 is

$$
h_{i_1,i_2,i_3,i_4} = \frac{1}{4!} \sum_{\pi\in\mathbf{\Pi}_4} \sum_{x\in[\ell_1], y\in[\ell_2]} \psi_{\pi(1)\pi(2)}(x,y) \psi_{\pi(3)\pi(4)}(x,y),
$$

where $\mathbf{\Pi}_4$ is the set of all permutations of $\{i_1, i_2, i_3, i_4\}$ and

$$
\psi_{ij}(x,y) = \mathbb{1}(X_{i,1} = x, Y_{i,1} = y) - \mathbb{1}(X_{i,1} = x)\mathbb{1}(Y_{j,1} = y).
$$

Importantly, for the binary case, we observe that

$$
h_{i_1,i_2,i_3,i_4} = g_{i_1,i_2,i_3,i_4} := \frac{4}{4!} \sum_{\pi\in\mathbf{\Pi}_4} \psi_{\pi(1)\pi(2)}(1,1) \psi_{\pi(3)\pi(4)}(1,1). \tag{45}
$$

This means that U_1 is the same as another U-statistic defined with the kernel g_{i_1,i_2,i_3,i_4}. Having this observation, we work with the U-statistic associated with g_{i_1,i_2,i_3,i_4}. We then follow basically

the same lines of the proofs in Neykov et al. (2021) and Appendix J of Kim et al. (2020) and see that the bound (42) is satisfied (more specifically, follow the same lines of the proof in Appendix J of Kim et al. (2020) by setting $d_1 = d_2 = 1$).

Proof of the inequality (37c). Similar to the proof of (29c), we can obtain

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\pi}}\big[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}, \mathcal{W}_N\big] &= \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \mathrm{Var}_{\boldsymbol{\pi}}[U_m^{\pi_m}|\boldsymbol{\sigma}, \mathcal{W}_N] \\ &\leq \mathbb{1}(\sigma_1 \geq 4)\sigma_1^2 \mathrm{Var}_{\boldsymbol{\pi}}[U_1^{\pi_1}|\boldsymbol{\sigma}, \mathcal{W}_N] + \sum_{m=2}^{M} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2,\end{aligned}$$

where the second inequality follows by $|U_m^{\pi_m}| \leq 1$. By taking the expectation, the second term can be smaller than $h^{4\gamma}$ by the condition (41). On the other hand, recall that for the binary case, U_1 is equivalent to the U-statistic based on the kernel g_{i_1,i_2,i_3,i_4} in (45). Thus, based on the results in Appendix I and Appendix J of Kim et al. (2020) (more specifically by letting $d_1 = d_2 = 1$ therein),

$$\begin{aligned}\mathbb{E}\{\mathrm{Var}_{\boldsymbol{\pi}}[U_1^{\pi_1}|\boldsymbol{\sigma}, \mathcal{W}_N]|\boldsymbol{\sigma}\} &\lesssim \sigma_1^{-2} \max\big\{q^2_{XY|\widetilde{Z}}(1,1|1), q^2_{X\cdot|\widetilde{Z}}(1|1) q^2_{\cdot Y|\widetilde{Z}}(1|1)\big\} \\ &\lesssim \sigma_1^{-2} h^{4\gamma}.\end{aligned}$$

Consequently, we obtain the desired bound $\mathbb{E}\big\{\mathrm{Var}_{\boldsymbol{\pi}}\big[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}, \mathcal{W}_N\big]\big\} \lesssim h_n^{4\gamma}$.

A.5.3 γ-Rényi Lipschitzness for $0 < \gamma$

The proof of this part is essentially the same as that in the case of γ-Hellinger Lipschitzness for $2 \leq \gamma$ (Appendix A.5.1). In particular, we consider the distribution setting in (33) and show that the considered distribution is Rényi smooth for all $\gamma > 0$. Once this is established, the other parts of the proof follow exactly the same lines of the proof as before in Appendix A.5.1.

Note that, for a sufficiently large M, $p_{X|Z}(x = 1|z)$ and $p_{X|Z}(x = 0|z)$ are bounded below and above by constants for any $z \in [0, 1]$, say,

$$p_{X|Z}(x = 1|z) \in [1/4, 3/4] \quad \text{and} \quad p_{X|Z}(x = 0|z) \in [1/4, 3/4]. \tag{46}$$

Therefore, when $0 < \gamma \leq 2$, there exist some constants $C_1, C_2 > 0$ such that

$$\begin{aligned}\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) &\overset{\text{(i)}}{\leq} \mathcal{D}_{\chi^2}(P_{X|Z=z}\|P_{X|Z=z'}) \\ &\overset{\text{(ii)}}{\leq} C_1\|P_{X|Z=z} - P_{X|Z=z'}\|_2^2 \\ &\overset{\text{(iii)}}{\leq} C_2|z - z'|^2,\end{aligned}$$

where step (i) uses the fact that Rényi divergence is nondecreasing in γ (e.g. Theorem 3 of Van Erven and Harremos, 2014) and the inequality (74). Additionally, step (ii) holds since we set $P_{X|Z=z'}$ is lower bounded by some constant and step (iii) follows directly by the construction of $P_{X|Z=z}$ in

(33). Similarly we let $P_{Y|Z=z} = P_{X|Z=z}$ and thus both $P_{X|Z=z}$ and $P_{Y|Z=z}$ are Rényi smooth when $0 < \gamma \leq 2$.

We now consider the case where $\gamma > 2$. Again, we make use of the distribution given in (33) and assume that $p_{X|Z}(x = 1|z)$ and $p_{X|Z}(x = 0|z)$ are bounded below and above by constants for any $z \in [0, 1]$ as in (46). Since X takes a binary value, Rényi divergence of $P_{X|Z=z}$ from $P_{X|Z=z'}$ simply becomes

$$\begin{aligned}
&\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) \\
&= \frac{1}{\gamma-1}\log\left\{\left(\frac{p_{X|Z}(x=1|z)}{p_{X|Z}(x=1|z')}\right)^{\gamma} p_{X|Z}(x=1|z') + \left(\frac{p_{X|Z}(x=0|z)}{p_{X|Z}(x=0|z')}\right)^{\gamma} p_{X|Z}(x=0|z')\right\}.
\end{aligned}$$

When both $z, z' \in (1/M, 1]$, Rényi divergence becomes zero and hence there is nothing to prove. Next assume that both $z, z' \in [0, 1/M]$. In this case, Rényi divergence can be written more explicitly as

$$\begin{aligned}
&\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) \\
&= \frac{1}{\gamma-1}\log\left\{\left(\frac{\frac{1}{2}-\frac{1}{2M}+z}{\frac{1}{2}-\frac{1}{2M}+z'}\right)^{\gamma}\left(\frac{1}{2}-\frac{1}{2M}+z'\right) + \left(\frac{\frac{1}{2}+\frac{1}{2M}-z}{\frac{1}{2}+\frac{1}{2M}-z'}\right)^{\gamma}\left(\frac{1}{2}+\frac{1}{2M}-z'\right)\right\} \\
&:= f(z).
\end{aligned}$$

By viewing the above expression as a function of z, Taylor's theorem yields that there exists $\xi \in (0, 1/M)$ such that

$$f(z) = f(z') + f'(z')(z - z') + f''(\xi)(z - z')^2.$$

The first term $f(z')$ is zero. We also note that the first derivative satisfies $f'(z') = 0$ since

$$f'(z) = \frac{1}{\gamma-1} \times \frac{\gamma\left(\frac{\frac{1}{2}-\frac{1}{2M}+z}{\frac{1}{2}-\frac{1}{2M}+z'}\right)^{\gamma-1} - \gamma\left(\frac{\frac{1}{2}+\frac{1}{2M}-z}{\frac{1}{2}+\frac{1}{2M}-z'}\right)^{\gamma-1}}{\left(\frac{\frac{1}{2}-\frac{1}{2M}+z}{\frac{1}{2}-\frac{1}{2M}+z'}\right)^{\gamma}\left(\frac{1}{2}-\frac{1}{2M}+z'\right) + \left(\frac{\frac{1}{2}+\frac{1}{2M}-z}{\frac{1}{2}+\frac{1}{2M}-z'}\right)^{\gamma}\left(\frac{1}{2}+\frac{1}{2M}-z'\right)}.$$

Furthermore, it can be seen that $\sup_{z\in[0,1/M]} |f''(z)| \leq C_{\gamma,M}$ where $C_{\gamma,M}$ is a positive constant depending only on γ and M. This shows that $\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) \leq C_{\gamma,M}(z - z')^2$ for all $z, z' \in [0, 1/M]$. Suppose now that $z \in [0, 1/M]$ and $z' \in (1/M, 1]$. In this case, since $p_{X|Z}(x = 1|z') = p_{X|Z}(x = z|1/M)$ and $p_{X|Z}(x = 0|z') = p_{X|Z}(x = 0|1/M)$ for all $z' \in (1/M, 1]$, the previous result implies that $\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) \leq C_{\gamma,M}(z - 1/M)^2 \leq C_{\gamma,M}(z - z')^2$. The final case where $z \in (1/M, 1]$ and $z' \in [0, 1/M]$ can be similarly handled, which concludes that $\mathcal{D}_{\gamma,R}(P_{X|Z=z}\|P_{X|Z=z'}) \leq C_{\gamma,M}(z - z')^2$ for all $z, z' \in [0, 1]$. Hence, by letting $P_{Y|Z=z} = P_{X|Z=z}$, both $P_{X|Z=z}$ and $P_{Y|Z=z}$ are Rényi smooth for $\gamma \geq 2$.

In summary, the distribution in (33) is Rényi smooth for $\gamma > 0$ with a different Lipschitz constant and therefore the result $\sup_{P_{X,Y,Z}\in\mathcal{P}_{0,R,\gamma,\delta}(L)} \mathbb{E}_{P^N_{X,Y,Z},N}[\phi^{\dagger}_{\mathrm{perm},n}] \geq 1 - \beta$ follows.

A.6 Proof of Theorem 5

We start by proving the type II error bound for $\phi_{\text{perm},2}$ and then turn to $\phi_{\text{perm},1}$.

• **Type II error of $\phi_{\text{perm},2}$.** We note that by Lemma 6, the type II error of $\phi_{\text{perm},2}$ can be expressed as

$$\mathbb{E}_{P^N_{X,Y,Z},N}[1-\phi_{\text{perm},2}] = \mathbb{P}_{P^N_{X,Y,Z},N}[T_{\text{CI},W} \le q_{1-\alpha}],$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of $T^{\boldsymbol{\pi}_1}_{\text{CI},W}, \ldots, T^{\boldsymbol{\pi}_K}_{\text{CI},W}$. Therefore, it is enough to work with the test that rejects the null when $T_{\text{CI},W} > q_{1-\alpha}$. For simplicity, we use the notation $\boldsymbol{\mathcal{W}}_{\text{weight}} := \{\boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}\}_{m\in[M]}$ to denote samples used for weights. Given this notation, we first claim that

$$\begin{aligned}
&\mathbb{P}_{P^N_{X,Y,Z},N}[T_{\text{CI},W} \le q_{1-\alpha}] \\
\le\ &\mathbb{P}_{P^N_{X,Y,Z},N}\left\{T_{\text{CI},W} < \zeta\sqrt{\min\{n,M\} + \mathbb{E}_{P^N_{X,Y,Z},N}[T_{\text{CI},W}|\boldsymbol{\mathcal{W}}_{\text{weight}}, \boldsymbol{\sigma}]}\right\} + \frac{1}{200}, \qquad (47)
\end{aligned}$$

where ζ is a sufficiently large constant as in the critical value of the test in Canonne et al. (2018). We then show that under the given condition,

$$\mathbb{P}_{P^N_{X,Y,Z},N}\left\{T_{\text{CI},W} < \zeta\sqrt{\min\{n,M\} + \mathbb{E}_{P^N_{X,Y,Z},N}[T_{\text{CI},W}|\boldsymbol{\mathcal{W}}_{\text{weight}}, \boldsymbol{\sigma}]}\right\} \le \frac{1}{200}, \qquad (48)$$

which establishes the desired result. To ease the notation, we suppress dependence on $P^N_{X,Y,Z}$ and N in the expectation and probability function, throughout this proof.

1. Proof of the first claim (47). To show the first claim (47), let us define an event

$$\mathcal{A} := \{q_{1-\alpha} \le \zeta\sqrt{\min(n,M) + \mathbb{E}_{P^N_{X,Y,Z},N}[T_{\text{CI},W}|\boldsymbol{\mathcal{W}}_{\text{weight}}, \boldsymbol{\sigma}]}\},$$

and show that

$$\mathbb{P}(\mathcal{A}) \ge \frac{199}{200}. \qquad (49)$$

If this is the case, then the first claim (47) follows by the union bound. To begin, let us simply denote the permutation distribution of $T^{\boldsymbol{\pi}}_{\text{CI},W}$ by

$$\mathbb{P}_{\boldsymbol{\pi}}\left[T^{\boldsymbol{\pi}}_{\text{CI},W} \le t\right] = \frac{1}{K}\sum_{\boldsymbol{\pi}_i\in\boldsymbol{\Pi}} \mathbb{1}\left[T^{\boldsymbol{\pi}_i}_{\text{CI},W} \le t\right].$$

More generally, we let $\mathbb{P}_{\boldsymbol{\pi}}[\cdot]$ be the probability measure in terms of $\boldsymbol{\pi}$ uniformly distributed over $\boldsymbol{\Pi}$ conditional on everything else. We denote the expectation and the variance operator under the permutation law by $\mathbb{E}_{\boldsymbol{\pi}}[\cdot]$ and $\text{Var}_{\boldsymbol{\pi}}[\cdot]$ for simplicity. As in the case of (27) for $T^{\boldsymbol{\pi}}_{\text{CI}}$, we make an observation that $T^{\boldsymbol{\pi}}_{\text{CI},W}$ is centered under the permutation law, i.e.

$$\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\text{CI},W}] = \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m\omega_m\mathbb{E}_{\boldsymbol{\pi}}[U_W(\boldsymbol{W}^{\pi_m}_m)] = 0,$$

which can be shown using the result in Appendix I of Kim et al. (2020). Therefore, for any $t > 0$, Chebyshev's inequality yields

$$\mathbb{P}_{\boldsymbol{\pi}}\{T^{\boldsymbol{\pi}}_{\mathrm{CI},W} \geq t\} \;\leq\; \frac{1}{t^2}\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big].$$

Then, by the definition of the $1-\alpha$ quantile, it holds that $q_{1-\alpha} \leq \sqrt{\alpha^{-1}\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big]}$ with probability one. Notice that $\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big]$ is a non-negative random variable, changing its value depending on $\{\boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}, \boldsymbol{W}_{XY,m}\}_{m=1}^{M}$. Then Markov's inequality verifies that the following inequality

$$\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big] < 400\mathbb{E}\big\{\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big]|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big\}$$

holds with probability at least $1 - 1/400$. Based on the results of Canonne et al. (2018), Neykov et al. (2021) and Kim et al. (2020), we further prove below that

$$400\alpha^{-1}\mathbb{E}\big\{\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big]|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big\} \leq \zeta^{1/2}\big(\min\{n, M\} + \mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]\big), \tag{50}$$

with probability at least $1 - 1/400$ by taking ζ sufficiently large (depending on α, but recall that α is assumed to be a fixed constant). This verifies the probability bound (49) for $\mathcal{A}$, completing the proof of the first claim (47).

Detail of the inequality (50). First of all, since $\pi_1, \ldots, \pi_M$ are independent, it can be verified that

$$\mathbb{E}_{\boldsymbol{\pi}}\big[(T^{\boldsymbol{\pi}}_{\mathrm{CI},W})^2\big] = \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2\omega_m^2\mathbb{E}_{\boldsymbol{\pi}}\big[U_W^2(\boldsymbol{W}_m^{\pi_m})\big]. \tag{51}$$

In order to analyze the expectation of $\mathbb{E}_{\boldsymbol{\pi}}\big[U_W^2(\boldsymbol{W}_m^{\pi_m})\big]$, we need some preliminary results. To this end, for $m = 1, \ldots, M$, let

$$g_{X,m}(x_1, x_2) := \sum_{x\in[\ell_1]} \frac{\mathbb{1}(x_1 = x)\mathbb{1}(x_2 = x)}{1 + a_{x,m}} \quad \text{and}$$

$$g_{Y,m}(y_1, y_2) := \sum_{y\in[\ell_2]} \frac{\mathbb{1}(y_1 = y)\mathbb{1}(y_2 = y)}{1 + a_{y,m}},$$

and for $p : [\ell_1] \times [\ell_2] \mapsto \mathbb{R}$, denote

$$\|p\|^2_{2,\boldsymbol{a}_m} := \sum_{x\in[\ell_1]}\sum_{y\in[\ell_2]} \frac{p^2(x, y)}{(1 + a_{x,m})(1 + a_{y,m})}.$$

Since $a_{x,m}, a_{y,m} \geq 0$, the above bivariate functions are non-negative and bounded by one. Therefore

$$\begin{aligned}
&\mathbb{E}[g^2_{X,m}(X_{1,m}, X_{2,m})g^2_{Y,m}(Y_{1,m}, Y_{2,m})|\boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}] \\
\leq\; &\mathbb{E}[g_{X,m}(X_{1,m}, X_{2,m})g_{Y,m}(Y_{1,m}, Y_{2,m})|\boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}] \;=\; \|P_{XY|Z=m}\|^2_{2,\boldsymbol{a}_m}.
\end{aligned}$$

Similar calculations show

$$\mathbb{E}[g_{X,m}^2(X_{1,m}, X_{2,m}) g_{Y,m}^2(Y_{3,m}, Y_{4,m}) | \boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}] \;\le\; \|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2$$

and

$$\begin{aligned} &\mathbb{E}[g_{X,m}^2(X_{1,m}, X_{2,m}) g_{Y,m}^2(Y_{1,m}, Y_{3,m}) | \boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}] \\ \le\; &\frac{1}{2}\|P_{XY|Z=m}\|_{2,\boldsymbol{a}_m}^2 + \frac{1}{2}\|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2, \end{aligned}$$

where the last inequality uses the inequality $xy \le \frac{1}{2}x^2 + \frac{1}{2}y^2$. In addition, using the triangle inequality, we have

$$\|P_{XY|Z=m}\|_{2,\boldsymbol{a}_m}^2 \;\le\; 2\|P_{X,Y|m} - P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2 + 2\|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2.$$

Since we consider half-permutation method (Remark 6) for $T_{\mathrm{CI},W}$, we can apply the bound for the expected variance of a permuted U-statistic in Kim et al. (2020) to $U_W^2(\boldsymbol{W}_m^{\pi_m})$. More specifically, by the argument made in Appendix I of Kim et al. (2020) along with the above results, we have

$$\begin{aligned} &\mathbb{1}(\sigma_m \ge 4)\mathbb{E}\big[\mathbb{E}_{\boldsymbol{\pi}}\big[U_W^2(\boldsymbol{W}_m^{\pi_m})\big] | \boldsymbol{W}_{X,m}, \boldsymbol{W}_{Y,m}, \sigma_m\big] \\ \lesssim\; &\mathbb{1}(\sigma_m \ge 4)\sigma_m^{-2} \max\big\{\|P_{XY|Z=m}\|_{2,\boldsymbol{a}_m}^2, \|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2\big\} \\ \lesssim\; &\mathbb{1}(\sigma_m \ge 4)\sigma_m^{-2}\big\{\|P_{X,Y|m} - P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2 + \|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2\big\}. \end{aligned}$$

This result together with (51) leads to the following bound

$$\mathbb{E}\Big[\mathbb{E}_{\boldsymbol{\pi}}\big[(T_{\mathrm{CI},W}^{\boldsymbol{\pi}})^2\big] | \boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\Big] \;\lesssim\; \big\{(\mathrm{I}) + (\mathrm{II})\big\},$$

where

$$\begin{aligned} (\mathrm{I}) \;&:=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\omega_m^2 \|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2, \\ (\mathrm{II}) \;&:=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\omega_m^2 \|P_{X,Y|m} - P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2. \end{aligned}$$

To simplify the first term, the result of Canonne et al. (2018) and Neykov et al. (2021) yields

$$\mathbb{E}[\|P_{X|Z=m} P_{Y|Z=m}\|_{2,\boldsymbol{a}_m}^2] \;\le\; \frac{1}{(1+t_{1,m})(1+t_{2,m})},$$

which further shows that

$$\mathbb{E}\big[(\mathrm{I}) | \boldsymbol{\sigma}\big] \;\le\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\frac{\omega_m^2}{(1+t_{1,m})(1+t_{2,m})} \;\lesssim\; \min\{N, M\}.$$

Recall that $N \sim \mathrm{Pois}(n)$, which highly concentrates around its mean. Thus, by Markov's inequality as well as the union bound, one can have $(\mathrm{I}) \lesssim \min\{n, M\}$ with probability at least $1 - 1/400$. For

the second term, since $\omega_m \le \sigma_m$, we have (II) $\le \mathbb{E}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big]$. Therefore, with probability at least $1 - 1/400$, it holds that

$$\text{(I)} + \text{(II)} \ \lesssim \ \min\{n, M\} + \mathbb{E}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big],$$

which proves the inequality (50).

2. Proof of the second claim (48). Next we prove the second claim (48) under the condition on ϵ imposed in Canonne et al. (2018). In particular, for $\ell_1 \ge \ell_2$ and $\inf_{Q\in\mathcal{P}_{0,[M]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \ge \varepsilon$, equation (1) of Canonne et al. (2018) assumes that ε satisfies

$$\begin{aligned} n \ \ge \ \zeta' \times \max\Bigg(& \min\left\{ \frac{M^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon}, \frac{M^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}} \right\}, \\ & \frac{M^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon}, \frac{M^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}}, \frac{M^{1/2}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon^2} \Bigg), \end{aligned} \tag{52}$$

where ζ' is a sufficiently large positive constant. A sufficient condition for this inequality in the form of $\varepsilon \gtrsim f(M, \ell_1, \ell_2, n)$ is

$$\begin{aligned} \varepsilon \ \ge \ \zeta'' \times \max\Bigg(& \frac{M^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{n}, \frac{M^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{n}, \frac{M^{1/2}\ell_1^{1/2}\ell_2^{1/4}}{n^{3/4}}, \\ & \frac{M^{1/4}\ell_1^{1/4}\ell_2^{1/4}}{n^{1/2}}, \frac{M^{3/4}\ell_1^{1/4}\ell_2^{1/4}}{n^{7/8}} \Bigg), \end{aligned}$$

where ζ'' is some large positive constant. Let us define an event $\mathcal{E}$ such that $\mathbb{E}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big] \gtrsim \sqrt{\zeta' \min\{n, M\}}$ and

$$\begin{aligned} & \mathrm{Var}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big] \\ \lesssim \ & \min\{n, M\} + \sqrt{\min\{n, M\}}\mathbb{E}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big] + \mathbb{E}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big]^{3/2}. \end{aligned}$$

If $\inf_{Q\in\mathcal{P}_{0,[M]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \ge \varepsilon$, Lemma 5.4 of Canonne et al. (2018) guarantees that $\mathbb{P}(\mathcal{E}) \ge 399/400$. Under this event, by choosing ζ' sufficiently large depending on ζ, we see that

$$\zeta\sqrt{\min\{n, M\} + \mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]} \ \le \ \frac{1}{2}\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}].$$

Thus

$$\begin{aligned} & \mathbb{P}\Big[T_{\mathrm{CI},W} < \zeta\sqrt{\min\{n, M\} + \mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]}\Big] \\ \le \ & \mathbb{P}\Big[T_{\mathrm{CI},W} \le \frac{1}{2}\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]|\mathcal{E}\Big] + \frac{1}{400} \\ \le \ & \frac{1}{200}, \end{aligned}$$

where the last inequality uses Chebyshev's inequality, as detailed in the proof of Lemma 5.6 in Canonne et al. (2018).

• **Type II error of $\phi_{\mathrm{perm},1}$.** The type II error bound for $\phi_{\mathrm{perm},1}$ follows similarly as $\phi_{\mathrm{perm},2}$. Indeed it is simpler to prove the result of $\phi_{\mathrm{perm},1}$ than $\phi_{\mathrm{perm},2}$ as T_{CI} does not involve randomness from weights. We note from Lemma 6 again that the type II error of $\phi_{\mathrm{perm},1}$ is equivalent to $\mathbb{P}_{P_{X,Y,Z}^N,N}[T_{\mathrm{CI}} \le q_{1-\alpha}]$ where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of $T_{\mathrm{CI}}^{\boldsymbol{\pi}^1}, \ldots, T_{\mathrm{CI}}^{\boldsymbol{\pi}^K}$ (with some abuse of notation). We then prove the results in two steps as before. In the first step, we show that

$$\mathbb{P}_{P_{X,Y,Z}^N,N}[T_{\mathrm{CI}} \le q_{1-\alpha}] \;\le\; \mathbb{P}_{P_{X,Y,Z}^N,N}\big[T_{\mathrm{CI}} < \zeta\sqrt{\min\{n,M\}}\big] + \frac{1}{200}. \tag{53}$$

In the second step, we verify that

$$\mathbb{P}_{P_{X,Y,Z}^N,N}\big[T_{\mathrm{CI}} < \zeta\sqrt{\min\{n,M\}}\big] \le \frac{1}{200}, \tag{54}$$

under the condition on ε imposed by Canonne et al. (2018).

1. Proof of the first claim (53). To start with the first claim (53), by Chebyshev's inequality along with Markov's inequality, we have

$$q_{1-\alpha} \le \sqrt{200\alpha^{-1}\mathbb{E}\big\{\mathbb{E}_{\boldsymbol{\pi}}\big[(T_{\mathrm{CI}}^{\boldsymbol{\pi}})^2\big]\big\}},$$

with probability at least $1 - 1/200$. Again, by the independence between $\pi_1, \ldots, \pi_M$ and following the analysis in the proof of Theorem 4 — especially the proof of the inequality (29c), the conditional expectation becomes

$$\begin{aligned}
\mathbb{E}\big\{\mathbb{E}_{\boldsymbol{\pi}}\big[(T_{\mathrm{CI}}^{\boldsymbol{\pi}})^2\big]|\boldsymbol{\sigma}\big\} \;&=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m^2\mathbb{E}\big[U_W^2(\boldsymbol{W}_m^{\pi_m})|\boldsymbol{\sigma}\big] \\
&\overset{\text{(i)}}{=} \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m^2\mathbb{E}\big[\mathrm{Var}_{\boldsymbol{\pi}}\{U_W(\boldsymbol{W}_m^{\pi_m})|\boldsymbol{W}_m,\boldsymbol{\sigma}\}|\boldsymbol{\sigma}\big] \\
&\overset{\text{(ii)}}{\lesssim} \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 1) \overset{\text{(iii)}}{\lesssim} \min\{N,M\},
\end{aligned} \tag{55}$$

where step (i) uses the fact that $\mathbb{E}_{\boldsymbol{\pi}}\{U_W(\boldsymbol{W}_m^{\pi_m})|\boldsymbol{W}_m,\boldsymbol{\sigma}\} = 0$, step (ii) follows based on the analysis in the proof of the inequality (29c) and step (iii) follows by combining the two inequalities $\sum_{m\in[M]}\mathbb{1}(\sigma_m \ge 1) \le \sum_{m\in[M]} 1 = M$ and $\sum_{m\in[M]}\mathbb{1}(\sigma_m \ge 1) \le \sum_{m\in[M]}\sigma_m = N$. By taking the expectation over $\boldsymbol{\sigma}$, we then have $\mathbb{E}\big\{\mathbb{E}_{\boldsymbol{\pi}}\big[(T_{\mathrm{CI}}^{\boldsymbol{\pi}})^2\big]\big\} \lesssim \min\{n,M\}$. Therefore, the inequality (53) holds by taking ζ sufficiently large and the union bound.

2. Proof of the second claim (54). The proof of this result directly follows by Section 3.1 of Canonne et al. (2018). In particular, for fixed ℓ_1 and ℓ_2, and $\inf_{Q\in\mathcal{P}_{0,[M]}}\mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z},Q) \ge \varepsilon$,

Canonne et al. (2018) assume that ε satisfies

$$n \;\geq\; \zeta' \times \max\left(\frac{M^{1/2}}{\varepsilon^2}, \min\left(\frac{M^{7/8}}{\varepsilon}, \frac{M^{6/7}}{\varepsilon^{8/7}}\right)\right), \tag{56}$$

where ζ' is a sufficiently large positive constant. A sufficient condition for this inequality in the form of $\varepsilon \gtrsim f(M,n)$ is

$$\varepsilon \;\geq\; \zeta'' \max\left(\frac{M^{1/4}}{n^{1/2}}, \frac{M^{7/8}}{n}, \frac{M^{3/4}}{n^{7/8}}\right),$$

for some large $\zeta'' > 0$ as mentioned in the main text. Under this condition, Section 3.1 of Canonne et al. (2018) shows that the second claim (54) holds by taking ζ sufficiently large. Finally, combining the two inequalities (53) and (54) yields the desired result. This completes the proof of Theorem 5.

A.7 Proof of Theorem 6

We first study the type II error bound of $\phi^{\dagger}_{\mathrm{perm},1}$ and then turn to the type II error bound of $\phi^{\dagger}_{\mathrm{perm},2}$.

• **Type II error of $\phi^{\dagger}_{\mathrm{perm},1}$.** The proof of this part is similar to that of Theorem 5, especially the result of $\phi_{\mathrm{perm},1}$. The only difference is that we are now dealing with the discretized probability $Q_{X,Y,\widetilde{Z}}$ defined in Section 4.2, rather than the original probability $P_{X,Y,Z}$. Since the upper bound for the expected variance $\mathbb{E}\{\mathbb{E}_{\boldsymbol{\pi}}[(T^{\boldsymbol{\pi}}_{\mathrm{CI}})^2]|\boldsymbol{\sigma}\}$ in (55) does not depend on the underlying discrete distribution, the same proof carries through and we can obtain the bound (53) for $Q_{X,Y,\widetilde{Z}}$ as well. In particular, since we set $\min\{n, M\} \leq M = \lceil n^{2/5} \rceil$, we have

$$\mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi^{\dagger}_{\mathrm{perm},1}] \;\leq\; \mathbb{P}_{Q^N_{X,Y,\widetilde{Z}},N}\left[T_{\mathrm{CI}} < \zeta n^{1/5}\right] + \frac{1}{200} + e^{-n/8},$$

where we note that $\mathbb{P}_{Q^N_{X,Y,\widetilde{Z}},N}\left[T_{\mathrm{CI}} < \zeta n^{1/5}\right] = \mathbb{P}_{P^N_{X,Y,Z},N}\left[T_{\mathrm{CI}} < \zeta n^{1/5}\right]$ since T_{CI} is defined only through the binned data. We also note that $e^{-n/8}$ comes from the truncation probability of N as in (26). This probability is essentially the type II error of the test of Neykov et al. (2021) by adjusting the value of ζ, and thus we have

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1],\mathrm{TV}}(L): \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \geq \varepsilon} \mathbb{P}_{P^N_{X,Y,Z},N}\left[T_{\mathrm{CI}} < \zeta n^{1/5}\right] \;\leq\; \frac{1}{100}.$$

See the proof of Theorem 5.2 in Neykov et al. (2021) for more details. Therefore, we conclude that $\phi^{\dagger}_{\mathrm{perm},1}$ has the same power guarantee as $\phi_{\mathrm{NBW},1}$.

• **Type II error of $\phi^{\dagger}_{\mathrm{perm},2}$.** Given that $\phi^{\dagger}_{\mathrm{perm},2}$ uses the same test statistic $T_{\mathrm{CI},W}$ as in Theorem 5, the proof of this result is similar to that of Theorem 5, especially the result of $\phi_{\mathrm{perm},2}$. The only difference is, again, that we now work with the discretized probability distribution $Q_{X,Y,\widetilde{Z}}$ defined

in Section 4.2. Following exactly the same lines in the proof of Theorem 5, there exists an absolute constant $\zeta' > 0$ such that

$$\begin{aligned}
&\mathbb{E}_{P^N_{X,Y,Z},N}[1 - \phi^{\dagger}_{\mathrm{perm},2}] \\
\leq\ &\mathbb{P}_{Q^N_{X,Y,\widetilde{Z}},N}\Big\{T_{\mathrm{CI},W} < \zeta'\sqrt{\min\{n, M\} + \mathbb{E}_{Q^N_{X,Y,\widetilde{Z}},N}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]}\Big\} + \frac{1}{200} + e^{-n/8},
\end{aligned}$$

where the second line also holds by replacing $Q^N_{X,Y,\widetilde{Z}}$ with $P^N_{X,Y,Z}$ in the probability and the expectation since $T_{\mathrm{CI},W}$ is defined only through the binned data.

To further bound the given probability, we recall the results in Neykov et al. (2021). By ignoring the dependence on $Q^N_{X,Y,\widetilde{Z}}$, the following two bounds hold with probability at least 399/400 under the given conditions for the guarantee (20):

$$\begin{aligned}
&\mathrm{Var}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big] \\
\lesssim\ &M + (\sqrt{M} + 1)\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}] + [T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]^{3/2}
\end{aligned}$$

and

$$\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}] \gtrsim \sqrt{\zeta M},$$

where $\zeta > 0$ is a sufficiently large constant given in $\phi_{\mathrm{NBW},2}$, which differs from ζ' above. See Lemma C.10 of Neykov et al. (2021) for details on these bounds. Let $\mathcal{E}$ be an event that both inequalities hold simultaneously. Under this event, by taking ζ sufficiently large depending on ζ', we have

$$\zeta'\sqrt{\min\{n, M\} + \mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]} \ \leq\ \frac{1}{2}\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}].$$

Then by Chebyshev's inequality and Lemma C.13 of Neykov et al. (2021),

$$\begin{aligned}
&\mathbb{P}\Big[T_{\mathrm{CI},W} < \zeta'\sqrt{\min\{n, M\} + \mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]}\Big] \\
\leq\ &\mathbb{P}\Big[T_{\mathrm{CI},W} \leq \frac{1}{2}\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]|\mathcal{E}\Big] + \frac{1}{400} \\
\leq\ &\frac{1}{200}.
\end{aligned}$$

This completes the proof of the second part.

A.8 Proof of Theorem 7

The proof of this result is essentially the same as that of the second part of Theorem 6, which builds on the proof of Theorem 5. The only difference is that we are now in a situation where X, Y, Z are all discretized. Under this binning scheme, the mean and the variance of $T_{\mathrm{CI},W}$ have been studied in Neykov et al. (2021). In particular, Lemma G.9 of Neykov et al. (2021) yields that the following two events hold under the conditions of the theorem with probability at least 399/400:

$$\mathrm{Var}\big[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}\big]$$

$$\lesssim \big\{M + (\sqrt{M}+1)\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}] + [T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}]^{3/2}\big\}$$

and

$$\mathbb{E}[T_{\mathrm{CI},W}|\boldsymbol{\mathcal{W}}_{\mathrm{weight}}, \boldsymbol{\sigma}] \gtrsim \sqrt{\zeta M},$$

where $\zeta > 0$ is a sufficiently large constant given in $\phi_{\mathrm{NBW},2}$. Building upon this result, we can proceed by following the same lines of the proof of Theorem 6 to complete the proof. We omit the details.

A.9 Proof of Proposition 1

The proof of this result follows by modifying the proof of the validity of usual permutation tests based on exchangeability (e.g. Lehmann and Romano, 2006). Here we provide a detail for completeness. Let $\mathcal{X} := \{(X_i, Y_i, \widetilde{Z}_i)\}_{i=1}^n \sim \widetilde{Q}^n_{X,Y,\widetilde{Z}}$ be a set of n i.i.d. draws from the CI projection $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$. For each $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \boldsymbol{\Pi}_{\mathrm{cyclic}}$, write $\mathcal{X}^{\boldsymbol{\pi}} = \{(X_i, Y_{\boldsymbol{\pi}(i)}, \widetilde{Z}_i)\}_{i=1}^n$ and $\mathcal{X}^{\boldsymbol{\pi}\circ\boldsymbol{\pi}'} = \{(X_i, Y_{\boldsymbol{\pi}\circ\boldsymbol{\pi}'(i)}, \widetilde{Z}_i)\}_{i=1}^n$ where $\boldsymbol{\pi}\circ\boldsymbol{\pi}'$ denotes the composition of two permutation maps $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$. Throughout this proof, we use the notation $T(\mathcal{X}^{\boldsymbol{\pi}}) = T^{\boldsymbol{\pi}}_{\mathrm{CI}}$ and $T(\mathcal{X}) = T_{\mathrm{CI}}$. Let $T^{(1)}(\mathcal{X}) \leq \ldots \leq T^{(K_*)}(\mathcal{X})$ be the ordered statistics of $T(\mathcal{X}^{\boldsymbol{\pi}_1}), \ldots, T(\mathcal{X}^{\boldsymbol{\pi}_{K_*}})$ where $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K_*}\} = \boldsymbol{\Pi}_{\mathrm{cyclic}}$. Note that the set $\{T(\mathcal{X}^{\boldsymbol{\pi}_1}), \ldots, T(\mathcal{X}^{\boldsymbol{\pi}_{K_*}})\}$ has the same components as $\{T(\mathcal{X}^{\boldsymbol{\pi}_1\circ\boldsymbol{\pi}}), \ldots, T(\mathcal{X}^{\boldsymbol{\pi}_{K_*}\circ\boldsymbol{\pi}})\}$ for any $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\mathrm{cyclic}}$. Therefore, $T^{(i)}(\mathcal{X}) = T^{(i)}(\mathcal{X}^{\boldsymbol{\pi}})$ for any $i = 1, \ldots, K_*$ and $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\mathrm{cyclic}}$.

Having the above preliminaries in place, suppose we reject the null when $T(\mathcal{X}) > T^{(k)}(\mathcal{X})$ where $k = \lceil (1-\alpha)K_* \rceil$. Then, by letting $\boldsymbol{\pi}_0$ have the uniform distribution on $\boldsymbol{\Pi}_{\mathrm{cyclic}}$, we have

$$\begin{aligned}\mathbb{P}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}\big\{T(\mathcal{X}) > T^{(k)}(\mathcal{X})\big\} &\overset{\text{(i)}}{=} \mathbb{P}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}\big\{T(\mathcal{X}^{\boldsymbol{\pi}_0}) > T^{(k)}(\mathcal{X}^{\boldsymbol{\pi}_0})\big\} \\ &\overset{\text{(ii)}}{=} \mathbb{P}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}\big\{T(\mathcal{X}^{\boldsymbol{\pi}_0}) > T^{(k)}(\mathcal{X})\big\},\end{aligned}$$

where step (i) uses the fact that $\mathcal{X}$ and $\mathcal{X}^{\boldsymbol{\pi}}$ have the same distribution for all $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\mathrm{cyclic}}$ under the law $\widetilde{Q}^n_{X,Y,\widetilde{Z}}$, and step (ii) holds due to $T^{(k)}(\mathcal{X}) = T^{(k)}(\mathcal{X}^{\boldsymbol{\pi}})$. Now since $\boldsymbol{\pi}_0$ is uniformly distributed on $\boldsymbol{\Pi}_{\mathrm{cyclic}}$, we have

$$\mathbb{P}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}\big\{T(\mathcal{X}^{\boldsymbol{\pi}_0}) > T^{(k)}(\mathcal{X})\big\} = \mathbb{E}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}\left[\frac{1}{K_*}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}} \mathbb{1}\big\{T(\mathcal{X}^{\boldsymbol{\pi}_i}) > T^{(k)}(\mathcal{X})\big\}\right] \leq \alpha,$$

where the last inequality holds since

$$\frac{1}{K_*}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}} \mathbb{1}\big\{T(\mathcal{X}^{\boldsymbol{\pi}_i}) > T^{(k)}(\mathcal{X})\big\} \leq \alpha.$$

Finally, by Lemma 6, the permutation test from Algorithm 2 can be equivalently written as $\phi_{\mathrm{perm},n} = \mathbb{1}\{T(\mathcal{X}) > T^{(k)}(\mathcal{X})\}$. Therefore, we have

$$\mathbb{E}_{\widetilde{Q}^n_{X,Y,\widetilde{Z}}}[\phi_{\mathrm{perm},n}] \leq \alpha.$$

Having this inequality, the other steps are exactly the same as the proofs of Theorem 2 and Theorem 3. This completes the proof of Proposition 1.

A.10 Proof of Theorem 8

The main idea of the proof of this result is similar to that of Theorem 6. However, it requires a much more involved analysis for the mean and the variance due to a sophisticated dependence structure among local permutations over smaller bins. For the single-binning method, there is a symmetric structure of the U-statistic under the (global) permutation, which allows us to show that the mean of the permuted U-statistic is zero and its variance is bounded by some tractable quantities (e.g. Theorem 5.1 of Kim et al., 2020). However, this is no longer the case for the double-binning method, which leads to a technical challenge.

Throughout this proof, we only deal with the type II error (or $1-$power) of $\phi_{\mathrm{perm},N}$ since the result of $\phi^{\dagger}_{\mathrm{perm},N}$ follows directly by the inequality (26). For a sufficiently large $\zeta > 0$ (depending on L), Neykov et al. (2021) prove that the test $\phi'_{\mathrm{NBW},1} = \mathbb{1}(T_{\mathrm{CI}} \geq \zeta n^{1/5})$ has a small type II error in the sense of (18) (indeed, without $e^{-n/8}$ term) when $\varepsilon \geq cn^{-2/5}$ for a sufficiently large c depending on $(\zeta, L, \ell_1, \ell_2)$. Moreover, since $\mathcal{P}'_{1,[0,1],\mathrm{TV}}(L) \subseteq \mathcal{P}_{1,[0,1],\mathrm{TV}}(L)$, it holds that

$$\sup_{P_{X,Y,Z} \in \mathcal{P}'_{1,[0,1],\mathrm{TV}}(L) : \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathcal{D}_{\mathrm{TV}}(P_{X,Y,Z}, Q) \geq \varepsilon} \mathbb{P}_{P^N_{X,Y,Z},N}[T_{\mathrm{CI}} \geq \zeta n^{1/5}] \leq \frac{1}{200}. \tag{57}$$

Now let $q_{1-\alpha}$ be the $1-\alpha$ quantile of the empirical distribution of $T^{\boldsymbol{\pi}_1}_{\mathrm{CI}}, \ldots, T^{\boldsymbol{\pi}_{K_*}}_{\mathrm{CI}}$ where $\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}$. Our main goal is to show that $q_{1-\alpha} \leq \zeta' n^{1/5}$ for a sufficiently large ζ' with high probability (say $1/200$) under the given conditions. Since the type II error of $\phi_{\mathrm{perm},N}$ can be written as $\mathbb{P}(T_{\mathrm{CI}} \leq q_{1-\alpha})$ in view of Lemma 6, this bound for the quantile together with the union bound implies that the type II error of $\phi_{\mathrm{perm},N}$ is smaller than $1/100$ by taking ζ sufficiently large.

We break down the proof into several pieces for readability.

- **Step 1. Bounding the quantile.** In the first step, we prove that the $1-\alpha$ quantile is upper bounded by

$$q_{1-\alpha} \lesssim \mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\mathrm{Var}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]},$$

  with high probability where $\boldsymbol{\pi} \sim F_{\boldsymbol{\pi}}$ defined in (59) below. Here we note that the expectation and the variance are with respect to both randomness from $\boldsymbol{\pi}$ and samples.

- **Step 2. Bounding the expectation.** In the second step, we upper bound the expectation of $T^{\boldsymbol{\pi}}_{\mathrm{CI}}$ where $\boldsymbol{\pi} \sim F_{\boldsymbol{\pi}}$ in (59) and prove that

$$\mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] = O\left(\frac{n}{M^2} + 1 + \frac{M^2}{n}\right),$$

- **Step 3. Bounding the variance.** In the third step, we upper bound the variance of $T^{\boldsymbol{\pi}}_{\mathrm{CI}}$ where $\boldsymbol{\pi} \sim F_{\boldsymbol{\pi}}$ in (59) and prove that

$$\mathrm{Var}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] = O\left(\frac{n^2}{M^4} + \frac{n}{M^2} + \frac{M^2}{n} + M\right).$$

- **Step 4. Completing the proof.** Lastly, we combine the previous results and prove the desired type II error bound for $\phi_{\mathrm{perm},N}$.

Step 1. Bounding the quantile. We start by bounding the $1-\alpha$ quantile

$$q_{1-\alpha} = \inf\left\{x \in \mathbb{R} : 1-\alpha \le \frac{1}{K_*}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \le x\big)\right\}.$$

Consider a subset of $\boldsymbol{\Pi}_{\mathrm{cyclic}}$ where we remove the identity permutation within each small bin. Let us denote the resulting subset by $\boldsymbol{\Pi}_{\mathrm{cyclic}}^{-\mathrm{id}}$ whose cardinality is $K_*^{-\mathrm{id}} := \prod_{(m,k)\in[M]\times[b]} \max\{\sigma_{m,k}-1, 1\}$. By letting α_* such that $1-\alpha_* = \min\big\{(1-\alpha)K_*/K_*^{-\mathrm{id}}, 1\big\}$, define

$$\begin{aligned} q_{1-\alpha_*}^{-\mathrm{id}} &= \inf\left\{x \in \mathbb{R} : \min\left\{(1-\alpha), \frac{K_*^{-\mathrm{id}}}{K_*}\right\} \le \frac{1}{K_*}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}^{-\mathrm{id}}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \le x\big)\right\} \\ &= \inf\left\{x \in \mathbb{R} : \min\left\{(1-\alpha)\frac{K_*}{K_*^{-\mathrm{id}}}, 1\right\} \le \frac{1}{K_*^{-\mathrm{id}}}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}^{-\mathrm{id}}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \le x\big)\right\}. \end{aligned}$$

Since $\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \le x\big)$ is a sum of non-negative numbers, we have $q_{1-\alpha} \le q_{1-\alpha_*}^{-\mathrm{id}}$. Therefore

$$\mathbb{P}(T_{\mathrm{CI}} \le q_{1-\alpha}) \;\le\; \mathbb{P}(T_{\mathrm{CI}} \le q_{1-\alpha_*}^{-\mathrm{id}}).$$

Let us define an event $\mathcal{A}$ as

$$\mathcal{A} = \left\{\frac{K_*}{K_*^{-\mathrm{id}}} \le \frac{1-\alpha/2}{1-\alpha}\right\}, \tag{58}$$

and let $q_{1-\alpha/2}^{-\mathrm{id}}$ be the $1-\alpha/2$ quantile of the empirical distribution

$$F_{\boldsymbol{\pi}}(x) := \frac{1}{K_*^{-\mathrm{id}}}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}_{\mathrm{cyclic}}^{-\mathrm{id}}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \le x\big). \tag{59}$$

Then, since $q_{1-\alpha_*}^{-\mathrm{id}} \le q_{1-\alpha/2}^{-\mathrm{id}}$ under $\mathcal{A}$, we have

$$\mathbb{P}(T_{\mathrm{CI}} \le q_{1-\alpha_*}^{-\mathrm{id}}) \;\le\; \mathbb{P}(T_{\mathrm{CI}} \le q_{1-\alpha/2}^{-\mathrm{id}}) + \mathbb{P}(\mathcal{A}^c).$$

In what follows, we use the notation $\mathbb{E}_{\boldsymbol{\pi}}[\cdot]$ and $\mathrm{Var}_{\boldsymbol{\pi}}[\cdot]$ to denote the expectation and variance operators with respect to $F_{\boldsymbol{\pi}}$ and let $\mathbb{P}_{\boldsymbol{\pi}}[\cdot] = \mathbb{E}_{\boldsymbol{\pi}}[\mathbb{1}(\cdot)]$. Given this notation, we further upper bound $q_{1-\alpha/2}^{-\mathrm{id}}$. To this end, we follow the proof of Lemma 3.1 of Kim et al. (2020). First, for $t > 0$, by Chebyshev's inequality,

$$\mathbb{P}_{\boldsymbol{\pi}}(T_{\mathrm{CI}}^{\boldsymbol{\pi}} - \mathbb{E}_{\boldsymbol{\pi}}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}] \ge t) \le \frac{1}{t^2}\mathrm{Var}_{\boldsymbol{\pi}}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}].$$

By setting the right-hand side to be $\alpha/2$, we see that $q^{-\mathrm{id}}_{1-\alpha/2}$ is deterministically bounded by

$$q^{-\mathrm{id}}_{1-\alpha/2} \le \mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\frac{2}{\alpha}\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]}.$$

Next applying Chebyshev's inequality and Markov's inequality yields

$$\mathbb{P}\left(\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] \ge \mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\frac{2}{\beta}\mathrm{Var}\big\{\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\}}\right) \le \frac{\beta}{2} \quad \text{and}$$

$$\mathbb{P}\left(\frac{2}{\alpha}\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] \ge \frac{4}{\alpha\beta}\mathbb{E}\big\{\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\}\right) \le \frac{\beta}{2}.$$

From the above observation together with the union bound, it holds with probability at least $1-\beta$ that

$$\begin{aligned} q^{-\mathrm{id}}_{1-\alpha/2} &\le \mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\frac{2}{\beta}\mathrm{Var}\big\{\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\}} + \sqrt{\frac{4}{\alpha\beta}\mathbb{E}\big\{\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\}} \\ &\le \mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\frac{32}{\alpha\beta}\mathrm{Var}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]}, \end{aligned}$$

where the last inequality uses $\mathrm{Var}\big\{\mathbb{E}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\} + \mathbb{E}\big\{\mathrm{Var}_{\boldsymbol{\pi}}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]\big\} \le \mathrm{Var}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]$. In summary, by taking $\alpha = 1/100$ and $\beta = 1/200$, we have obtained that

$$\mathbb{P}(T_{\mathrm{CI}} \le q_{1-\alpha}) \le \mathbb{P}\Big(T_{\mathrm{CI}} \lesssim \mathbb{E}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}] + \sqrt{\mathrm{Var}[T^{\boldsymbol{\pi}}_{\mathrm{CI}}]}\Big) + \frac{1}{200} + \mathbb{P}(\mathcal{A}^c). \tag{60}$$

Step 2. Upper bounding the expectation. Recall that

$$T^{\boldsymbol{\pi}}_{\mathrm{CI}} = \sum_{m\in[M]} \mathbb{1}(\sigma_m \ge 4)\sigma_m U(\boldsymbol{W}^{\boldsymbol{\pi}}_m),$$

where $\boldsymbol{\pi}$ is uniformly distributed on $\boldsymbol{\Pi}^{-\mathrm{id}}_{\mathrm{cyclic}}$. When $\sigma_m < 4$, $\mathbb{1}(\sigma_m \ge 4)\sigma_m U(\boldsymbol{W}^{\boldsymbol{\pi}}_m)$ becomes zero, which does not contribute to $T^{\boldsymbol{\pi}}_{\mathrm{CI}}$. Hence, we assume $\sigma_m \ge 4$ and show that

$$\mathbb{E}[U(\boldsymbol{W}^{\boldsymbol{\pi}}_m)|\sigma_m] = O\left(\frac{1}{M^2} + \frac{1}{M\sigma_m} + \frac{1}{\sigma_m^2}\right) \quad \text{for each } m \in [M]. \tag{61}$$

Recall $U(\boldsymbol{W}^{\boldsymbol{\pi}}_m)$ is the U-statistic with the kernel that takes the form

$$\begin{aligned} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} := \sum_{x\in[\ell_1],y\in[\ell_2]} &\big\{\mathbb{1}(X_{i,m} = x, Y_{\pi_{i,m}} = y) - \mathbb{1}(X_{i,m} = x)\mathbb{1}(Y_{\pi_{j,m}} = y)\big\}\times \\ &\big\{\mathbb{1}(X_{k,m} = x, Y_{\pi_{k,m}} = y) - \mathbb{1}(X_{k,m} = x)\mathbb{1}(Y_{\pi_{\ell,m}} = y)\big\}, \end{aligned}$$

where $\{i,j,k,\ell\}$ are distinct integers between 1 and σ_m, and $\{\pi_{i,m}, \pi_{j,m}, \pi_{k,m}, \pi_{\ell,m}\}$ are components of $\boldsymbol{\pi}$ that are used to permute the corresponding components of $\{Y_{i,m}, Y_{j,m}, Y_{k,m}, Y_{\ell,m}\}$. Due to a

complicated local structure of $\boldsymbol{\pi}$, it is not trivial to compute the expectation of this kernel. This is in contrast to the single-binning case where this expectation of the kernel is zero under the local permutation law. We instead upper bound the expectation using the following observation. Suppose that there is no collision, meaning that neither of $Y_{i,m}, Y_{k,m}$ corresponds to $Y_{\pi_{i,m}}, Y_{\pi_{j,m}}, Y_{\pi_{k,m}}, Y_{\pi_{\ell,m}}$. In this case, $\{X_{i,m}, X_{j,m}, Y_{\pi_{i,m}}, Y_{\pi_{j,m}}, Y_{\pi_{k,m}}, Y_{\pi_{\ell,m}}\}$ are mutually independent, and the expectation of the kernel is bounded above by

$$\begin{aligned}\left|\mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}|\sigma_m]\right| \;\le\; \Bigg| \max_{k_1,k_2,k_3,k_4\in[b]} \Bigg[\sum_{x\in[\ell_1],y\in[\ell_2]} \{p_{m,k_1}(x)q_{m,k_1}(y) - p_{m,k_1}(x)q_{m,k_2}(y)\}\times \\ \{p_{m,k_3}(x)q_{m,k_3}(y) - p_{m,k_3}(x)q_{m,k_4}(y)\}\Bigg]\Bigg|,\end{aligned} \tag{62}$$

where $p_{m,k}(x) := \mathbb{P}(X = x|Z \in B_{m,k})$ and $q_{m,k}(y) := \mathbb{P}(Y = y|Z \in B_{m,k})$. Furthermore, since we assume $P_{X,Y,Z} \in \mathcal{P}'_{1,[0,1],\mathrm{TV}(L)}$,

$$\begin{aligned}&\mathcal{D}_{\mathrm{TV}}(q_{m,k_1}, q_{m,k_2}) \\ =\;& \frac{1}{2}\sum_{y\in[\ell_2]} \left| \int_{B_{m,k_1}} p_{Y|Z}(y|z) d\widetilde{P}_{Z|Z\in B_{m,k_1}}(z) - \int_{B_{m,k_2}} p_{Y|Z}(y|z') d\widetilde{P}_{Z|Z\in B_{m,k_2}}(z') \right| \\ \le\;& \frac{1}{2}\int_{B_{m,k_1}}\int_{B_{m,k_2}} \sum_{y\in[\ell_2]} \left|p_{Y|Z}(y|z) - p_{Y|Z}(y|z')\right| d\widetilde{P}_{Z|Z\in B_{m,k_1}}(z) d\widetilde{P}_{Z|Z\in B_{m,k_2}}(z') \\ \le\;& \frac{1}{2}\max_{z\in B_{m,k_1}, z'\in B_{m,k_2}} \delta(z,z') \le \frac{C_1}{M},\end{aligned}$$

for some constant $C_1 > 0$. Therefore $\left|\mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}]\right|$ is further bounded by

$$\begin{aligned}&\left|\mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}|\sigma_m]\right| \\ \le\;& \left| \max_{k_1,k_2,k_3,k_4\in[b]} \left[\sum_{x\in[\ell_1]} p_{m,k_1}(x)p_{m,k_3}(x) \sum_{y\in[\ell_2]} |q_{m,k_1}(y) - q_{m,k_2}(y)| \sum_{y\in[\ell_2]} |q_{m,k_3}(y) - q_{m,k_4}(y)| \right]\right| \\ \le\;& 4\max_{k_1,k_2,k_3,k_4\in[b]} \mathcal{D}_{\mathrm{TV}}(q_{m,k_1}, q_{m,k_2}) \times \mathcal{D}_{\mathrm{TV}}(q_{m,k_3}, q_{m,k_4}) \\ \le\;& \frac{C_2}{M^2}.\end{aligned}$$

Suppose that there is only one collision, i.e. there exists only one dependent pair of (X, Y) among $\{X_{i,m}, X_{j,m}, Y_{\pi_{i,m}}, Y_{\pi_{j,m}}, Y_{\pi_{k,m}}, Y_{\pi_{\ell,m}}\}$. Since we deal with a cyclic permutation excluding the identity map, there are only a few possibilities of one collision. In particular, a cyclic permutation, which is not the identity map, is a derangement (i.e. no point will be at its original place). This guarantees that $X_{i,m}$ and $Y_{\pi_{i,m}}$ are independent for each i and m, which simplifies our analysis.

Case 1) Suppose that $X_{i,m}$ and $Y_{\pi_{j,m}}$ collide. In this case, $\{X_{k,m}, Y_{\pi_{k,m}}, Y_{\pi_{\ell,m}}\}$ are still independent, which leads to the bound of the expectation

$$\begin{aligned}
\big|\mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}|\sigma_m]\big| \;&:=\; \Big| \sum_{x\in[\ell_1],y\in[\ell_2]} \mathbb{E}\big\{\mathbb{1}(X_{i,m}=x, Y_{\pi_{i,m}}=y) - \mathbb{1}(X_{i,m}=x)\mathbb{1}(Y_{\pi_{j,m}}=y)\big\}\times \\
&\qquad\qquad \mathbb{E}\big\{\mathbb{1}(X_{k,m}=x, Y_{\pi_{k,m}}=y) - \mathbb{1}(X_{k,m}=x)\mathbb{1}(Y_{\pi_{\ell,m}}=y)\big\}\Big| \\
&\le\; 2 \sum_{x\in[\ell_1],y\in[\ell_2]} \big|\mathbb{E}\big\{\mathbb{1}(X_{k,m}=x, Y_{\pi_{k,m}}=y) - \mathbb{1}(X_{k,m}=x)\mathbb{1}(Y_{\pi_{\ell,m}}=y)\big\}\big| \\
&\le\; 2 \sup_{k_1,k_2\in[b]} \sum_{x\in[\ell_1],y\in[\ell_2]} p_{m,k_1}(x) \times \big|q_{m,k_1}(y) - q_{m,k_2}(y)\big| \\
&\le\; \frac{C_3}{M},
\end{aligned}$$

where the last inequality follows similarly as the previous analysis.

Case 2) Next, suppose that $X_{i,m}$ and $Y_{\pi_{k,m}}$ collide. In this case, $Y_{\pi_{i,m}}$ and $Y_{\pi_{j,m}}$ are independent of the others. Using this property, the expectation is bounded by

$$\begin{aligned}
\big|\mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}|\sigma_m]\big| \;&:=\; \Big| \sum_{x\in[\ell_1],y\in[\ell_2]} \mathbb{E}\big\{\mathbb{1}(Y_{\pi_{i,m}}=y) - \mathbb{1}(Y_{\pi_{j,m}}=y)\big\}\mathbb{E}[\mathbb{1}(X_{k,m}=x)]\times \\
&\qquad\qquad \mathbb{E}\big[\mathbb{1}(X_{i,m}=x)\big\{\mathbb{1}(Y_{\pi_{k,m}}=y) - \mathbb{1}(Y_{\pi_{\ell,m}}=y)\big\}\big]\Big| \\
&\le\; 2 \sup_{k_1,k_2\in[b]} \sum_{x\in[\ell_1],y\in[\ell_2]} p_{m,k_1}(x) \times \big|q_{m,k_1}(y) - q_{m,k_2}(y)\big| \\
&\le\; \frac{C_3}{M}.
\end{aligned}$$

The other cases with one collision can be bounded similarly by $O(M^{-1})$.

Now consider another representation of the U-statistic:

$$U(\boldsymbol{W}^{\boldsymbol{\pi}}_m) \;=\; \frac{(\sigma_m-4)!}{\sigma_m!} \sum_{\{i,j,k,\ell\}\in \boldsymbol{i}_4^{\sigma_m}} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell},$$

where $\boldsymbol{i}_4^{\sigma_m}$ denotes the set of all 4-tuples drawn without replacement from $[\sigma_m]$. Next we decompose the summation into three cases depending on whether there exists a collision among

$$\{(X_{i,m}, Y_{\pi_{i,m}}), (X_{j,m}, Y_{\pi_{j,m}}), (X_{k,m}, Y_{\pi_{k,m}}), (X_{\ell,m}, Y_{\pi_{\ell,m}})\}. \tag{63}$$

In particular,

$$\sum_{\{i,j,k,\ell\}\in \boldsymbol{i}_4^{\sigma_m}} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} \;=\; \sum_{\{i,j,k,\ell\}\in \boldsymbol{i}_4^{\sigma_m}\cap A^0_{\text{collision}}} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} + \sum_{\{i,j,k,\ell\}\in \boldsymbol{i}_4^{\sigma_m}\cap A^1_{\text{collision}}} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}$$

$$+ \sum_{\{i,j,k,\ell\} \in \boldsymbol{i}_4^{\sigma_m} \cap A^{1+}_{\text{collision}}} \varphi^{\boldsymbol{\pi}}_{i,j,k,\ell},$$

where 1) $\boldsymbol{i}_4^{\sigma_m} \cap A^0_{\text{collision}}$ is the subset of $\boldsymbol{i}_4^{\sigma_m}$ such that there exists no collision, 2) $\boldsymbol{i}_4^{\sigma_m} \cap A^1_{\text{collision}}$ is the subset of $\boldsymbol{i}_4^{\sigma_m}$ such that there exists only one collision, and 3) $\boldsymbol{i}_4^{\sigma_m} \cap A^{1+}_{\text{collision}}$ is the subset of $\boldsymbol{i}_4^{\sigma_m}$ such that there exists more than one collision.

Now we upper bound each summation. First of all, by the previous analysis,

$$\frac{(\sigma_m - 4)!}{\sigma_m!} \sum_{\{i,j,k,\ell\} \in \boldsymbol{i}_4^{\sigma_m} \cap A^0_{\text{collision}}} \mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} | \sigma_m] \;=\; O\left(\frac{1}{M^2}\right).$$

Next, since $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{-\text{id}}_{\text{cyclic}}$ is a derangement by construction, if we pick four observations (63) with no collision (i.e. $\{i,j,k,\ell\}$ and $\{\pi_i, \pi_j, \pi_k, \pi_\ell\}$ have no shared component) out of σ_m observations, there are at least $\sigma_m(\sigma_m - 2)(\sigma_m - 4)(\sigma_m - 6)$ ways. Thus in order to have a collision, we have at most $O(\sigma_m^3)$ choices. Moreover, as analyzed before, the expectation of the kernel is bounded by $O(M^{-1})$ if there is one collision. Therefore

$$\frac{(\sigma_m - 4)!}{\sigma_m!} \sum_{\{i,j,k,\ell\} \in \boldsymbol{i}_4^{\sigma_m} \cap A^1_{\text{collision}}} \mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} | \sigma_m] \;=\; O\left(\frac{1}{M\sigma_m}\right).$$

Lastly, there are at most $\asymp \sigma_m^2$ cases to have more than one collision. In this case, using the fact that the kernel $\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell}$ is bounded between -1 and 1, we have

$$\frac{(\sigma_m - 4)!}{\sigma_m!} \sum_{\{i,j,k,\ell\} \in \boldsymbol{i}_4^{\sigma_m} \cap A^{1+}_{\text{collision}}} \mathbb{E}[\varphi^{\boldsymbol{\pi}}_{i,j,k,\ell} | \sigma_m] \;=\; O\left(\frac{1}{\sigma_m^2}\right).$$

These three bounds imply the claim (61). Therefore

$$\mathbb{E}[T^{\boldsymbol{\pi}}_{\text{CI}}] \;=\; O\left(\frac{n}{M^2} + 1 + \sum_{m \in [M]} \mathbb{E}\left[\mathbb{1}(\sigma_m \geq 4)\sigma_m^{-1}\right]\right).$$

By recalling $\sigma_m \sim \text{Pois}(np_m)$, we have the following inequalities

$$\mathbb{E}\left[\mathbb{1}(\sigma_m \geq 4)\frac{1}{\sigma_m}\right] \;\leq\; \frac{5}{4}\mathbb{E}\left[\mathbb{1}(\sigma_m \geq 4)\frac{1}{\sigma_m + 1}\right] \;\leq\; \frac{5}{4}\mathbb{E}\left[\frac{1}{\sigma_m + 1}\right]. \tag{64}$$

Furthermore,

$$\mathbb{E}\left[\frac{1}{\sigma_m + 1}\right] \;\leq\; \frac{1 - e^{-np_m}}{np_m} \;\leq\; \frac{M}{c_{\text{low}} n},$$

where the last inequality uses the condition $p_m \geq c_{\text{low}} M^{-1}$ for all $m \in [M]$. Therefore, we have

$$\mathbb{E}[T^{\boldsymbol{\pi}}_{\text{CI}}] \;=\; O\left(\frac{n}{M^2} + 1 + \frac{M^2}{n}\right).$$

Step 3. Upper bounding the variance. Next we upper bound the variance of the U-statistic $U(\boldsymbol{W}_m^{\boldsymbol{\pi}})$. To compute the variance of $U(\boldsymbol{W}_m^{\boldsymbol{\pi}})$, first set up some notation. Similarly as in Section 4.3, let

$$\psi_{ij}^{m,\boldsymbol{\pi}}(x,y) = \mathbb{1}(X_{i,m} = x, Y_{\pi_{i,m}} = y) - \mathbb{1}(X_{i,m} = x)\mathbb{1}(Y_{\pi_{j,m}} = y),$$

and define a symmetrized kernel function as

$$h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}} = \frac{1}{4!}\sum_{\pi\in\boldsymbol{\Pi}_4}\sum_{x\in[\ell_1],y\in[\ell_2]}\psi_{\pi(1)\pi(2)}^{m,\boldsymbol{\pi}}(x,y)\psi_{\pi(3)\pi(4)}^{m,\boldsymbol{\pi}}(x,y)$$

where $\boldsymbol{\Pi}_4$ is the set of all permutations of $\{i_1,i_2,i_3,i_4\}$. Then the U-statistic can be written as

$$U(\boldsymbol{W}_m^{\boldsymbol{\pi}}) \;=\; \frac{1}{\binom{\sigma_m}{4}}\sum_{i_1<i_2<i_3<i_4:(i_1,i_2,i_3,i_4)\in[\sigma_m]} h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}},$$

and it variance has the form as

$$\mathrm{Var}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m] \;=\; \frac{1}{\binom{\sigma_m}{4}^2}\sum_{\substack{i_1<i_2<i_3<i_4\\:(i_1,i_2,i_3,i_4)\in[\sigma_m]}}\sum_{\substack{j_1<j_2<j_3<j_4\\:(j_1,j_2,j_3,j_4)\in[\sigma_m]}}\mathrm{Cov}(h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}}, h_{j_1,j_2,j_3,j_4}^{m,\boldsymbol{\pi}}).$$

Here and hereafter, we often ignore the conditional event σ_m when we write the expectation and the (co)variance operator for simplicity.

Now we decompose the summation into a few different parts depending on the number of common indices between $\{i_1,i_2,i_3,i_4\}$ and $\{j_1,j_2,j_3,j_4\}$. In particular, let A_{com}^k be the set of indices $\{i_1,i_2,i_3,i_4,j_1,j_2,j_3,j_4\} \in [\sigma_m]$ such that $i_1<i_2<i_3<i_4$, $j_1<j_2<j_3<j_4$ and there are k elements in common between $\{i_1,i_2,i_3,i_4\}$ and $\{j_1,j_2,j_3,j_4\}$. Using this notation, the variance can be written as

$$\begin{aligned}
&\mathrm{Var}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m]\\
=\;& \frac{1}{\binom{\sigma_m}{4}^2}\sum_{A_{\mathrm{com}}^0}\mathrm{Cov}(h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}}, h_{j_1,j_2,j_3,j_4}^{m,\boldsymbol{\pi}}) + \frac{1}{\binom{\sigma_m}{4}^2}\sum_{A_{\mathrm{com}}^1}\mathrm{Cov}(h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}}, h_{j_1,j_2,j_3,j_4}^{m,\boldsymbol{\pi}})\\
+\;& \frac{1}{\binom{\sigma_m}{4}^2}\sum_{A_{\mathrm{com}}^{1+}}\mathrm{Cov}(h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}}, h_{j_1,j_2,j_3,j_4}^{m,\boldsymbol{\pi}}).
\end{aligned}$$

Since 1) the number of cases in which $\{i_1,i_2,i_3,i_4\}$ and $\{j_1,j_2,j_3,j_4\}$ have more than one common index is at most $\asymp \sigma_m^6$ and 2) that the kernel is bounded, the last term in the above display is bounded by

$$\frac{1}{\binom{\sigma_m}{4}^2}\sum_{A_{\mathrm{com}}^{1+}}\mathrm{Cov}(h_{i_1,i_2,i_3,i_4}^{m,\boldsymbol{\pi}}, h_{j_1,j_2,j_3,j_4}^{m,\boldsymbol{\pi}}) \;=\; O\left(\sigma_m^{-2}\right).$$

Next consider the cases in which $\{i_1,i_2,i_3,i_4\}$ and $\{j_1,j_2,j_3,j_4\}$ have the exactly one common index. Without loss of generality, we let $i_1 = j_1$ and consider the resulting pairs

$$\{(X_{i_1,m}, Y_{\pi_{i_1,m}}), (X_{i_2,m}, Y_{\pi_{i_2,m}}), (X_{i_3,m}, Y_{\pi_{i_3,m}}), (X_{i_4,m}, Y_{\pi_{i_4,m}}),$$

$$(X_{j_2,m}, Y_{\pi_{j_2,m}}), (X_{j_3,m}, Y_{\pi_{j_3,m}}), (X_{j_4,m}, Y_{\pi_{j_4,m}})\}.$$

For now, we assume that there is no collision among the above set. Then we can treat them as if they are i.i.d. observations with $X \perp\!\!\!\perp Y$. Conditional on this assumption of no collision, we can follow the proof of Lemma 5.1 of Neykov et al. (2021) and show

$$\mathrm{Cov}(h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}, h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}) \;=\; O(M^{-2}). \tag{65}$$

To prove this, first let $\boldsymbol{\tau} := \{\tau(1), \tau(2), \tau(3), \tau(4)\}$ be sampled uniformly from the set of all permutations of $\{i_1, i_2, i_3, i_4\}$. Similarly let $\boldsymbol{\tau}' := \{\tau'(1), \tau'(2), \tau'(3), \tau'(4)\}$ be sampled uniformly from the set of all permutations of $\{j_1, j_2, j_3, j_4\}$. Then by treating $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$ to be independent random vectors, we can re-write the expectation as

$$\begin{aligned}
&\mathbb{E}[h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4} h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}] \\
&= \mathbb{E}\Bigg[\sum_{x\in[\ell_1],y\in[\ell_2]} \psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y) \sum_{x'\in[\ell_1],y'\in[\ell_2]} \psi^{m,\boldsymbol{\pi}}_{\tau'(1)\tau'(2)}(x',y')\psi^{m,\boldsymbol{\pi}}_{\tau'(3)\tau'(4)}(x',y')\Bigg].
\end{aligned}$$

Using the condition that there exists exactly one common index between $\{\tau(1), \tau(2), \tau(3), \tau(4)\}$ and $\{\tau'(1), \tau'(2), \tau'(3), \tau'(4)\}$ (without loss of generality, let the shared common index be between $\{\tau(1), \tau(2)\}$ and $\{\tau'(1), \tau'(2)\}$), the above expectation becomes

$$\begin{aligned}
&\sum_{x\in[\ell_1],y\in[\ell_2]} \sum_{x'\in[\ell_1],y'\in[\ell_2]} \mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\psi^{m,\boldsymbol{\pi}}_{\tau'(1)\tau'(2)}(x',y')\big]\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\big]\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau'(3)\tau'(4)}(x',y')\big] \\
\leq\; &\sqrt{\sum_{x\in[\ell_1],y\in[\ell_2]} \sum_{x'\in[\ell_1],y'\in[\ell_2]} \Big\{\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\psi^{m,\boldsymbol{\pi}}_{\tau'(1)\tau'(2)}(x',y')\big]\Big\}^2} \times \\
&\sqrt{\sum_{x\in[\ell_1],y\in[\ell_2]} \sum_{x'\in[\ell_1],y'\in[\ell_2]} \Big\{\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\big]\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau'(3)\tau'(4)}(x',y')\big]\Big\}^2},
\end{aligned}$$

where the second step follows by Cauchy–Schwarz inequality. Now, based on the notation used in (62)

$$\begin{aligned}
\sum_{x\in[\ell_1],y\in[\ell_2]} \Big\{\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\big]\Big\}^2 \;&\leq\; \max_{k_1,k_2\in[b]} \sum_{x\in[\ell_1],y\in[\ell_2]} \Big\{p_{m,k_1}(x)q_{m,k_1}(x) - p_{m,k_1}(x)q_{m,k_2}(x)\Big\}^2 \\
&\leq\; \max_{k_1,k_2\in[b]} \mathcal{D}^2_{\mathrm{TV}}(q_{m,k_1}, q_{m,k_2}) \;\leq\; \frac{C_1}{M^2}.
\end{aligned} \tag{66}$$

This bound gives

$$\mathbb{E}[h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4} h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}] \;=\; O\left(M^{-2}\right). \tag{67}$$

Next we look at the product of the expectations

$$\mathbb{E}[h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}] \times \mathbb{E}[h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}]$$

$$= \mathbb{E}\left[\sum_{x\in[\ell_1],y\in[\ell_2]} \psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\right] \times \mathbb{E}\left[\sum_{x'\in[\ell_1],y'\in[\ell_2]} \psi^{m,\boldsymbol{\pi}}_{\tau'(1)\tau'(2)}(x',y')\psi^{m,\boldsymbol{\pi}}_{\tau'(3)\tau'(4)}(x',y')\right].$$

By focusing on the first term, Cauchy–Schwarz inequality together with the previous bound (66) yields

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{x\in[\ell_1],y\in[\ell_2]} \psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\right]\right| \\
&\leq \sqrt{\sum_{x\in[\ell_1],y\in[\ell_2]} \left\{\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(1)\tau(2)}(x,y)\big]\right\}^2} \times \sqrt{\sum_{x\in[\ell_1],y\in[\ell_2]} \left\{\mathbb{E}\big[\psi^{m,\boldsymbol{\pi}}_{\tau(3)\tau(4)}(x,y)\big]\right\}^2} \\
&\leq \frac{C_2}{M^2}.
\end{aligned}$$

The other term can be similarly analyzed and thus

$$\mathbb{E}[h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}] \times \mathbb{E}[h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}] \;=\; O\left(M^{-4}\right).$$

Combining the above with the bound (67) yields the claim (65).

Suppose that there are at least one collision among

$$\begin{aligned}
&\{(X_{i_1,m},Y_{\pi_{i_1,m}}),(X_{i_2,m},Y_{\pi_{i_2,m}}),(X_{i_3,m},Y_{\pi_{i_3,m}}),(X_{i_4,m},Y_{\pi_{i_4,m}}),\\
&(X_{j_1,m},Y_{\pi_{j_1,m}}),(X_{j_2,m},Y_{\pi_{j_2,m}}),(X_{j_3,m},Y_{\pi_{j_3,m}}),(X_{j_4,m},Y_{\pi_{j_4,m}})\}.
\end{aligned}$$

where $\{i_1,i_2,i_3,i_4,j_1,j_2,j_3,j_4\} \in A^1_{\mathrm{com}}$. The number of such cases is at most $O(\sigma_m^6)$. Therefore

$$\frac{1}{\binom{\sigma_m}{4}^2}\sum_{A^1_{\mathrm{com}}} \mathrm{Cov}(h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}, h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}) \;=\; O\left(\frac{1}{\sigma_m M^2}+\frac{1}{\sigma_m^2}\right).$$

Lastly we focus on the case when $\{i_1,i_2,i_3,i_4,j_1,j_2,j_3,j_4\} \in A^0_{\mathrm{com}}$. First of all, when there is no collision, two kernels become independent and thus $\mathrm{Cov}(h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}, h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}) = 0$. When there is one collision, we are basically in the same situation where $\{i_1,i_2,i_3,i_4,j_1,j_2,j_3,j_4\} \in A^1_{\mathrm{com}}$ and there is no collision. Therefore we can re-use the previous bound for the covariance (65). Furthermore, since there are at most $O(\sigma_m^7)$ cases of one collision, $O(\sigma_m^6)$ of two collisions and so on, we have

$$\frac{1}{\binom{\sigma_m}{4}^2}\sum_{A^0_{\mathrm{com}}} \mathrm{Cov}(h^{m,\boldsymbol{\pi}}_{i_1,i_2,i_3,i_4}, h^{m,\boldsymbol{\pi}}_{j_1,j_2,j_3,j_4}) \;=\; O\left(\frac{1}{\sigma_m M^2}+\frac{1}{\sigma_m^2}\right).$$

Therefore, combining the results yields

$$\mathrm{Var}[U(\boldsymbol{W}^{\boldsymbol{\pi}}_m)|\sigma_m] \;=\; O\left(\frac{1}{\sigma_m M^2}+\frac{1}{\sigma_m^2}\right).$$

Given the above bound, the variance of the combined test statistic is

$$\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}] \;=\; \mathbb{E}\big[\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big] + \mathrm{Var}\big[\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big].$$

Let us upper bound these two terms separately. Starting with the first term, note that

$$\begin{aligned}\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}] \;&=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \times \mathrm{Var}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m] \\ &=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \times O\left(\frac{1}{\sigma_m M^2} + \frac{1}{\sigma_m^2}\right),\end{aligned}$$

which results in

$$\mathbb{E}\big[\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big] = O\left(\frac{n}{M^2} + M\right). \tag{68}$$

For the second term,

$$\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}] \;=\; \sum_{m\in[M]} \mathbb{1}(\sigma_m \geq 4)\sigma_m \times \mathbb{E}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m].$$

Since $\sigma_1, \ldots, \sigma_M$ are independent,

$$\begin{aligned}\mathrm{Var}\big[\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big] \;&=\; \sum_{m\in[M]} \mathrm{Var}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m \times \mathbb{E}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m]\big] \\ &\leq\; \sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2 \times \big\{\mathbb{E}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m]\big\}^2\big].\end{aligned}$$

Recall from the claim (61) that

$$\mathbb{E}[U(\boldsymbol{W}_m^{\boldsymbol{\pi}})|\sigma_m] = O\left(\frac{1}{M^2} + \frac{1}{M\sigma_m} + \frac{1}{\sigma_m^2}\right) \quad \text{for each } m \in [M]. \tag{69}$$

Therefore

$$\begin{aligned}\mathrm{Var}\big[\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big] \;\leq\; &\frac{C_1}{M^4} \sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2\big] + \frac{C_2}{M^2} \sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\big] \\ &+ C_3 \sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m^{-2}\big].\end{aligned}$$

Now using the property that $\sigma_m \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(np_m)$ where $\sum_{m=1}^M p_m = 1$ and $p_m \geq 0$, we can upper bound

$$\sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m^2\big] \leq \sum_{m\in[M]} np_m + n^2 p_m^2 \leq n + n^2 \quad \text{and}$$

$$\sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\big] \leq \sum_{m\in[M]} \mathbb{E}[\sigma_m] \leq n.$$

Furthermore, the inequality (64) implies that

$$\sum_{m\in[M]} \mathbb{E}\big[\mathbb{1}(\sigma_m \geq 4)\sigma_m^{-2}\big] \;\leq\; O\left(\frac{M^2}{n}\right).$$

Putting pieces together yields

$$\mathrm{Var}\big[\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}|\boldsymbol{\sigma}]\big] \;\leq\; O\left(\frac{n^2}{M^4} + \frac{n}{M^2} + \frac{M^2}{n}\right).$$

Combining the above with (68) concludes

$$\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}] \;=\; O\left(\frac{n^2}{M^4} + \frac{n}{M^2} + \frac{M^2}{n} + M\right).$$

Step 4. Completing the proof. Now by taking $M = \lceil n^{2/5} \rceil$, the previous bounds for the mean and the variance yield

$$\mathbb{E}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}] + \sqrt{\mathrm{Var}[T_{\mathrm{CI}}^{\boldsymbol{\pi}}]} \;=\; O\big(n^{1/5}\big).$$

This together with the inequality (60) gives

$$\mathbb{P}(T_{\mathrm{CI}} \leq q_{1-\alpha}) \;\leq\; \mathbb{P}\big(T_{\mathrm{CI}} \lesssim n^{1/5}\big) + \frac{1}{200} + \mathbb{P}(\mathcal{A}^c).$$

Finally, we prove that $\mathbb{P}(\mathcal{A}^c)$ converges to zero uniformly over a class of distributions of $(X, Y, Z) \in [\ell_1] \times [\ell_2] \times [0, 1]$ where the density of Z is lower bounded by some small constant $c_{\mathrm{low}} > 0$. In particular, we recall the event $\mathcal{A}$ in (58) and define

$$\rho_n := \sup_{P_{X,Y,Z} \in \mathcal{P}'_{1,[0,1],\mathrm{TV}}(L)} \mathbb{P}_{P_{X,Y,Z},N}^{N}(\mathcal{A}^c). \tag{70}$$

We then show that $\rho_n \to 0$ under the conditions in Theorem 8.

By letting $\sigma_{m,k}$ be the sample size within the bin $B_{m,k}$ and $\tau_\alpha > \log^{-1}((1-\alpha/2)/(1-\alpha))$, let us define another event $\mathcal{E}$ in which $\min_{(m,k)\in[M]\times[b]} \sigma_{m,k} \geq \tau_\alpha \times (Mb)$. Note that, under Poissonization, each $\sigma_{m,k} \sim \mathrm{Pois}(np_{m,k})$ where $np_{m,k} \geq c_{\mathrm{low}} n/(Mb)$ since the density of Z is bounded below by c_{low}. We further assume that $c_{\mathrm{low}} n/(Mb) \geq 2\tau_\alpha \times (Mb)$. Then, by Lemma 7 in Appendix B.3, the event $\mathcal{E}$ holds with probability at least

$$1 - \sum_{(m,k)\in[M]\times[b]} e^{-\frac{np_{m,k}}{6}} \geq 1 - Mb \times e^{-\frac{c_{\mathrm{low}} n}{6Mb}}.$$

This probability converges to one as $n \to \infty$ as long as M, b are chosen such that $n/(Mb) \to \infty$. Therefore, we can work under the assumption that $\min_{(m,k)\in[M]\times[b]} \sigma_{m,k} \geq \tau_\alpha \times (Mb)$. Under this event $\mathcal{E}$ with $\tau_\alpha(Mb) \geq 2$ (this can be assumed as $Mb \to \infty$), $\mathbb{P}(\mathcal{A}^c|\mathcal{E})$ is equivalent to

$$\begin{aligned}
&\mathbb{P}\left(\frac{\prod_{(m,k)\in[M]\times[b]} \max\{\sigma_{m,k}-1, 1\}}{\prod_{(m,k)\in[M]\times[b]} \max\{\sigma_{m,k}, 1\}} \leq \frac{1-\alpha}{1-\alpha/2} \middle| \mathcal{E} \right) \\
\leq\ & \mathbb{P}\left(\prod_{(m,k)\in[M]\times[b]} (1-\sigma_{m,k}^{-1}) \leq \frac{1-\alpha}{1-\alpha/2} \middle| \mathcal{E} \right) \\
\leq\ & \mathbb{1}\left(\left(1-\tau_\alpha^{-1}(Mb)^{-1}\right)^{Mb} \leq \frac{1-\alpha}{1-\alpha/2} \right).
\end{aligned}$$

Given that $\tau_\alpha > \log((1-\alpha/2)/(1-\alpha))$ and $(1-a/n)^n \to e^{-a}$ for any constant a, the last display converges to zero as $Mb \to \infty$. Finally, as we argued before, we have $\mathbb{P}(\mathcal{E}^c) \to 0$. Therefore

$$\begin{aligned}
\mathbb{P}(\mathcal{A}^c) &= \mathbb{P}(\mathcal{A}^c|\mathcal{E})\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{A}^c|\mathcal{E}^c)\mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{P}(\mathcal{A}^c|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \to 0,
\end{aligned}$$

as desired. Since we set $\alpha = 0.01$, we can take $\tau_\alpha = 200 > \log^{-1}((1-\alpha/2)/(1-\alpha)) \approx 198.5$ and the above conditions are satisfied when $n/(Mb)^2 \geq 400 c_{\text{low}}^{-1}$ and $Mb \to \infty$ as $n \to \infty$. These are the conditions imposed in Theorem 8 and thus we complete the proof of Theorem 8.

B Auxiliary lemmas and additional results

This section collects some auxiliary lemmas used for the main proofs as well as additional results.

B.1 Metrics and inequalities

The next lemma connects the Hellinger distance between binned distributions $Q_{X,Y|\widetilde{Z}=m}$ and $\widetilde{Q}_{X,Y|\widetilde{Z}=m}$, defined in Section 4.2, to that between the underlying conditional distributions. This lemma is the basis for the proof of Theorem 2 and Theorem 3.

Lemma 3 (A bound on the Hellinger distance)**.** *For each $m = 1, \ldots, M$, we let Z_m denote a random variable from the conditional distribution $Z|Z \in B_m$, and we let $Z_m^{'}, Z_m^{''}$ be i.i.d. copies of Z_m. Then under $X \perp\!\!\!\perp Y|Z$, it holds that*

$$\begin{aligned}
&\mathcal{D}_{\text{H}}^2\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \\
\leq\ & 6\mathbb{E}_{Z_m, Z_m^{'}, Z_m^{''}}\big[\mathcal{D}_{\text{H}}^2\big(P_{X|Z=Z_m}, P_{X|Z=Z_m^{'}}\big) \times \mathcal{D}_{\text{H}}^2\big(P_{Y|Z=Z_m}, P_{Y|Z=Z_m^{''}}\big)\big].
\end{aligned}$$

It is interesting to observe that Lemma 3 connects the distance between the joint conditional distributions of the binned data to the distances between the marginal conditional distributions of the unbinned data. In particular, the upper bound consists of the product between two distances,

which allows us to control the type I error more tightly. The constant factor of 6 is a convenient choice from our analysis and it is by no means tight. To prove this result, we use Jensen's inequality for a vector valued function and the proof can be found in Appendix D.1.

The next lemma presents an inclusion property of the generalized Hellinger distance, which is useful for the proof of Theorem 2.

Lemma 4 (Generalized Hellinger distance)**.** *For any $1 \le \gamma_1 \le \gamma_2$, we have*

$$\mathcal{D}_{\gamma_2,\mathrm{H}}^{\gamma_2}(P,Q) \le \mathcal{D}_{\gamma_1,\mathrm{H}}^{\gamma_1}(P,Q). \tag{71}$$

On the other hand, for any $\gamma \ge 1$ and $1 \le \alpha \le 2$, it holds that

$$\mathcal{D}_{\gamma,\mathrm{H}}(P,Q) \le 2^{1-1/\gamma+1/(\gamma\alpha)}\mathcal{D}_{\gamma\alpha,\mathrm{H}}(P,Q). \tag{72}$$

We note that the first inequality (71) was proved in Kamps (1989) and, as far as we are aware, the second inequality (72) is new in the literature. In particular, the inequality (72) generalizes the well-known inequality between the TV distance and the Hellinger distance:

$$\mathcal{D}_{\mathrm{TV}}(P,Q) \le \sqrt{2}\mathcal{D}_{\mathrm{H}}(P,Q),$$

by taking $\gamma = 1$ and $\alpha = 2$. The proof of Lemma 4 can be found in Appendix D.2. As a direct consequence of Lemma 4, we have the following corollary.

Corollary 1 (A bound on the Hellinger distance)**.** *Suppose that $\gamma > 2$ is within the range of $2^k < \gamma \le 2^{k+1}$ for some positive integer k. Then*

$$\mathcal{D}_{\mathrm{H}}(P,Q) \le 2^k\mathcal{D}_{\gamma,\mathrm{H}}(P,Q). \tag{73}$$

Next we state a few inequalities related to Rényi divergence, which will be useful for the proof of Theorem 3. Let $\mathcal{D}_{\mathrm{KL}}(P\|Q) = \int p\log(p/q)d\mu$ and $\mathcal{D}_{\chi^2}(P\|Q) = \int(p-q)^2/qd\mu$ be the KL divergence and the χ^2 divergence of P from Q, respectively. Then we have the identities:

$$\begin{aligned}
\mathcal{D}_{1/2,\text{Rényi}}(P\|Q) &= -2\log\{1-\mathcal{D}_{\mathrm{H}}^2(P,Q)/2\},\\
\mathcal{D}_{1,\text{Rényi}}(P\|Q) &= \mathcal{D}_{\mathrm{KL}}(P\|Q),\\
\mathcal{D}_{2,\text{Rényi}}(P\|Q) &= \log\{1+\mathcal{D}_{\chi^2}(P\|Q)\}.
\end{aligned}$$

Notice that $\mathcal{D}_{\gamma,\text{Rényi}}(P\|Q)$ is nondecreasing in γ and thus, by the inequality $\log t \le t-1$ for all $t > 0$, we have

$$\mathcal{D}_{\mathrm{H}}^2(P,Q) \le \mathcal{D}_{1/2,\text{Rényi}}(P\|Q) \le \mathcal{D}_{\mathrm{KL}}(P\|Q) = R_1(f\|g) \le \mathcal{D}_{2,\text{Rényi}}(P\|Q) \le \mathcal{D}_{\chi^2}(P\|Q). \tag{74}$$

Other notable inequalities are

$$\frac{\gamma_1}{1-\gamma_1}\frac{1-\gamma_2}{\gamma_2}\mathcal{D}_{\gamma_2,\text{Rényi}}(P\|Q) \le \mathcal{D}_{\gamma_1,\text{Rényi}}(P\|Q) \le \mathcal{D}_{\gamma_2,\text{Rényi}}(P\|Q) \quad \text{for } 0<\gamma_1<\gamma_2<1.$$

Based on the above inequalities, we can upper bound the Hellinger distance by Rényi divergence as follows.

Lemma 5 (Rényi divergence). *For any $\gamma \geq 1/2$, we have*

$$\mathcal{D}_{\mathrm{H}}^2(P, Q) \leq \mathcal{D}_{\gamma,\text{Rényi}}(P\|Q).$$

On the other hand, for $0 < \gamma < 1/2$, it holds that

$$\mathcal{D}_{\mathrm{H}}^2(P, Q) \leq \mathcal{D}_{1/2,\text{Rényi}}(P\|Q) \leq \frac{1-\gamma}{\gamma}\mathcal{D}_{\gamma,\text{Rényi}}(P\|Q).$$

The proof of the above lemma can be found in Van Erven and Harremos (2014).

B.2 Connection between permutation p-value and quantile

The next lemma provides an alternative expression of the permutation test in terms of the quantile of the permutation distribution, which is more intuitive to analyze.

Lemma 6 (Quantile). *Recall the permutation p-value p_{perm} in (3). Let $q_{1-\alpha}$ be the $1-\alpha$ quantile of the empirical distribution of $T_{\text{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\text{CI}}^{\boldsymbol{\pi}_K}$. Then it holds that*

$$\mathbb{1}(p_{\text{perm}} > \alpha) = \mathbb{1}(T_{\text{CI}} \leq q_{1-\alpha}).$$

Furthermore, by letting $k = \lceil (1-\alpha)K \rceil$, $q_{1-\alpha}$ is the same as the kth order statistic $T_{\text{CI}}^{\boldsymbol{\pi}_{(k)}}$ of $T_{\text{CI}}^{\boldsymbol{\pi}_1}, \ldots, T_{\text{CI}}^{\boldsymbol{\pi}_K}$.

The proof of this result follows by the definition of the quantile function and the details can be found in Appendix D.3.

B.3 Poissonization

In the following proposition, we demonstrate that the local permutation test under Poissonization has the same validity as before.

Proposition 2 (Validity under Poissonization). *Let $\mathcal{P}_0' \subset \mathcal{P}_0$ be the collection of distributions for which it holds that $\sup_{P_{X,Y,Z} \in \mathcal{P}_0'} \mathbb{E}_{P_{X,Y,Z}^n}[\phi_{\text{perm},n}] \leq \alpha + C_1 n^{-\epsilon}$ where $C_1 > 0$ is a universal constant, $\epsilon > 0$ and $n \geq 1$. We further let $\phi_{\text{perm},n} = 0$ when $n = 0$. Then under Poissonization where $N \sim \text{Pois}(n)$, the local permutation test has the type I error guarantee as*

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_0'} \mathbb{E}_{P_{X,Y,Z},N}^N[\phi_{\text{perm},N}] \leq \alpha + C_2(\epsilon) \cdot n^{-\epsilon} + ne^{-n/12},$$

where $C_2(\epsilon) > 0$ is a constant that only depends on ϵ.

The proof of Proposition 2 uses the fact that a Poisson random variable is tightly concentrated around its mean, and the details can be found in Appendix D.4. Suppose we choose the maximum diameter h such that the upper bounds in Theorem 2 and Theorem 3 converge to α at a certain rate. Then Proposition 2 essentially implies that the same convergence rate of the type I error can be obtained under Poissonization as well.

The following lemma provides a concentration bound for the minimum of Poisson random variables.

Lemma 7 (Poisson concentration). *For $m = 1, \ldots, M$, let X_m be a Poisson random variable with parameter $\lambda_m \geq c_0$. Then*

$$\mathbb{P}\Big(\min_{m \in [M]} X_m \leq c_0/2 \Big) \leq \sum_{m=1}^{M} e^{-\frac{\lambda_m}{6}}.$$

B.4 Unbounded support of Z

In this subsection, we discuss how to deal with unbounded support of Z for type I error control. In particular, we demonstrate that it is possible to obtain similar validity results as in Theorem 2 and Theorem 3, up to a vanishing term in the sample size, even when the support of the conditional variable Z is unbounded.

To fix ideas, suppose Z is a univariate random variable and let $\{B_1, \ldots, B_M\}$ denote a partition of $[a, b]$ for some $a, b \in \mathbb{R}$. We further denote the complement of $\cup_{i=1}^{M} B_i$ by $B_{M+1} = (-\infty, a) \cup (b, \infty)$ so that $\cup_{i=1}^{M+1} B_i = \mathbb{R}$. Under this setting, the permutation test that builds on partitions $\{B_1, \ldots, B_{M+1}\}$ has the same type I error guarantee as those in Theorem 2 and Theorem 3 except that we have an additional term to control: $nq_{\widetilde{Z}}(m+1) \times \mathcal{D}^2_{\mathrm{H}}(Q_{X,Y|\widetilde{Z}=m+1}, \widetilde{Q}_{X,Y|\widetilde{Z}=m+1})$. In particular, since the Hellinger distance is bounded by one, this extra term goes to zero as long as $nq_{\widetilde{Z}}(m+1) \to 0$ as $n \to \infty$. Therefore, the problem boils down to controlling $nq_{\widetilde{Z}}(m+1)$. We discuss two strategies to handle this depending on whether there exists prior knowledge on the distribution of Z.

1. **Prior information on Z.** Suppose that we have prior knowledge on the distribution of Z. To fix ideas, suppose that Z has an exponential tail such that $\mathbb{P}(|Z| > t) \leq Ce^{-t}$ for some $C > 0$ and for all $t \geq 0$. Then by choosing b sufficiently larger than $\log n$ and by letting $a = -b$, it is guaranteed that the extra term vanishes as $nq_{\widetilde{Z}}(m+1) \to 0$.

2. **Sample splitting.** When there is no information on the distribution of Z a priori, we may choose a and b in a data-dependent way. In detail, we split the sample into two sets of size n_1 and n_2, denoted by D_1 and D_2, respectively. Let $Z_{(1)}$ and $Z_{(n_1)}$ be the minimum and the maximum order statistic of the first set D_1 and let $a = Z_{(1)}$ and $b = Z_{(n_1)}$. We then work with the second set D_2 to perform the local permutation test. Provided that Z has a density with respect to the Lebesgue measure, it can be seen that $\mathbb{P}(Z < Z_{(1)}) = \frac{1}{n_1+1}$ and $\mathbb{P}(Z > Z_{(n_1)}) = \frac{1}{n_1+1}$ where Z is independent of D_1. Note that $q_{\widetilde{Z}}(m+1) = \mathbb{P}(Z < Z_{(1)}|D_1) + \mathbb{P}(Z > Z_{(n_1)}|D_1)$ and therefore $\mathbb{E}_{D_1}[q_{\widetilde{Z}}(m+1)] = \frac{2}{n_1+1}$. In view of (6), the type I error is bounded by

$$\begin{aligned}
&\mathbb{E}_{P^n_{X,Y,Z}}[\phi_{\mathrm{perm},n_2}] = \mathbb{E}_{D_1}[\mathbb{E}_{D_2}[\phi_{\mathrm{perm},n_2|D_1}]] \\
\leq\ & \alpha + \mathbb{E}_{D_1}\left[\left\{ 2n_2 \sum_{m=1}^{M+1} q_{\widetilde{Z}}(m) \times \mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \right\}^{1/2} \right] \\
\leq\ & \alpha + \underbrace{\mathbb{E}_{D_1}\left[\left\{ 2n_2 \sum_{m=1}^{M} q_{\widetilde{Z}}(m) \times \mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \right\}^{1/2} \right]}_{\text{(I)}}
\end{aligned}$$

$$+ \underbrace{\mathbb{E}_{D_1}\left[\left\{2n_2 q_{\widetilde{Z}}(M+1) \times \mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y|\widetilde{Z}=M+1}, \widetilde{Q}_{X,Y|\widetilde{Z}=M+1}\big)\right\}^{1/2}\right]}_{\text{(II)}}.$$

The first term (I) can be handled similarly as in the bounded case. For the second term (II), it can be seen that (II) $\leq \sqrt{\frac{4n_2}{n_1+1}}$ by using the fact that the Hellinger distance is bounded by one together with Jensen's inequality. Thus, the second term is asymptotically vanishing as long as $n_2/n_1 \to 0$ as $n \to \infty$.

C Definitions

We recall a Hölder smoothness class from Neykov et al. (2021), which is used to analyze the local permutation test for continuous data.

Definition 5 (Hölder Smoothness). *Let $s > 0$ be a fixed real number, and let $\lfloor s \rfloor$ denote the maximum integer strictly smaller than s. Denote by $\mathcal{H}^{2,s}(L)$, the class of functions $f : [0,1]^2 \mapsto \mathbb{R}$, which posses all partial derivatives up to order $\lfloor s \rfloor$ and for all $x, y, x', y' \in [0,1]$ we have*

$$\sup_{k \leq \lfloor s \rfloor} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x,y) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x',y') \right| \leq L((x-x')^2 + (y-y')^2))^{\frac{s-\lfloor s \rfloor}{2}},$$

and in addition

$$\sup_{k \leq \lfloor s \rfloor} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x,y) \right| \leq L.$$

The next definition describes a cyclic permutation used in Section 6.

Definition 6 (Cyclic permutation). *For a set $\mathcal{X}$ with finite elements $a_0, a_2, \ldots, a_{n-1}$, a cyclic permutation π is a bijective function from $\mathcal{X}$ to $\mathcal{X}$ such that for some integer k, $\pi(a_i) = a_{(i+k) \bmod n}$.*

D Proofs of the auxiliary lemmas

This section collects the proof of the lemmas in Appendix B.

D.1 Proof of Lemma 3

To ease the notation, we suppress $(x, y|m)$ in $q_{XY|\widetilde{Z}}(x, y|m)$ and write $q_{XY|\widetilde{Z}}$. Similarly, we write $q_{X\cdot|\widetilde{Z}}(x|m)$ and $q_{\cdot Y|\widetilde{Z}}(y|m)$ by $q_{X\cdot|\widetilde{Z}}$ and $q_{\cdot Y|\widetilde{Z}}$, respectively. Using this shorthand notation, consider an alternative expression of the Hellinger distance and its upper bounds

$$2\mathcal{D}^2_{\mathrm{H}}\big(Q_{X,Y|\widetilde{Z}=m}, \widetilde{Q}_{X,Y|\widetilde{Z}=m}\big) \;=\; \int\int \big(\sqrt{q_{XY|\widetilde{Z}}} - \sqrt{q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}}\big)^2 d\mu_X d\mu_Y$$

$$
\begin{aligned}
&= \int\int \frac{\left(q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}\right)^2}{\left(\sqrt{q_{XY|\widetilde{Z}}} + \sqrt{q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}}\right)^2} d\mu_X d\mu_Y \\
&\leq \int\int \frac{\left(q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}\right)^2}{q_{XY|\widetilde{Z}} + q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}} d\mu_X d\mu_Y \\
&\leq 3\int\int \frac{\left(q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}\right)^2}{q_{XY|\widetilde{Z}} + 3q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}} d\mu_X d\mu_Y.
\end{aligned}
$$

Under $X \perp\!\!\!\perp Y|Z$, the term inside of the last expression is equivalent to

$$
\frac{\left(q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}\right)^2}{q_{XY|\widetilde{Z}} + 3q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}} = \frac{\left\{\mathbb{E}_{Z_m}[(p_{X|Z_m} - q_{X\cdot|\widetilde{Z}})(p_{Y|Z_m} - q_{\cdot Y|\widetilde{Z}})]\right\}^2}{\mathbb{E}_{Z_m}[(p_{X|Z_m} + q_{X\cdot|\widetilde{Z}})(p_{Y|Z_m} + q_{\cdot Y|\widetilde{Z}})]},
$$

where we denote $p_{X|Z_m} = p_{X|Z}(x|Z_m)$ and $p_{Y|Z_m} = p_{Y|Z}(y|Z_m)$. Next we apply Jensen's inequality several times and obtain that

$$
\begin{aligned}
\frac{\left\{\mathbb{E}_{Z_m}[(p_{X|Z_m} - q_{X\cdot|\widetilde{Z}})(p_{Y|Z_m} - q_{\cdot Y|\widetilde{Z}})]\right\}^2}{\mathbb{E}_{Z_m}[(p_{X|Z_m} + q_{X\cdot|\widetilde{Z}})(p_{Y|Z_m} + q_{\cdot Y|\widetilde{Z}})]} &\overset{\text{(i)}}{\leq} \mathbb{E}_{Z_m}\left[\frac{(p_{X|Z_m} - q_{X\cdot|\widetilde{Z}})^2 (p_{Y|Z_m} - q_{\cdot Y|\widetilde{Z}})^2}{(p_{X|Z_m} + q_{X\cdot|\widetilde{Z}})(p_{Y|Z_m} + q_{\cdot Y|\widetilde{Z}})}\right] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}_{Z_m, Z'_m, Z''_m}\left[\frac{(p_{X|Z_m} - p_{X|Z'_m})^2 (p_{Y|Z_m} - p_{Y|Z''_m})]^2}{(p_{X|Z_m} + p_{X|Z'_m})(p_{Y|Z_m} + p_{Y|Z''_m})}\right]
\end{aligned}
$$

where we detail step (i) and step (ii) at the end of this section. Applying Tonelli's theorem yields

$$
\begin{aligned}
&3\int\int \frac{\left(q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}\right)^2}{q_{XY|\widetilde{Z}} + 3q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}}} d\mu_X d\mu_Y \\
\leq\ & 3\mathbb{E}_{Z_m, Z'_m, Z''_m}\left[\int \frac{(p_{X|Z_m} - p_{X|Z'_m})^2}{p_{X|Z_m} + p_{X|Z'_m}} d\mu_X \times \int \frac{(p_{Y|Z_m} - p_{Y|Z''_m})^2}{p_{Y|Z_m} + p_{Y|Z''_m}} d\mu_Y\right] \\
\leq\ & 12\mathbb{E}_{Z_m, Z'_m, Z''_m}\left[\mathcal{D}^2_{\mathrm{H}}\big(P_{X|Z=Z_m}, P_{X|Z=Z'_m}\big) \times \mathcal{D}^2_{\mathrm{H}}\big(P_{Y|Z=Z_m}, P_{Y|Z=Z''_m}\big)\right],
\end{aligned}
$$

where the last inequality holds since $(x+y) \leq (\sqrt{x}+\sqrt{y})^2 \leq 2(x+y)$ for $x, y \geq 0$. Hence the result follows.

Details of step (i) and step (ii)

Throughout this part, we assume that $q_{X\cdot|\widetilde{Z}} > 0$ and $q_{\cdot Y|\widetilde{Z}} > 0$. Otherwise we have $q_{XY|\widetilde{Z}} - q_{X\cdot|\widetilde{Z}} q_{\cdot Y|\widetilde{Z}} = 0$ and the bound becomes trivial. Based on the identity

$$
\frac{\left\{\mathbb{E}_{Z_m}[(p_{X|Z} - q_{X\cdot|\widetilde{Z}})(p_{Y|Z} - q_{\cdot Y|\widetilde{Z}})]\right\}^2}{\mathbb{E}_{Z_m}[(p_{X|Z} + q_{X\cdot|\widetilde{Z}})(p_{Y|Z} + q_{\cdot Y|\widetilde{Z}})]}
$$

$$= \frac{\{\mathbb{E}_{Z_m}[p_{X|Z_m}p_{Y|Z_m}] - q_{\cdot Y|\widetilde{Z}}\mathbb{E}_{Z_m}[p_{X|Z_m}] - q_{X\cdot|\widetilde{Z}}\mathbb{E}_{Z_m}[p_{Y|Z_m}] + q_{X\cdot|\widetilde{Z}}q_{\cdot Y|\widetilde{Z}}\}^2}{\mathbb{E}_{Z_m}[p_{X|Z_m}p_{Y|Z_m}] + q_{\cdot Y|\widetilde{Z}}\mathbb{E}_{Z_m}[p_{X|Z_m}] + q_{X\cdot|\widetilde{Z}}\mathbb{E}_{Z_m}[p_{Y|Z_m}] + q_{X\cdot|\widetilde{Z}}q_{\cdot Y|\widetilde{Z}}},$$

it is enough to show that the function

$$f(x,y,z) = \frac{(x-ay-bz+ab)^2}{x+ay+bz+ab}$$

is convex for $x,y,z \geq 0$ and $a,b>0$. Then the result follows by Jensen's inequality for a vector valued function. To this end, we prove that the Hessian matrix of $f(x,y,z)$ is positive semidefinite. Some calculations show that

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{8(ay+bz)^2}{(x+bz+ay+ab)^3}, \\
\frac{\partial^2 f}{\partial y^2} &= \frac{8a^2(x+ab)^2}{(x+bz+ay+ab)^3}, \\
\frac{\partial^2 f}{\partial z^2} &= \frac{8b^2(x+ab)^2}{(x+bz+ay+ab)^3}, \\
\frac{\partial^2 f}{\partial x \partial y} &= -\frac{8a(x+ab)(ay+bz)}{(x+bz+ay+ab)^3}, \\
\frac{\partial^2 f}{\partial x \partial z} &= -\frac{8b(x+ab)(ay+bz)}{(x+bz+ay+ab)^3}, \\
\frac{\partial^2 f}{\partial y \partial z} &= -\frac{8ab(x+ab)^2}{(x+bz+ay+ab)^3},
\end{aligned}$$

and thus the Hessian matrix becomes

$$\mathbf{H}_f = \frac{8}{(x+bz+ay+ab)^3}\begin{bmatrix} ay+bz \\ -a(x+ab) \\ -b(x+ab) \end{bmatrix}\begin{bmatrix} ay+bz & -a(x+ab) & -b(x+ab) \end{bmatrix}.$$

This completes the proof of step (i). Step (ii) follows similarly by conditional Jensen's inequality together with the observation that $\frac{(a-x)^2}{b+x}$ is convex for $x \geq 0$ and $a,b \geq 0$.

D.2 Proof of Lemma 4

The first inequality (71) is stated in Corollary 2 of Kamps (1989). Hence we focus on the second inequality (72). We first claim that, for any $1 \leq \alpha \leq \beta$ such that $1/\alpha + 1/\beta = 1$,

$$|x-y| \leq |x^{1/\alpha} - y^{1/\alpha}| \times |x^{1/\beta} + y^{1/\beta}|. \tag{75}$$

Without loss of generality, assume that $x > y \geq 0$ (when $x = y$, there is nothing to prove). Then, by expanding the terms, the claim becomes equivalent to

$$x^{1/\alpha}y^{1/\beta} \geq x^{1/\beta}y^{1/\alpha}.$$

By taking the log on both sides,

$$\left(\frac{1}{\alpha}-\frac{1}{\beta}\right)\log x \geq \left(\frac{1}{\alpha}-\frac{1}{\beta}\right)\log y,$$

which is true since $1 \leq \alpha \leq \beta$ and $x > y \geq 0$. This verifies the inequality (75). Leveraging this inequality, observe that

$$\int \left|p^{1/\gamma}-q^{1/\gamma}\right|^{\gamma} d\mu \leq \int \left|p^{1/(\gamma\alpha)}-q^{1/(\gamma\alpha)}\right|^{\gamma}\left|p^{1/(\gamma\beta)}+q^{1/(\gamma\beta)}\right|^{\gamma} d\mu$$

for any $1 \leq \alpha \leq \beta$ and $1/\alpha + 1/\beta = 1$. By Hölder's inequality, we further have

$$\left[\int \left|p^{1/\gamma}-q^{1/\gamma}\right|^{\gamma} d\mu\right]^{1/\gamma} \leq \left[\int \left|p^{1/(\gamma\alpha)}-q^{1/(\gamma\alpha)}\right|^{\gamma\alpha} d\mu\right]^{1/(\gamma\alpha)} \left[\int \left|p^{1/(\gamma\beta)}+q^{1/(\gamma\beta)}\right|^{\gamma\beta} d\mu\right]^{1/(\gamma\beta)}.$$

Since p, q are density functions, the second term in the product is bounded above by

$$\left[\int \left|p^{1/(\gamma\beta)}+q^{1/(\gamma\beta)}\right|^{\gamma\beta} d\mu\right]^{1/(\gamma\beta)} = \left[2^{\gamma\beta}\int \left|\frac{p^{1/(\gamma\beta)}+q^{1/(\gamma\beta)}}{2}\right|^{\gamma\beta} d\mu\right]^{1/(\gamma\beta)} \leq 2,$$

where the last step follows by Jensen's inequality (note that $\gamma\beta \geq 1$). Therefore

$$\begin{aligned}\mathcal{D}_{\gamma,\mathrm{H}}(P,Q) = \left[\frac{1}{2}\int \left|p^{1/\gamma}-q^{1/\gamma}\right|^{\gamma} d\mu\right]^{1/\gamma} &\leq 2^{1-1/\gamma}\left[\int \left|p^{1/(\gamma\alpha)}-q^{1/(\gamma\alpha)}\right|^{\gamma\alpha} d\mu\right]^{1/(\gamma\alpha)} \\ &= 2^{1-1/\gamma+1/(\gamma\alpha)}\mathcal{D}_{\gamma\alpha,\mathrm{H}}(P,Q).\end{aligned}$$

Note that by the restriction on α and β, the parameter α should be in the range of $1 \leq \alpha \leq 2$. This completes the proof of Lemma 4.

D.3 Proof of Lemma 6

By the definition of the quantile function, $q_{1-\alpha}$ can be written as

$$\begin{aligned}q_{1-\alpha} &= \inf\left\{x \in \mathbb{R} : 1-\alpha \leq \frac{1}{K}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} \leq x\big)\right\} \\ &= \inf\left\{x \in \mathbb{R} : 1-\alpha \leq \frac{1}{K}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} < x\big)\right\} \\ &= T_{\mathrm{CI}}^{\boldsymbol{\pi}_{(k)}},\end{aligned}$$

where we recall $k = \lceil (1-\alpha)K \rceil$. Given this representation and by noting that the permutation distribution is discrete, we see that $T_{\mathrm{CI}} \leq q_{1-\alpha}$ holds if and only if

$$\frac{1}{K}\sum_{\boldsymbol{\pi}_i \in \boldsymbol{\Pi}} \mathbb{1}\big(T_{\mathrm{CI}}^{\boldsymbol{\pi}_i} < T_{\mathrm{CI}}\big) < 1-\alpha,$$

which is equivalent to $p_{\mathrm{perm}} > \alpha$. This completes the proof.

D.4 Proof of Proposition 2

Since we assume that $\phi_{\mathrm{perm},k} = 0$ when $k = 0$, it holds that

$$\sup_{P_{X,Y,Z}\in\mathcal{P}_0'}\left\{\sum_{k=0}^{\infty}\mathbb{P}(N=k)\mathbb{E}_{P_{X,Y,Z}^k}[\phi_{\mathrm{perm},k}]\right\} = \sup_{P_{X,Y,Z}\in\mathcal{P}_0'}\left\{\sum_{k=1}^{\infty}\mathbb{P}(N=k)\mathbb{E}_{P_{X,Y,Z}^k}[\phi_{\mathrm{perm},k}]\right\}$$
$$\leq \alpha + C_1\sum_{k=1}^{\infty}\mathbb{P}(N=k)k^{-\epsilon}.$$

Furthermore, by transforming k to $k = k' + 1$, we can simplify the infinite sum as

$$\sum_{k=1}^{\infty}\mathbb{P}(N=k)k^{-\epsilon} = \sum_{k=1}^{\infty}\frac{n^k e^{-n}}{k!}k^{-\epsilon} = n\sum_{k'=0}^{\infty}\mathbb{P}(N=k')(k'+1)^{-\epsilon-1}$$
$$= n\mathbb{E}[(N+1)^{-\epsilon-1}],$$

where we recall that $N \sim \mathrm{Pois}(n)$. Now decompose and upper bound the expectation by

$$n\mathbb{E}[(N+1)^{-\epsilon-1}] = n\mathbb{E}[(N+1)^{-\epsilon-1}\mathbb{1}\{N \leq n/2\}] + n\mathbb{E}[(N+1)^{-\epsilon-1}\mathbb{1}\{N > n/2\}]$$
$$\leq n\mathbb{P}(N \leq n/2) + n\left[\frac{1}{n/2+1}\right]^{1+\epsilon}$$
$$\leq ne^{-n/12} + 2^{1+\epsilon}n^{-\epsilon},$$

where the last inequality uses a Chernoff bound for a Poisson random variable (e.g. Canonne, 2017). This completes the proof of Proposition 2.

D.5 Proof of Lemma 7

Let X have a Poisson distribution with parameter $\lambda > 0$. Then by using a Poisson tail bound (e.g. Canonne, 2017), we have for $x > 0$,

$$\mathbb{P}(X \leq \lambda - x) \leq e^{-\frac{x^2}{\lambda+x}}.$$

Now, by setting $x = \lambda/2$, the above inequality guarantees that

$$\mathbb{P}(X \leq \lambda/2) \leq e^{-\frac{\lambda}{6}}. \tag{76}$$

By the union bound along with (76), we have

$$\mathbb{P}\Big(\min_{m\in[M]} X_m \leq c/2\Big) \leq \sum_{m=1}^{M}\mathbb{P}(X_m \leq \lambda_m/2) \leq \sum_{m=1}^{M} e^{-\frac{\lambda_m}{6}},$$

which proves the result.