LOCAL PERMUTATION TESTS FOR CONDITIONAL INDEPENDENCE

BY ILMUN KIM1 MATEY NEYKOV2 SIVARAMAN BALAKRISHNAN2
AND LARRY WASSERMAN2

1Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, ilmun@yonsei.ac.kr
2Department of Statistics and Data Science, Carnegie Mellon University, mneykov@stat.cmu.edu; siva@stat.cmu.edu; larry@stat.cmu.edu

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 [16, 33] and causal inference [58, 46, 26]. 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 [63, 14], 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 [54], 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 exchangeability of the data. 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 [20, 18, 52, 43], 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. [13] and Neykov et al. [43] which study minimax-optimal CI tests and the work of Kim et al. [30] 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, [30] 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 41, 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. [20]. In particular, Fukumizu et al. [20] propose a test based on the empirical Hilbert–Schmidt norm of the conditional cross-covariance operator. Zhang et al. [66] introduce another kernel-based test attempting to measure partial correlations, which in turn characterize CI [e.g. 15]. Strobl et al. [59] use random Fourier features to approximate kernel computations, and propose a more computationally efficient version of the test of [66]. Other CI measures proposed by Doran et al. [18] and Huang et al. [25] are motivated by the kernel maximum mean discrepancy for two-sample testing [21]. In particular, the CI measure introduced by [25] compares whether \(Y \mid X, Z\) and \(Y \mid X\) have the same distribution, and their measure can be viewed as a kernelized version of the CI measure of [2]. Recently, Sheng and Sriperumbudur [55] and Part and Muandet [44] propose kernel CI measures that are closely connected to the Hilbert–Schmidt independence criterion [22]. Sheng and Sriperumbudur [55] also discuss a connection between their CI measure to the conditional distance correlation proposed by Wang et al. [62].

Another category of methods relies on estimating regression functions. Consider random variables \(X\) and \(Y\), and their regression residuals on \(Z\), denoted by \(\epsilon_{X,Z} := X - \mathbb{E}[X \mid Z]\) and \(\epsilon_{Y,Z} := Y - \mathbb{E}[Y \mid Z]\). The underlying idea of regression-based methods is that the expected value of \(\epsilon_{X,Z}\epsilon_{Y,Z}\) is zero if \(X \perp \perp Y \mid Z\), and not necessarily zero if \(X \not\perp \not\perp Y \mid Z\). Thus, one can utilize 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. [65] for a discussion of the relationship between \(\epsilon_{X,Z} \perp \perp \epsilon_{Y,Z}\) and \(X \not\perp \not\perp Y \mid Z\). Given that there exist a variety of successful regression algorithms to estimate \(\mathbb{E}[X \mid Z]\) and \(\mathbb{E}[Y \mid Z]\), 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 [54] propose the generalized covariance measure, which has been extended to functional linear models by Lundborg et al. [38]. The methods proposed by Zhang et al. [66] and Strobl et al. [59] 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 [48, 37, 42] 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, [7, 56] design nonparametric tests by leveraging the success of generative adversarial networks. [53, 52] 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 [9, 8, 57, 45, 47]. A metric-based approach is also common in the literature, including tests based on the conditional Hellinger distance [60] and conditional mutual information [49]. The above methods are mainly for continuous data, whereas there are numerous CI tests available for discrete data as well [1, 64, 28, 13, 40, 43, 3]. A more extensive review of CI tests can be found in [36].

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 to 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 [60, 24, 66, 62, 59, 54, 67]. Due to technical hurdles, this line of work often focuses on a pointwise (rather than uniform) type I error guarantee with a few exceptions [54, 38].
• **Model-X framework.** Formulated by Candès et al. [11], 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 [11] or via a permutation method as in [10]. 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. [10] 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 [39, 20, 18, 53] 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 [13], [43] and [30]. Canonne et al. [13] 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. [43] extend the work of [13] to the case where $Z$ is a continuous and bounded random variable. However, both tests considered in [13] and [43] 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 [30] 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 [13] and [43].

1.2. **Our contributions.** We now outline our contributions.

• **Hardness result of CI testing (Section 3).** By leveraging the recent work of [6, 5], 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. 30]. However, this is not the case for the continuous CI testing problem. As pointed out by Shah and Peters [54], 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. [13]. Theorem 5 then shows that the corresponding local permutation tests have the same power guarantee as the tests in [13]. 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. [43] 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 [43].

• **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 both the discrete case and the continuous 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 [13] and [43], 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 in Section 8. All technical proofs are relegated to the supplementary material [31]. In [31], we also provide additional simulation results and a practical guideline of binning scheme.
Step 1. Discretization of $Z$

**Fig 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 $\tilde{Z}$. Given a product distribution $P_{X,Y,Z}^n$, the corresponding distribution smoothed over the bins is denoted by $Q_{X,Y,Z}^n$ (see Section 4.2 for a more precise description). Next, we consider another distribution $\tilde{Q}_{X,Y,Z}^n$, which is the CI projection of $Q_{X,Y,Z}^n$ onto the space where $X \perp Y | \tilde{Z}$. The validity of the location permutation test is essentially determined by the TV distance $\delta_n$ (or its upper bound) between $Q_{X,Y,Z}^n$ and $\tilde{Q}_{X,Y,Z}^n$. 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 $\delta_n$ 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 [43]. 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 $X \perp Y | Z$ 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 $Z$ and briefly discuss an extension to unbounded support in Appendix C.4 of [31]. Let $\{B_1, \ldots, B_M\}$ denote a partition of $Z$ such that $Z = \cup_{m=1}^M B_m$ and $\sigma := \{\sigma_1, \ldots, \sigma_M\}$ denote the sample sizes within bins $\{B_1, \ldots, B_M\}$. Furthermore, let $W_m$ denote the set of the pairs of $(X_i, Y_i)$ that belong to the $m$th bin. More formally, by letting $(X_{i,m}, Y_{i,m})$ be the $i$th pair in the $m$th bin, we write $W_m := \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\}$ when $\sigma_m \geq 1$ and otherwise $W_m = \emptyset$.

Given this binned data, we consider a generic test statistic for CI testing, which maps from $W_1, \ldots, W_M$ to $\mathbb{R}$. That is, for some function $f : W_1, \ldots, W_M \mapsto \mathbb{R}$, we compute our test statistic as:

\[ T_{CI} = f(W_1, \ldots, W_M). \]
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 $W_1, \ldots, W_M$, 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, let us consider a permutation $\pi_m = \{\pi_m(1), \ldots, \pi_m(\sigma_m)\}$ of $\{1, \ldots, \sigma_m\}$ and denote $W_{\pi_m} = \{(X_{1,m}, Y_{\pi_m(1),m}), \ldots, (X_{\sigma_m,m}, Y_{\pi_m(\sigma_m),m})\}$ for $m = 1, \ldots, M$. Notice that when $\sigma_m = 0$, there is nothing to permute and we set $W_{\pi_m} = \emptyset$. The test statistic computed using the locally permuted data set is denoted by

\[
T_{CI}^{\pi} = f(W_{\pi_1}, \ldots, W_{\pi_M}).
\]

Let us further denote the set of all possible such local permutations $\pi = \{\pi_1, \ldots, \pi_M\}$ by $\Pi$ 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 $\alpha$

1. For each $\pi \in \Pi$, compute $T_{CI}^{\pi}$ as in (2) and denote the resulting statistics by $T_{CI}^{\pi_1}, \ldots, T_{CI}^{\pi_K}$.
2. By comparing the statistic $T_{CI}$ in (1) with the permuted ones, calculate the $p$-value as

\[
p_{perm} = \frac{1}{K} \sum_{\pi \in \Pi} I\{T_{CI}^{\pi} \geq T_{CI}\}.
\]

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

The local permutation procedure, like other randomized or permutation procedures [e.g. Chapter 15 of 35], 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 [35], 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 [19] and we refer to Corollary 6.1 of [29] or Proposition 4 of [51] 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 [23] for completeness. For a nominal level $\alpha$, we denote $k = K - \lceil K\alpha \rceil$ where $\lfloor K\alpha \rfloor$ is the largest integer less than or equal to $K\alpha$. In addition let $K^+$ and $K^0$ be the numbers of $T_{CI}^{\pi_1}, \ldots, T_{CI}^{\pi_K}$, which are greater than or equal to $T_{CI}$, respectively. Given $a = (K\alpha - K^+)/K^0$ and the $k$th order statistic $T_{CI}^{(k)}$ of $T_{CI}^{\pi_1}, \ldots, T_{CI}^{\pi_K}$, we define $\phi_{perm,n,a} = 1, a$ or $0$ depending on whether $T_{CI} > T_{CI}^{(k)}$, $T_{CI} = T_{CI}^{(k)}$ or $T_{CI} < T_{CI}^{(k)}$, respectively. Then under the exchangeability assumption of the permuted statistics, it holds that $\mathbb{E}[\phi_{perm,n,a}] = \alpha$, whereas $\phi_{perm,n}$ from Algorithm 1 has a weaker guarantee that $\mathbb{E}[\phi_{perm,n}] \leq \alpha$ 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 \( \mu \) 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. 27], includes the TV distance and the Hellinger distance as special cases.

**DEFINITION 2.1** (Generalized Hellinger distances). Given \( \gamma \geq 1 \), the generalized Hellinger distance with parameter \( \gamma \) between \( P \) and \( Q \) is defined as

\[
D_{\gamma,H}(P,Q) = \left( \frac{1}{2} \int |p^{1/\gamma} - q^{1/\gamma}|^\gamma d\mu \right)^{1/\gamma}.
\]

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 \( D_{TV}(P,Q) \) and \( D_H(P,Q) \), respectively. The generalized Hellinger distance has the monotonicity property that \( D_{\gamma_1,H}(P,Q) \leq D_{\gamma_2,H}(P,Q) \) for \( 1 \leq \gamma_1 \leq \gamma_2 \) [Corollary 3 of 27]. This monotonic relationship generalizes the well-known inequality between the TV and Hellinger distances, namely \( D_H^2(P,Q) \leq D_{TV}(P,Q) \) [e.g. Chapter 4 of 34].

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

**DEFINITION 2.2** (Rényi divergences). For \( \gamma \in (0, \infty) \), Rényi divergence of order \( \gamma \) of \( P \) from \( Q \) is defined as

\[
D_{\gamma,R}(P\|Q) = \begin{cases} 
\frac{1}{\gamma - 1} \log \left\{ \int \left( \frac{p}{q} \right)^\gamma q d\mu \right\}, & \text{if } \gamma \neq 1, \\
\int p \log \left( \frac{p}{q} \right) d\mu, & \text{if } \gamma = 1.
\end{cases}
\]

In the above definition, some notable values of \( \gamma \) 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 \( \chi^2 \) divergence \( (\gamma = 2) \) as stated in Appendix C of [31]. We refer the reader to [61] and [50] 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 [54], further revisited by [43], CI testing is intrinsically difficult in the following sense. Let \( \mathcal{P}_{R,d} \) be the set of all distributions for \( (X,Y,Z) \) on \( \mathbb{R}^{d_x+d_y+d_z} \), and let \( \mathcal{P}_K \subset \mathcal{P}_{R,d} \) be the subset of \( \mathcal{P}_{R,d} \) whose support is defined within a \( L_\infty \) 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 \( X \perp Y \mid Z \) 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 \( P_{X,Y,Z}^n \), the result of [54] states that any valid CI test \( \phi \) for the class of null distributions \( \mathcal{P}_{0,K} \) should satisfy \( \mathbb{E}_{P_{X,Y,Z}^n}[\phi] \leq \alpha \) 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 $\mathcal{P}_0 \subset \mathcal{P}_{\mathbb{Z}^d}$ be the subset of distributions such that $X \perp Y | Z$, and define $\mathcal{P}_1 = \mathcal{P}_{\mathbb{R}^d} \setminus \mathcal{P}_0$. Further, let $\mathcal{P}_{0,\text{disc}} \subset \mathcal{P}_0$ 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 $\rho_{J,P} := \mathbb{P}\{Z_1, \ldots, Z_J \text{ are distinct}\}$, where $Z_1, \ldots, Z_J$ are i.i.d. samples from the marginal distribution of $Z$. Suppose that a test $\phi$ satisfies $\sup_{P_{X,Y,Z} \in \mathcal{P}_{0,\text{disc}}} \mathbb{E}_{P_{X,Y,Z}}[\phi] \leq \alpha$ for $\alpha \in (0, 1)$. Then for any $P_{X,Y,Z} \in \mathcal{P}_1$, the power of $\phi$ is bounded above by

$$\mathbb{E}_{P_{X,Y,Z}}[\phi] \leq \alpha \times \rho_{J,P} + (1 - \rho_{J,P}) + \frac{n(n - 1)}{J}.$$ (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 $\alpha$. We make this intuition more precise in Remark 1 of the supplementary material [31] 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 $\phi$ 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 $\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,\text{no-atom}}} \mathbb{E}_{P_{X,Y,Z}}[\phi] \leq \alpha$ for any $\phi$ such that $\sup_{P_{X,Y,Z} \in \mathcal{P}_0} \mathbb{E}_{P_{X,Y,Z}}[\phi] \leq \alpha$. 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 [54] 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 [54].

- On the other hand, the proof of Theorem 1 is much simpler than that of the hardness result of [54], being highly motivated by the recent impossibility results in distribution-free conditional predictive inference e.g., Lemma A.1 of [6] and Lemma 1 of [5]. 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 $\phi$ 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 [21] (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 $\alpha$, 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 B.1 of [31].

In summary, the hardness result of [54] 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 $z \in Z$. 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 [43] defined as follows.

**Definition 4.1 ($\gamma$-Hellinger Lipschitzness).** Let $\mathcal{P}_{0,H,\gamma,\delta}(L) \subset \mathcal{P}_0$ be the collection of distributions $P_{x,y,z}$ such that for all $z, z' \in Z$,

$$D_{\gamma,H}(P_{X|Z=z}, P_{X|Z=z'}) \leq L\delta(z, z') \quad \text{and} \quad D_{\gamma,H}(P_{Y|Z=z}, P_{Y|Z=z'}) \leq L\delta(z, z'),$$

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

Another smoothness assumption is made based on Rényi divergence.

**Definition 4.2 ($\gamma$-Rényi Lipschitzness).** Let $\mathcal{P}_{0,R,\gamma,\delta}(L) \subset \mathcal{P}_0$ be the collection of distributions $P_{x,y,z}$ such that for all $z, z' \in Z$,

$$D_{\gamma,R}^{1/2}(P_{X|Z=z}\|P_{X|Z=z'}) \leq L\delta(z, z') \quad \text{and} \quad D_{\gamma,R}^{1/2}(P_{Y|Z=z}\|P_{Y|Z=z'}) \leq L\delta(z, z'),$$

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

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 us denote by $(X^n, Y^n, Z^n)$ the binned version of $(X^n, Y^n, Z^n)$ where $Z \in \{1, \ldots, M\}$ is a discrete random variable with probability $\mathbb{P}(\tilde{Z} = m) = \mathbb{P}(Z \in B_m)$ for $m = 1, \ldots, M$. For simplicity, let us write $q_Z(m) = \mathbb{P}(\tilde{Z} = m)$ and denote the conditional distribution of $Z|\tilde{Z} \in B_m$ by
Then \( (X, Y, \tilde{Z}) \) has its density function \( q_{X,Y,\tilde{Z}}(x, y, m) = q_{XY|\tilde{Z}}(x, y|m)q_{Z}(m) \) where

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

and we denote the corresponding joint distribution by \( Q_{X,Y,\tilde{Z}} \).

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 \( (X^n, Y^n, \tilde{Z}^n) \). Furthermore, since our test function \( \phi_{perm,n} \) is computed only through the binned data, we observe that

\[
(5) \quad \mathbb{E}_{P_{X,Y,Z}}^{\phi_{perm,n}} = \mathbb{E}_{Q_{X,Y,\tilde{Z}}}^{\phi_{perm,n}}.
\]

To proceed further, we consider a product density

\[
q_{X,Y,\tilde{Z}}(x, y, m) = q_{X|\tilde{Z}}(x|m)q_{Y|\tilde{Z}}(y|m)q_{Z}(m),
\]

where \( q_{X|\tilde{Z}}(x|m) \) and \( q_{Y|\tilde{Z}}(y|m) \) are the marginals of \( q_{XY|\tilde{Z}}(x, y|m) \), and denote the corresponding joint distribution by \( \tilde{Q}_{X,Y,\tilde{Z}} \). By construction, \( (X^n, Y^n, \tilde{Z}^n) \) from \( \tilde{Q}_{X,Y,\tilde{Z}} \) satisfies \( X \perp Y|\tilde{Z} \) and therefore exchangeability of \( Y \) within each bin yields

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

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 \( \phi_{perm,n} \) is bounded above by

\[
\mathbb{E}_{P_{X,Y,Z}}^{\phi_{perm,n}} \leq \alpha + \mathcal{D}_{TV}(Q_{X,Y,\tilde{Z}}, \tilde{Q}_{X,Y,\tilde{Z}}).
\]

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 \( \phi_{perm,n,a} \) (Remark 1) is exact under the law of \( \tilde{Q}_{X,Y,\tilde{Z}}^n \), one can establish a stronger result that

\[
\left| \mathbb{E}_{P_{X,Y,Z}}^{\phi_{perm,n,a}} - \alpha \right| \leq \mathcal{D}_{TV}(Q_{X,Y,\tilde{Z}}, \tilde{Q}_{X,Y,\tilde{Z}}),
\]

where \( \tilde{R} \sim \text{Uniform}[0, 1] \) 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 \( Q_{X,Y,\tilde{Z}}^n \) and \( \tilde{Q}_{X,Y,\tilde{Z}}^n \) 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 \( \phi_{perm,n} \) is bounded above by

\[
(6) \quad \mathbb{E}_{P_{X,Y,Z}}^{\phi_{perm,n}} \leq \alpha + \left\{ \frac{2n}{M} \sum_{m=1}^{M} q_{Z}(m) \times \mathcal{D}_{H}^2(Q_{X,Y|\tilde{Z}=m}, \tilde{Q}_{X,Y|\tilde{Z}=m}) \right\}^{1/2},
\]

where the second term on the right-hand side is simply \( \sqrt{2n}\mathcal{D}_{H}(Q_{X,Y,\tilde{Z}}, \tilde{Q}_{X,Y,\tilde{Z}}) \).
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
\[
D_{TV}(Q^n_{X,Y,Z},\tilde{Q}^n_{X,Y,Z}) \to 0 \quad \text{if and only if} \quad nD^2_{H}(Q_{X,Y,Z},\tilde{Q}_{X,Y,Z}) \to 0.
\]
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
\[
h := \max_{1 \leq m \leq M} \sup_{z,z' \in B_m} \delta(z,z'). \tag{7}
\]
We first state the result under \(\gamma\)-Hellinger Lipschitzness in the next theorem and then consider \(\gamma\)-Rényi Lipschitzness in Theorem 3.

**THEOREM 2 (Validity of \(\phi_{\text{perm},n}\) under \(\gamma\)-Hellinger Lipschitzness).** For any \(\alpha \in (0,1)\), the type I error of \(\phi_{\text{perm},n}\) under \(\gamma\)-Hellinger Lipschitzness is bounded above by
\[
\sup_{P_{X,Y,Z} \in P_{0,H,\gamma,s(L)}} \mathbb{E}_{P_{X,Y,Z}}[\phi_{\text{perm},n}] \leq \begin{cases} 
\alpha + C_\gamma n^{1/2}L^{\gamma}h^{\gamma}, & \text{if } \gamma \in [1,2], \\
\alpha + C_\gamma n^{1/2}L^2 h^2, & \text{if } \gamma > 2,
\end{cases}
\tag{8}
\]
where \(C_\gamma\) is a constant that only depends on \(\gamma\).

A few remarks are in order.

**REMARK 3.**

- The result of Theorem 2 shows that once we assume that \(L\) and \(\gamma\) are fixed in the sample size, a sufficient condition for the validity of \(\phi_{\text{perm},n}\) is \(h = o(n^{-1/2\gamma})\) if \(\gamma \in [1,2]\) and \(h = o(n^{-1/4})\) if \(\gamma > 2\) under \(\gamma\)-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 and in Appendix F of [31].

- In most practical applications, the smoothness parameter \(\gamma\) 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 \(\gamma \geq 1\), 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 \(\alpha\).
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-Rényi Lipschitzness is stated in terms of original distributions. The bulk of the effort in proving Theorem 2 lies in connecting the Hellinger distance between $Q_{X,Y|Z} = m$ and $Q_{X,Y|Z} = n$ to the γ-Rényi 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 B.3 of [31].

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

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

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

where $C_{\gamma}$ is a constant that only depends on $\gamma$.

In contrast to Theorem 2, the above result indicates that the smoothness parameter $\gamma$ in Rényi Lipschitzness does not affect the type 1 error of $\phi_{perm,n}$ 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 γ-Lipschitzness parameter whose underlying divergence is lower bounded by $\gamma > 0$ (see Lemma 3 of [31]). In other words, the conditional distribution in $\mathcal{P}_{0,R,\gamma,\delta}(L)$ is at least as smooth as the one in $\mathcal{P}_{0,H,\gamma=2,\delta}(L)$, 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 $\chi^2$ 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. 13, 4, 43]. 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 1 of the supplementary material [31], 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

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

to denote the expectation operator with respect to $P_{X,Y,Z}^{N}$ where $N$ is a random sample size. We similarly write $\mathbb{P}_{P_{X,Y,Z}^{N},N}[\cdot]$ to denote $\mathbb{E}_{P_{X,Y,Z}^{N},N}[\mathbb{1}(\cdot)]$. Suppose now that

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{0}^{L}} \mathbb{E}_{P_{X,Y,Z}}[\phi_{perm,n}] \leq \alpha + C'n^{-\epsilon}
$$

for some constants $C, \epsilon > 0$. Then under the same condition, the permutation test under Poissonization (i.e. $N \sim \text{Pois}(n)$) satisfies

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{0}^{L}} \mathbb{E}_{P_{X,Y,Z}^{N}}[\phi_{perm,n}] \leq \alpha + C'n^{-\epsilon}
$$

where $C_{\epsilon}$ is a constant that only depends on $\epsilon$. See Proposition 1 of [31] for a more rigorous statement.

**Remark 5 (Double Robustness).** A close look at the proof of Theorem 2 (and also Theorem 3) shows that the excess type I error depends on the product of two Hellinger distances: (i) one between $P_{X|Z=z}$ and $P_{X|Z=z'}$ and (ii) the other between $P_{Y|Z=z}$ and $P_{Y|Z=z'}$. This error bound reveals that the local permutation test can still be valid even if only one of the conditional distributions is sufficiently smooth. Just to build an intuition for this double robustness property, suppose that $Y$ and $Z$ are independent so that $P_{Y|Z=z}$ is a constant function of $z$. In this case, the null of conditional independence $X \perp Y|Z$ becomes equivalent...
to the null of unconditional independence $X \perp Y$ under which the local permutation test is valid regardless of the smoothness condition for $P_{X|Z=z}$. Our proofs of the validity results formalize this intuition and demonstrate the double robustness of location permutation tests in their type I error control.

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 6.

Suppose that $X$ and $Y$ are discrete random variables supported on $[\ell_1] \times [\ell_2]$ for some positive integers $\ell_1$ and $\ell_2$. 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 $W_1, \ldots, W_M$ defined in Section 2.2. To study lower bounds, we work with the weighted sum of U-statistics proposed by [13] and [43]. Let

$$\psi_{ij}^m(x,y) = \mathbb{1}(X_i,m = x, Y_i,m = y) - \mathbb{1}(X_i,m = x) \mathbb{1}(Y_i,m = y),$$

and define a kernel function as

$$h_{i_1,i_2,i_3,i_4}^m = \frac{1}{4!} \sum_{\pi \in \Pi_4} \sum_{x \in \ell_1, y \in \ell_2} \psi_{\pi(1)\pi(2)}^m(x,y) \psi_{\pi(3)\pi(4)}^m(x,y),$$

where $\Pi_4$ is the set of all permutations of $\{i_1, i_2, i_3, i_4\}$. By linearity of expectations, it is seen that $h_{i_1,i_2,i_3,i_4}^m$ is an unbiased estimator of the squared $L_2$ norm between $Q_{X,Y|Z=m}$ and $Q_{X,Y|Z=m}$. Given this kernel and by recalling $W_m = \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\}$, the resulting U-statistic is calculated as

$$U(W_m) := \frac{1}{\binom{\sigma_m}{4}} \sum_{i_1 < i_2 < i_3 < i_4, (i_1,i_2,i_3,i_4) \in [\sigma_m]} h_{i_1,i_2,i_3,i_4}^m.$$ 

The final statistic is a weighted sum of $U(W_1), \ldots, U(W_M)$ given by

$$T_{\text{CI}} := \sum_{m \in [M]} 1(\sigma_m \geq 4) \sigma_m U(W_m).$$

Several properties of $T_{\text{CI}}$, such as minimax power optimality, have been studied under Poissonization by [13] and [43]. To fully benefit from their results, we work with a modified local permutation test: First draw $N \sim \text{Pois}(n/2)$ 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

$$\phi_{\text{perm}}^N, := \phi_{\text{perm},N} \times \mathbb{1}(N \leq n),$$

where $\phi_{\text{perm},N}$ denotes the local permutation test using $T_{\text{CI}}$ (11) computed based on the $N$ samples chosen before. By Proposition 1 of [31] along with the inequality $\phi_{\text{perm}}^N \leq \phi_{\text{perm},N}$, it is clear that $\phi_{\text{perm}}^N$ is asymptotically valid whenever the upper bounds in (8) and (9) converge to $\alpha$. The next theorem provides complementary results establishing lower bounds for the type I error of $\phi_{\text{perm}}^N$. 


Lipschitzness. Consider $\phi$ness. On the other hand, choose $h_n \to 0$ such that $\sqrt{n} h_n \to \infty$ for $\gamma \in [1, 2]$ and $\sqrt{n} h_n^2 \to \infty$ for $\gamma > 2$ under $\gamma$-Hellinger Lipschitzness. On the other hand, choose $h_n \to 0$ such that $\sqrt{n} h_n^2 \to \infty$ for $\gamma > 0$ under $\gamma$-Rényi Lipschitzness. Consider $\phi_{\text{perm}, n}$ based on $T_{\text{CI}}$ in (11). Then there exist constants $n_0, L > 0$, where $L$ depends only on $\gamma$, such that for all $n \geq n_0$, the following two inequalities hold:

$$
\sup_{P_{X, Y, Z} \in \mathcal{P}_{0, n, \gamma, \delta}(L)} \mathbb{E}_{P_{X, Y, Z}, N} [\phi_{\text{perm}, n}^\dagger] \geq 1 - \beta \quad \text{and} \quad
\sup_{P_{X, Y, Z} \in \mathcal{P}_{0, n, \gamma, \delta}(L)} \mathbb{E}_{P_{X, Y, Z}, N} [\phi_{\text{perm}, n}^\dagger] \geq 1 - \beta.
$$

Let us provide several comments on Theorem 4.

**Remark 6.**

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

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

This means that, in order to verify that the type I error of $\phi_{\text{perm}, n}^\dagger$ is inflated, it suffices to show that $\phi_{\text{perm}, n}^\dagger$ is asymptotically powerful against the above alternative (13). With this observation in place, our main task is to construct a distributional setting where $\phi_{\text{perm}, n}^\dagger$ 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 $\phi_{\text{perm}, n}^\dagger$ 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 $T_{\text{CI}}$ approximates $\sigma_1 U(W_1)$, which is much easier to handle. It is then sufficient to study the permutation test based on $\sigma_1 U(W_1)$ 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 [30] where the authors investigate the permutation test based on $U(W_1)$ for unconditional independence testing.

- We expect that $\phi_{\text{perm}, n}$ (i.e. without Poissonization) also achieves the same error bounds in Theorem 4 as it always uses more samples than $\phi_{\text{perm}, n}^\dagger$ by definition. See empirical evidence in Figure 4. However, we found it challenging to analyze $T_{\text{CI}}$ without Poissonization, especially its variance, because of a non-trivial dependence between the summands of $T_{\text{CI}}$. Due to this technical difficulty, we focus on the Poisson-sampling scheme as in [13] and [43] and leave the detailed analysis of $\phi_{\text{perm}, n}$ to future work. Nevertheless, the concentration property of a Poisson random variable allows us to say a certain negative result on $\phi_{\text{perm}, n}$ 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. 12]. Thus it is guaranteed that one can find a fixed sample size $\bar{n}$ such that $cn \leq \bar{n} \leq Cn$ and Theorem 4 holds for $\phi_{\text{perm}, \bar{n}}$ 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 \( \varepsilon \) far away from the null in terms of the TV distance. Our main interest is then to characterize \( \varepsilon \) 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 [13] and demonstrate the power property of the local permutation procedure based on the same test statistics. [13] propose two test statistics \([13] \) proposed another statistic building on the flattening idea of [17]. The latter statistic can be viewed as the sum of weighted U-statistics as observed by [43].

Weighted U-statistic. To proceed, let us formally write down the weighted U-statistic. First recall that \( W_m = \{(X_{1,m}, Y_{1,m}), \ldots, (X_{\sigma_m,m}, Y_{\sigma_m,m})\} \) is the set of pairs of \((X_i, Y_i)\) with \( Z_i = m \). Suppose that the sample size of \( W_m \) is \( \sigma_m \geq 4 \) and \( \sigma_m = 4 + 4t_m \) for some \( t_m \in \mathbb{N} \). Following the notation in [43], let \( t_{1,m} := \min\{t_m, \ell_1\} \) and \( t_{2,m} := \min\{t_m, \ell_2\} \). We then randomly split the data \( W_m \) into three sets \( W_{X,m}, W_{Y,m}, \) and \( W_{XY,m} \) of size \( t_{1,m}, t_{2,m} \) and \( 2t_m + 4 \), respectively, where \( W_{X,m} := \{X_{i,m} : i \in [t_{1,m}]\} \), \( W_{Y,m} := \{Y_{i,m} : t_{1,m} + 1 \leq i \leq t_{1,m} + t_{2,m}\} \) and \( W_{XY,m} := \{(X_{i,m}, Y_{i,m} : 2t_m + 1 \leq i \leq \sigma_m\} \). 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 \( W_{X,m} \) and, similarly, \( a_{y,m} \) is the number of occurrences \( y \) in \( W_{Y,m} \). Next let \( h_{i_1,i_2,i_3,i_4}^{m,a} \) denote a weighted kernel function defined as

\[
    h_{i_1,i_2,i_3,i_4}^{m,a} := \frac{1}{4!} \sum_{\pi \in \Pi_4} \sum_{x \in [\ell_1], y \in [\ell_2]} \frac{\psi_{\pi(1)}(x,y)\psi_{\pi(2)}^m(x,y)}{1 + a_{xy,m}}.
\]

Given this kernel, we compute the weighted U-statistic for each \( 1 \leq m \leq M \) as

\[
    U_W(W_m) := \frac{1}{\binom{2t_m+4}{2}} \sum_{1 \leq i_1 < i_2 < i_3 < i_4 \leq \min(\sigma_m, \ell_1) \min(\sigma_m, \ell_2)} h_{i_1,i_2,i_3,i_4}^{m,a},
\]

where \((i_1, i_2, i_3, i_4) \in W_{XY,m}\) stands for taking four observations from \( W_{XY,m} \). 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 \( U_W(W_1), \ldots, U_W(W_M) \) given by

\[
    T_{CI,W} := \sum_{m \in [M]} \mathbb{I}(\sigma_m \geq 4) \sigma_m \omega_m U_W(W_m).
\]
Tests of Canonne et al. [13]. For both test statistics $T_{CI}$ in (11) and $T_{CI,W}$ in (14), Canonne et al. [13] suggest that one rejects the null when the test statistic is larger than $\zeta \sqrt{\min(n,M)}$ where $\zeta$ 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 $\phi_{CDKS,1} := 1(T_{CI} \geq \zeta \sqrt{\min(n,M)})$ denote the test based on the unweighted statistic and similarly let $\phi_{CDKS,2} := 1(T_{CI,W} \geq \zeta \sqrt{\min(n,M)})$ denote the test based on the weighted test statistic computed based on $N \sim \text{Pois}(n)$ 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 $X \perp Y|Z$ and $\mathcal{P}_{1,[M]} := \mathcal{P}_{[M]} \setminus \mathcal{P}_{0,[M]}$. Canonne et al. [13] first consider the regime where $\ell_1$ and $\ell_2$ are fixed and $\zeta$ satisfies a certain condition recalled in Equation (33) of [31], and then show that

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[M]}: \inf_Q \mathbb{D}_{TV}(P_{X,Y,Z,Q}) \geq \varepsilon} \mathbb{E}_{P^{\text{CI}}_{X,Y,Z,N}}[1 - \phi_{CDKS,1}] \leq \frac{1}{100}.
$$

For example, the condition for $\varepsilon$ is fulfilled if $\varepsilon \geq c \max\{M^{1/4}/n^{1/2}, M^{7/8}/n, M^{3/4}/n^{7/8}\}$ for some large constant $c > 0$. In the second regime where $\ell_1$ and $\ell_2$ can vary, the authors consider a more involved condition for $\varepsilon$ depending on $(\ell_1, \ell_2, M, n)$, and show that

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[M]}: \inf_Q \mathbb{D}_{TV}(P_{X,Y,Z,Q}) \geq \varepsilon} \mathbb{E}_{P^{\text{CI}}_{X,Y,Z,N}}[1 - \phi_{CDKS,2}] \leq \frac{1}{100}.
$$

The condition for $\varepsilon$ in this second regime is recalled in Equation (28) of [31] for completeness.

Main results for the discrete case. In contrast to [13], 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 [13]. 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 $T_{CI,W}$.

REMARK 7 (Full- versus. half-permutation). For the weighted test statistic $T_{CI,W}$, 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 $W_m$, 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 $W_{XY,m}$, 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 $W_{X,m}$, $W_{Y,m}$ and $W_{XY,m}$. On the other hand, the half-permutation approach preserves the independence between $W_{XY,m}$ and $\{W_{X,m}, W_{Y,m}\}$ 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 [30, 32] to analyze two-sample and (unconditional) independence tests.

We are now ready to state the main results of this subsection. As in [13], suppose that we draw $N$ i.i.d. samples from $P_{X,Y,Z} \in \mathcal{P}_{[M]}$ where $N \sim \text{Pois}(n)$. Given these samples, let $\phi_{\text{perm},1}$ be the local permutation test based on the unweighted test statistic $T_{CI}$ (11) through the full-permutation procedure described in Remark 7. Similarly, we let $\phi_{\text{perm},2}$ be the local

### LOCAL PERMUTATION TESTS FOR CONDITIONAL INDEPENDENCE

17

- [13] Canonne et al. [13] first consider the regime where the full-permutation procedure described in Remark 7. Similarly, we let $\phi_{\text{perm},1}$ be the local permutation test based on the unweighted test statistic $T_{CI}$ (11) through the full-permutation procedure described in Remark 7. Similarly, we let $\phi_{\text{perm},2}$ be the local permutation test based on the weighted test statistic computed based on $N \sim \text{Pois}(n)$ 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 $X \perp Y|Z$ and $\mathcal{P}_{1,[M]} := \mathcal{P}_{[M]} \setminus \mathcal{P}_{0,[M]}$. Canonne et al. [13] first consider the regime where $\ell_1$ and $\ell_2$ are fixed and $\zeta$ satisfies a certain condition recalled in Equation (33) of [31], and then show that

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[M]}: \inf_Q \mathbb{D}_{TV}(P_{X,Y,Z,Q}) \geq \varepsilon} \mathbb{E}_{P^{\text{CI}}_{X,Y,Z,N}}[1 - \phi_{CDKS,1}] \leq \frac{1}{100}.
$$

For example, the condition for $\varepsilon$ is fulfilled if $\varepsilon \geq c \max\{M^{1/4}/n^{1/2}, M^{7/8}/n, M^{3/4}/n^{7/8}\}$ for some large constant $c > 0$. In the second regime where $\ell_1$ and $\ell_2$ can vary, the authors consider a more involved condition for $\varepsilon$ depending on $(\ell_1, \ell_2, M, n)$, and show that

$$
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[M]}: \inf_Q \mathbb{D}_{TV}(P_{X,Y,Z,Q}) \geq \varepsilon} \mathbb{E}_{P^{\text{CI}}_{X,Y,Z,N}}[1 - \phi_{CDKS,2}] \leq \frac{1}{100}.
$$

The condition for $\varepsilon$ in this second regime is recalled in Equation (28) of [31] for completeness.

- [30] Main results for the discrete case. In contrast to [13], 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 [13]. 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 $T_{CI,W}$.

### REMARK 7 (Full- versus. half-permutation). For the weighted test statistic $T_{CI,W}$, 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 $W_m$, 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 $W_{XY,m}$, 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 $W_{X,m}$, $W_{Y,m}$ and $W_{XY,m}$. On the other hand, the half-permutation approach preserves the independence between $W_{XY,m}$ and $\{W_{X,m}, W_{Y,m}\}$ 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 [30, 32] to analyze two-sample and (unconditional) independence tests.
permutation test base on the weighted test statistic \(T_{CI,W}\) (14) through the half-permutation procedure described in Remark 7. 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 \(\phi_{perm,1}\) and \(\phi_{perm,2}\), described above. In the setting of discrete \((X, Y, Z)\), \(\phi_{perm,1}\) and \(\phi_{perm,2}\) have the same type II error guarantee as in (15) and (16), respectively.

The implications of Theorem 5 are as follows.

**Remark 8.**

- 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 but “fixed” number in the open interval \((0, 1)\). This change will only affect the constant factor in the condition for separation \(\varepsilon\) given in Equation (29) of [31].

- To make the given tests feasible for a fixed sample size, one can apply the truncation trick as in (12) and consider \(\phi_{perm,1}^\dagger = \phi_{perm,1} \times I(N \leq n)\) and \(\phi_{perm,2}^\dagger = \phi_{perm,2} \times I(N \leq n)\) where \(N \sim \text{Pois}(n/2)\). As discussed before, these modified tests have smaller type I errors than \(\phi_{perm,1}\) and \(\phi_{perm,2}\) based on \(N \sim \text{Pois}(n/2)\), respectively, and have the same power guarantee up to \(e^{-n/8}\) factor. See equation (3) of [31] for more details.

- Canonne et al. [13] further prove that the condition (29) of [31] 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 [13] whenever the tests of [13] 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 [13] 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 [30]. 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 B.6 of [31].

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 [43] and investigate the power of local permutation tests for continuous data. The idea of [43] is to carefully discretize \(Z\) into several bins and apply the tests based on \(T_{CI}\) and \(T_{CI,W}\) as if the original data were discrete. Neykov et al. [43] 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 [43].

**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 $X \perp Y | Z$ and $\mathcal{P}_{1,[0,1]} := \mathcal{P}_{[0,1]} \setminus \mathcal{P}_{0,[0,1]}$. Furthermore, let $\mathcal{P}_{1,[0,1]}$, $TV(L) \subset \mathcal{P}_{1,[0,1]}$ be the subset of alternative distributions that satisfies the TV smoothness:

\begin{equation}
D_{TV}(P_{X,Y|Z=z},P_{X,Y|Z=z'}) \leq L |z - z'| \quad \text{for all } z, z' \in [0, 1].
\end{equation}

To describe the result of [43], draw $N \sim \text{Pois}(n/2)$. If $N \leq n$, 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 $N \leq n$, compute the unweighted test statistic $T_{CI}$ (11) using the binned data set $W_1, \ldots, W_M$ where $M = \lceil n^{2/5} \rceil$ and $\sum_{i=1}^M \sigma_i = N$. For a sufficiently large constant $\zeta$ depending on $L$, the corresponding test of [43] is defined as

$$\phi_{NBW,1} := 1(T_{CI} \geq \zeta n^{1/5}) \times 1(N \leq n).$$

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 $\varepsilon \geq cn^{-2/5}$,

\begin{equation}
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1]}, TV(L)} \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathbb{E}_{P_{X,Y,Z,N}}[1 - \phi_{NBW,1}] \leq \frac{1}{100} + e^{-n/8}.
\end{equation}

The type I error of $\phi_{NBW,1}$ 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 $\ell_1$ and $\ell_2$ potentially increase with $n$. To avoid this issue, Neykov et al. [43] follow the idea of [13] and propose another test based on the weighted test statistic $T_{CI,W}$ (14) where $M = \lceil \frac{n^{2/5}}{(\ell_1 \ell_2)^{1/5}} \rceil$. More formally, for a sufficiently large $\zeta$ depending on $L$, the second test is defined as

\begin{equation}
\phi_{NBW,2} := 1(T_{CI,W} \geq \sqrt{\zeta M}) \times 1(N \leq n).
\end{equation}

Again, let $c$ be a sufficiently large constant depending on $(\zeta, L)$ such that $\varepsilon \geq c\frac{(\ell_1 \ell_2)^{1/5}}{n^{2/5}}$. Under this condition for $\varepsilon$ and an extra condition that $M \ell_1 \lesssim n$ for $\ell_1 \geq \ell_2$, Theorem 5.5 of [43] guarantees that

\begin{equation}
\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,[0,1]}, TV(L)} \inf_{Q \in \mathcal{P}_{0,[0,1]}} \mathbb{E}_{P_{X,Y,Z,N}}[1 - \phi_{NBW,2}] \leq \frac{1}{100} + e^{-n/8}.
\end{equation}

Furthermore, the authors prove that no CI test can be uniformly powerful in the TV distance when $\varepsilon$ is much less than $\frac{(\ell_1 \ell_2)^{1/5}}{n^{2/5}}$. That means, $\phi_{NBW,1}$ and $\phi_{NBW,2}$ achieve minimax optimal rate, while the optimality of $\phi_{NBW,1}$ is only guaranteed when $\ell_1$ and $\ell_2$ are bounded. However the optimal power of $\phi_{NBW,2}$ over a broader regime comes at the cost of decreasing the size of null distributions. In fact, the type I error of $\phi_{NBW,2}$ is guaranteed over a $\chi^2$-smooth class of null distributions, which is smaller than the class of null distributions considered for $\phi_{NBW,1}$. See [43] for more details.

Having described the results of [43], 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 size $N \leq n$ where $N \sim \text{Pois}(n/2)$, let us denote the local permutation tests based on $T_{CI}$ and $T_{CI,W}$ by $\phi_{\text{perm},1}$ and $\phi_{\text{perm},2}$, respectively. As described in Remark 7, it greatly simplifies the power analysis for $T_{CI,W}$ when we consider the half-permutation method. Hence, unlike $\phi_{\text{perm},1}$ that builds on Algorithm 1, the $p$-value of $\phi_{\text{perm},2}$ 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

$$\phi_{\text{perm},1}^\dagger := \phi_{\text{perm},1} \times 1(N \leq n) \quad \text{and} \quad \phi_{\text{perm},2}^\dagger := \phi_{\text{perm},2} \times 1(N \leq n).$$

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 $\phi_{\text{perm},1}^\dagger$ and $\phi_{\text{perm},2}^\dagger$ described above. In the setting of discrete $(X, Y)$ and continuous $Z$, $\phi_{\text{perm},1}^\dagger$ and $\phi_{\text{perm},2}^\dagger$ have the same type II error guarantee as in (18) and (20), respectively.

**Remark 9.**

- We highlight once again that the local permutation tests $\phi_{\text{perm},1}^\dagger$ and $\phi_{\text{perm},2}^\dagger$ do not require the knowledge on unspecified constant $\zeta$ in $\phi_{\text{NBW},1}$ and $\phi_{\text{NBW},2}$, 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 $\mathcal{P}_{0,H;\gamma,\delta}(L)$ in Definition 4.1, 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 $\phi_{\text{perm},1}^\dagger$ is valid when $\gamma > 5/4$. On the other hand, $\phi_{\text{NBW},1}$ can control the type I error even when $\gamma = 1$ [see Theorem 5.2 of 43]. Similarly, in order for $\phi_{\text{perm},2}^\dagger$ to be valid under the same $\chi^2$-smoothness condition in Theorem 5.5 of [43], we require that $n^{1/2}h^{2} = (\xi_{L})^{2/5} / n^{3/10} \to 0$, which was not needed for $\phi_{\text{NBW},2}$. 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 $\phi_{\text{perm},1}^\dagger$ is upper bounded by that of $\phi_{\text{NBW},1}$, 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 $\phi_{\text{perm},1}^\dagger$ is small. The proof for $\phi_{\text{perm},2}^\dagger$ follows similarly. The details can be found in Appendix B.7 of [31].

**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 $X \perp Y \mid Z$. Let $\mathcal{P}_{1,[0,1]^3,TV}$ be the subset of $\mathcal{P}_{[0,1]^3} := \mathcal{P}_{[0,1]^3} \setminus \mathcal{P}_{0,[0,1]^3}$, which satisfies the TV smoothness condition in (17). In addition, we assume that, for any $P_{X,Y,Z} \in \mathcal{P}_{1,[0,1]^3,TV}$, 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 $[0, 1]^2 \to \mathbb{R}$ with parameter $s$ defined in Definition D.1 of the supplementary material [31].

In order to apply $T_{CI,W}$ to continuous data, we need to further discretize $X, Y$ into several bins. For this purpose, for a given $\hat{M} > 0$ (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 $g : [0, 1] \to \{1, \ldots, M'\}$ 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 $\phi_{\text{NBW},2}$ in (19), but with a different choice of $M = \lceil n^{2s/(5s+2)} \rceil$. According to Theorem 5.6 of [43], the resulting test achieves minimax optimal power. In particular, they show that there exists a sufficiently large constant
c depending on $\zeta$, $L$, and if $\varepsilon \geq cn^{-2s/(5s+2)}$, then

\begin{equation}
\sup_{P_{X,Y,Z} \in P_{\{0,1\}^3}} \mathbb{E}_{P_{X,Y,Z}^n} \left[ \inf_{Q \in P_{\{0,1\}^3}} D_{TV}(P_{X,Y,Z}, Q) \right] \geq \varepsilon \implies \mathbb{E}_{P_{X,Y,Z}^n} \left[ 1 - \phi_{\text{NBW},2} \right] \leq \frac{1}{100} + e^{-n/8}.
\end{equation}

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

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

The same points in Remark 9 apply to Theorem 7. While $\phi_{\text{perm},2}^{\dagger}$ has the same optimality as $\phi_{\text{NBW},2}$ in terms of power, we need to restrict the class of null distributions further to rigorously control the type I error. In particular, under $\gamma$-Hellinger Lipschitzness with $\gamma \geq 2$, the underlying conditional density function should be smooth enough to ensure that $n^{1/h^2} = n^{-4s/(5s+2)+1/2} \to 0$, 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 [43]. 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 $Q_{X,Y,Z}^n$ and its CI projection $\tilde{Q}_{X,Y,Z}^n$ 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 $Z$. 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 $W_{m,k}$ 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 $i$th pair in $B_{m,k}$, we write $W_{m,k} := \{(X_{1,m,k}, Y_{1,m,k}), \ldots, (X_{\sigma_{m,k},m,k}, Y_{\sigma_{m,k},m,k})\}$ when $\sigma_{m,k} \geq 1$ and otherwise $W_{m,k} = \emptyset$. 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 $T_{CI}^\pi$ 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 D.2 of [31]. 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 \(\Pi_{\text{cyclic}}\). 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 \(Z\): \(\{B_1, \ldots, B_M\}\), sub-partitions of each \(B_m : \{B_{m,1}, \ldots, B_{m,b}\}\) for \(m = 1, \ldots, M\), a test statistic \(T_{\text{C1}}\), a nominal level \(\alpha\)

1. For each \(\pi \in \Pi_{\text{cyclic}}\), compute \(T_{\pi}^{\text{CI}}\) as in (2) and denote the resulting statistics by \(T_{\pi}^{\text{CI}}\).
2. By comparing the statistic \(T_{\text{C1}}\) in (1) with the permuted ones, calculate the \(p\)-value as

\[ p_{\text{perm}} = \frac{1}{K_*} \sum_{\pi \in \Pi_{\text{cyclic}}} 1 \{T_{\pi}^{\text{CI}} \geq T_{\text{C1}}\}. \]

3. Given the nominal level \(\alpha \in (0, 1)\), define the test function \(\phi_{\text{perm}, n} = \mathbb{I}(p_{\text{perm}} \leq \alpha)\) 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, type I error control of this new approach depends only on the sub-bins and all of the results in Section 4.2 continue to hold with the maximum diameter (7) defined over the sub-bins. Since the maximum diameter of these finer bins is smaller than that of the coarser bins, 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 \(\Pi_{\text{cyclic}}\), rather than all possible local permutations. However, this change does not affect the validity as stated below.

**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 1 of [31]. 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 1 in [31]. 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 \(\ell_1, \ell_2\) are fixed. In addition, we recall the definition of \(\mathcal{P}_{1,[0,1],TV}(L)\) from Section 5.2 and consider a subset of \(\mathcal{P}_{1,[0,1],TV}(L)\), denoted by \(\mathcal{P}_{1,[0,1],TV}(L)\), 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 \(\mathcal{P}_{1,[0,1],TV}(L)\) satisfies

\[ \mathcal{D}_{TV}(P_X|Z=z, P_X|Z=z') \leq L|z-z'| \quad \text{and} \quad \mathcal{D}_{TV}(P_Y|Z=z, P_Y|Z=z') \leq L|z-z'|, \]

for all \(z, z' \in [0, 1]\).
To describe type II error results, we consider the test statistic $T_{CI}$ used in $\phi_{NBW,1}$. 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 $N \sim \text{Pois}(n/2)$, we compute $\phi_{\text{perm},N}$ based on $T_{CI}$ using Algorithm 2 at significance level $\alpha = 0.01$ (for simplicity) and define $\phi_{\text{perm},n}^\dagger := \phi_{\text{perm},N} \times 1\{N \leq n\}$ 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 $\phi_{\text{perm},n}^\dagger$ defined above with the number of larger bins $M = \lceil n^{2/5} \rceil$. Suppose that we choose $b$ such that $n/(Mb)^2 > 400c_\text{low}^{-1}$ as $n \to \infty$. Suppose further that $\varepsilon \geq cn^{-2/5}$ for a sufficiently large $c$ depending on $(L, \ell_1, \ell_2)$. Then

$$\sup_{P_{X,Y,Z} \in \mathcal{P}_{1,\ell_1,\ell_2}(L): \text{inf}_{Q \in \mathcal{P}_0,\ell_1,\ell_2} D_{\text{TV}}(P_{X,Y,Z},Q) \geq \varepsilon} \mathbb{E}_{P_{X,Y,Z} \in \mathcal{P}^\dagger_{1,\ell_1,\ell_2}(L)}[1 - \phi_{\text{perm},n}^\dagger] \leq \frac{1}{100} + \varepsilon^{-n/8} + \rho_n,$$

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

We first note that an explicit form of $\rho_n$ in the above result can be found in Equation (47) given in [31]. Next, observe that the above double-binning test achieves the same minimax separation rate as $\phi_{\text{NBW},1}$ without being dependent on an unspecified constant in its critical value. Its guarantee holds over a smaller set of alternative distributions, namely $\mathcal{P}^\dagger_{1,\ell_1,\ell_2}(L)$ (in contrast, the result of [43] 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 $\gamma$-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 10 (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 $\phi_{\text{NBW},1}$. While we believe that a similar power result can be developed based on other test statistics including $T_{CI,W}$ used in $\phi_{\text{NBW},2}$, 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 $p_{\text{perm}}$ in (3). In particular, we draw $B$ permutations, denoted by $\pi_1', \ldots, \pi_B'$, from $\Pi$ with replacement. Then the permutation p-value using the test statistic $T_{CI}$ is defined as

$$p_{\text{perm}} := \frac{1}{B+1} \left[ \sum_{i=1}^B 1\{T_{CI}^{\pi_i'} \geq T_{CI}\} + 1 \right].$$

It is well-known that $p_{\text{perm}} \leq \alpha$ is a valid test in finite-sample scenarios, whenever $T_{CI}, T_{CI}^{\pi_1'}, \ldots, T_{CI}^{\pi_B'}$ are exchangeable under the null. Furthermore, $p_{\text{perm}}$ can be arbitrarily close to $p_{\text{perm}}$ 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 $\alpha = 0.05$. 

Marginal density of $X$ used in construction of our lower bound result (Theorem 4). In particular, we consider the local permutation test in Section 4. To generate the data, we consider a distributional setup for any given $M$.

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 $(X, Y) \perp Z$. 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 Equation (2) of [31] with $\epsilon = n^{-1}$, and the conditional density of $X|Z = z$ in Equation (10) of [31]. We further let $Y|Z = z$ have the same conditional density as $X|Z = z$ for all $z \in [0, 1]$, while satisfying $X \perp Y|Z$. As proved in Appendix B.5.1 and Appendix B.5.3 of [31], the considered distribution satisfies $\gamma$-Hellinger Lipschitzness with any fixed $\gamma \geq 1$ as well as $\gamma$-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 $\alpha$ as long as $nM^{-4} \rightarrow 0$. 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 $nM^{-4} \rightarrow \infty$. 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 $nM^{-4} \rightarrow 0$. On the other hand, the error tends to increase when $nM^{-4} \rightarrow \infty$, 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 $\hat{p}_{\text{perm}}$ in (22) but by drawing $B$ cyclic permutations from $\Pi_{\text{cyclic}}$ 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:

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

The considered distribution depends on the parameter $\theta$, which controls the smoothness of the conditional probability mass function. In particular, the conditional marginals (23) become more wiggly as $\theta$ 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 $\theta$ 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

\begin{align*}
    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} \quad \\
    p_{XY|Z}(X = 1, Y = 0 | z) &= p_{XY|Z}(X = 0, Y = 1 | z) = 4f(z)/5 - f(z)^2.
\end{align*}

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. Lastly, we refer readers to Appendix F of [31] for additional simulation results and a practical guideline on the choice of binning scheme.

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 [54], motivates us to consider reasonable assumptions under which CI testing becomes possible. Under certain smoothness assumptions,
Fig 3: Illustration of the power and type I error of the local permutation test using $T_{C_1}$. 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 $nM^{-4} \to \infty$.

Fig 4: Illustration of the type I error and power of the single- and double-binning tests using $T_{C_1}$. 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 $\theta$. 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.

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 [13, 43] 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.
Our work leaves several interesting directions of future work, including results under different metrics, adaptive binning schemes, depoissonization and minimax power results for multivariate $Z$. We refer to Appendix A of [31] for a detailed discussion.

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 acknowledges support from the Yonsei University Research Fund of 2021-22-0332 as well as support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384). IK also acknowledges support from EPSRC grant EP/N031938/1. The authors would like to thank the reviewers for their constructive comments and suggestions. The authors are also grateful to Richard J. Samworth, Rajen Shah and Anton Rask Lundborg for their helpful discussion on Theorem 1.

REFERENCES


