The Annals of Statistics
2021, Vol. 49, No. 4, 2151–2177
https://doi.org/10.1214/20-AOS2030

MINIMAX OPTIMAL CONDITIONAL INDEPENDENCE TESTING

BY MATEY NEYKOV[*], SIVARAMAN BALAKRISHNAN[†] AND LARRY WASSERMAN[‡]

Department of Statistics & Data Science, Carnegie Mellon University, [*]mneykov@stat.cmu.edu; [†]siva@stat.cmu.edu; [‡]larry@stat.cmu.edu

We consider the problem of conditional independence testing of X and Y given Z where X, Y and Z are three real random variables and Z is continuous. We focus on two main cases—when X and Y are both discrete, and when X and Y are both continuous. In view of recent results on conditional independence testing [*Ann. Statist.* **48** (2020) 1514–1538], one cannot hope to design nontrivial tests, which control the type I error for all absolutely continuous conditionally independent distributions, while still ensuring power against interesting alternatives. Consequently, we identify various, natural smoothness assumptions on the conditional distributions of $X, Y|Z = z$ as z varies in the support of Z, and study the hardness of conditional independence testing under these smoothness assumptions. We derive matching lower and upper bounds on the critical radius of separation between the null and alternative hypotheses in the total variation metric. The tests we consider are easily implementable and rely on binning the support of the continuous variable Z. To complement these results, we provide a new proof of the hardness result of Shah and Peters [*Ann. Statist.* **48** (2020) 1514–1538].

1. Introduction. Conditional independence (CI) testing is a fundamental problem, with widespread applications throughout statistics. From being a foundation of basic concepts such as sufficiency and ancillarity [13], to its applications in estimation and inference for graphical models [23, 24] and in causal inference and causal discovery [27, 33, 41], the concept of conditional independence and conditional independence testing play a central role in the fields of statistics, machine learning and related areas. A large body of work has focussed on CI testing under the assumption of joint Gaussianity. In this setting, CI testing corresponds to testing whether certain partial correlations between the variables are zero. Since partial correlations are (relatively) easy to estimate, the Gaussian assumption gives a shortcut to CI testing, but if the model is non-Gaussian this can lead to misleading conclusions as variables could be conditionally dependent even with zero partial correlation. In practice, the Gaussian assumption is unlikely to hold exactly and many applications call for the additional flexibility provided by nonparametric CI testing.

In this paper, we consider CI testing from a nonparametric perspective. Following Dawid [13], given three random vectors $(X, Y, Z) \in \mathbb{R}^{d_X+d_Y+d_Z}$ we will denote the CI of X and Y given Z by $X \perp\!\!\!\perp Y|Z$. In the case when $d_X = d_Y = d_Z = 1$ and Z is a continuous random variable supported on $[0, 1]$, we construct nonparametric tests which are capable of testing the null hypothesis $X \perp\!\!\!\perp Y|Z$ versus the alternative $X \not\perp\!\!\!\perp Y|Z$. The variables X and Y are allowed to be either both discrete or both continuous supported on $[0, 1]$. It was recently argued in a precise mathematical sense [31] that CI testing is a statistically hard task for absolutely continuous (with respect to the Lebesgue measure) random variables—namely if one wants to have a test that controls the type I error for all absolutely continuous triplets (X, Y, Z) such that $X \perp\!\!\!\perp Y|Z$, such a test cannot have power against any alternative. This discouraging result

Received July 2020.
MSC2020 subject classifications. 62G10.
Key words and phrases. Conditional independence, minimax optimality, hypothesis testing.

demystified the fact that despite a large body of literature on the subject, no fully satisfactory CI tests had been developed for continuous random variables.

Concurrently with the paper of Shah and Peters [31], the work of Canonne et al. [12] constructed tests for CI of *discrete* distributions (X, Y, Z) which are minimax optimal in certain regimes. Part of the effort of this paper is devoted to extending the ideas of Canonne et al. [12] to the case when Z is an absolutely continuous random variable on $[0, 1]$.

In order to characterize the difficulty of CI testing in this setting, we adopt the minimax perspective [21, 22]. Naturally, if an alternative distribution is very close to a null distribution (in a certain metric such as the total variation metric) it will be very difficult to test for CI given a finite number of n samples. By discarding distributions under the alternative that are "ε_n-close" to the null hypothesis, we are able to set up a well-defined testing problem. The goal in minimax hypothesis testing is then to characterize the optimal "critical radius" ε_n, that is, the smallest ε_n at which it is possible to reliably distinguish the null from the ε_n-separated alternative, as a function of the sample size n. This standard step of discarding "near-null distributions" is insufficient as one cannot hope to design a nontrivial test which controls the type I error for *all* conditionally independent absolutely continuous triplets [31]. In order to make the problem of CI testing well-posed, we further impose certain natural smoothness assumptions on the conditional distributions of $X, Y|Z = z$ as z varies in the support of Z, and establish upper and lower bounds on the critical radius of conditional independence testing under these smoothness assumptions.

1.1. *Related work.* As we mentioned earlier, there is a large body of work on independence and CI testing. We focus our review on the literature most relevant to our approach. It is worth noting that almost all relevant works considered here, with the notable exception of Canonne et al. [12] who consider minimax CI testing for discrete distributions, do not take a minimax perspective to the problem. We are not aware of tests that achieve the minimax rates for testing CI with a continuous random variable Z other than the ones that we develop in this paper. In addition, we would like to note that the ideas introduced by Cannone et al. [12] are instrumental in the development of the minimax rates in the present work. In particular, [12] offer a variety of results in the discrete X, Y, Z conditional independence testing, including lower and upper bounds on the sample complexity. We borrow key constructs from this work, particularly an unbiased estimator of the L_2^2 distance, and tools to analyze its variance and expectation under Poisson sampling in order to come up with upper bounds for our estimators.

Given knowledge of the conditional distribution of $X|Z$, Berrett et al. [9] develop a permutation-based test for testing the null hypothesis of CI. We note that from a minimax perspective knowing $X|Z$ changes the problem of CI testing significantly and we do not address this CI testing variant here. The works [7, 8] propose a partial copula approach, which needs estimators of the conditional distributions of $X|Z$ and $Y|Z$. Since estimation is typically more costly than testing, we anticipate that such a procedure does not attain minimax optimal rates for the critical radius. In a setting different from the present paper, Song [32] proposes a CI test for two variables given a single index of a random vector via "Rosenblatt transforms," which are multivariate extensions of the probability integral transform. The techniques in this work also involve estimation of certain conditional distributions via kernel smoothing. Huang [20] proposes a nonparametric CI test using the so called maximal nonlinear conditional correlation. The author proves that under the null hypothesis given that certain conditions hold, the test achieves asymptotic normality. This work once again requires kernel smoothed estimates of certain conditional expectations and is therefore unlikely to result in minimax optimal tests of CI. In an interesting paper, Györfy and Walk [18] extend the independence testing results of Gretton and Györfi [17] to the CI case, and propose strongly

consistent nonparametric tests. We believe however that there is a gap in one of the proofs of this work, which would otherwise seem to contradict the CI hardness results of Shah and Peters [31]. In particular, in the proof of Theorem 1 of Györfi and Walk [18], it is claimed that the following expression is 0:

$$\left|\mathbb{P}(X \in A, Y \in B, Z \in C) - \frac{\mathbb{P}(X \in A, Z \in C)\mathbb{P}(Y \in A, Z \in C)}{\mathbb{P}(Z \in C)}\right|,$$

under the null hypothesis of independence, where A, B, C are elements of a partition of the domains of X, Y, Z, respectively. Note that this need not hold in general for conditionally independent distributions since averaging over Z does not necessarily preserve independence. In fact, this is one of the major complications that we have to deal with in our proofs.

Patra, Sen and Székely [26] design a novel nonparametric residual between a random variable and a random vector and use it to develop tests of CI with the help of the bootstrap. An innovative approach to nonparametric CI testing using a nearest neighbor bootstrap and converting the testing problem to a classification problem was recently proposed by Sen et al. [29]. Fukumizu et al. [16] give a measure of CI of random variables, based on normalized cross-covariance operators on reproducing kernel Hilbert spaces. Different reproducing kernel based methods were proposed by Zhang et al. [41] and Doran et al. [15], respectively. The recent work of Shah and Peters [31], along with the hardness result, proposes CI tests based on the so called generalized covariance measure which is a measure related to the normalized residuals of regressing X and Y on Z. In another recent paper, Azadkia and Chatterjee [3] propose a novel measure of CI which takes values in $[0, 1]$, where the measure takes the value 0 when the variables are conditionally independent, and is equal to 1 when the Y is a measurable function of X given Z.

There is also a significant amount of work on CI testing in the econometrics literature (see, for instance, [34–36, 39]). Su and White [35] give a Hellinger distance based approach to CI testing, which employs a plug in based estimate using kernel smoothed estimates of the joint and conditional densities of X, Y, Z. In follow-up work, Su and White [34] propose estimating a functional involving the difference of two conditional characteristic functions. They show asymptotic normality under the null hypothesis and explore the power of the test based on this estimator under local alternatives. In the work [36], the authors propose an empirical likelihood based approach to CI testing. Wang and Hong [39] develop a new test based on characteristic functions, which achieves faster rates against certain local alternatives in comparison to the test developed by Su and White [34].

So far we have discussed works which focus on nonparametric CI testing in the continuous case. It is noteworthy that there are also numerous CI tests in the discrete case as well. See, for example, the works [1, 12, 28, 40] as well as references therein.

1.2. *Summary of results.* We will now informally summarize the main findings of our work. For the most part, this paper is focused on the following two cases:

1. When X and Y are discrete supported on $[\ell_1] \times [\ell_2]$ for some integers ℓ_1, ℓ_2 (here, $[\ell_1] = \{1, 2, \ldots, \ell_1\}$ and similarly for $[\ell_2]$), and when Z has an absolutely continuous (with respect to the Lebesgue measure) distribution supported on $[0, 1]$,
2. When all three variables (X, Y, Z) have an absolutely continuous (with respect to the Lebesgue measure) distribution supported on $[0, 1]$.

We study the minimax rate for the critical radius ε_n which we define as the separation between the null and alternative hypothesis, in the total variation (TV) distance, required to reliably distinguish them. Formally, we consider distinguishing

$$H_0 : p_{X,Y,Z} \quad \text{s.t. } X \perp\!\!\!\perp Y | Z \quad \text{versus}$$

$$H_1 : p_{X,Y,Z} \quad \text{s.t. } \inf_{q \text{ in } H_0} \|p_{X,Y,Z} - q\|_1 \geq \varepsilon_n.$$

TABLE 1
This is a summary of the minimax results obtained in the main text of our paper

	X, Y		
	Discrete on $[\ell_1] \times [\ell_2]$, ℓ_1, ℓ_2 fixed	Discrete on $[\ell_1] \times [\ell_2]$	Continuous
ε_n-Upper Bounds	$n^{-2/5}$	$\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$, given $\frac{\ell_1^4}{\ell_2} \lesssim n^3$	$n^{-2s/(5s+2)}$
ε_n-Lower Bounds	$n^{-2/5}$	$\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$	$n^{-2s/(5s+2)}$

In addition, we remove distributions under H_0 and H_1 which are not smooth enough, that is, $p_{X,Y|Z=z}$ is not a smooth function of z (for precise definitions, refer to Section 2.3). Given this set-up, our interest is in finding the smallest possible ε_n such that even in the worst-case scenario for distributions under H_0 and under H_1 the sum of the type I and type II errors can be controlled under a prespecified threshold.

1. Let us first discuss the case when X and Y are discrete on $[\ell_1] \times [\ell_2]$ where ℓ_1 and ℓ_2 are fixed integers which are not allowed to scale with n. In this setting, we show that

$$\varepsilon_n \asymp n^{-2/5}.$$

That is we show matching minimax lower and upper bounds at the optimal rate of the critical radius which is given by $n^{-2/5}$. Here, we use $\asymp$ to mean equal up to a positive absolute constant.

2. Next, consider the more general case when ℓ_1 and ℓ_2 are allowed to scale with n. Then we are able to show that

$$\varepsilon_n \gtrsim \frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}} \wedge 1,$$

and we have a matching upper bound (i.e., a test) whenever, for $\ell_1 \geq \ell_2$, we have $\frac{\ell_1^4}{\ell_2} \lesssim n^3$. We further show that this latter condition holds whenever $\ell_1 \asymp \ell_2$. Here, and throughout this paper, $\gtrsim$ and $\lesssim$ mean inequalities up to a positive absolute constant.

3. Finally, in the fully continuous case we show that

$$\varepsilon_n \asymp n^{-2s/(5s+2)},$$

where s denotes the Hölder smoothness parameter of the conditional density $p_{X,Y|Z}$ under the alternative hypothesis.

Our results are also summarized in Table 1. The tests used to achieve the upper bounds for the above minimax rates, are computationally tractable and we implement them and provide some numerical results. Our tests do not require kernel smoothing. They are rather calculated based on binning the support of Z (and X and Y when they are continuous) into a certain sample-size dependent number of bins. For each Z-bin, a (weighted) U-statistic is calculated and the resulting statistics are summed up according to appropriate weighting across the Z-bins. Roughly, the U-statistics target the L_2^2 distance between $p_{X,Y|Z}$ and $p_{X|Z}p_{Y|Z}$ within each of the Z-bins (or in the weighted U-statistic case a distance similar to the chi-square distance between $p_{X,Y|Z}$ and $p_{X|Z}p_{Y|Z}$). This strategy also reveals the need to impose certain smoothness assumptions on the conditional distribution of $p_{X,Y|Z=z}$ in z since otherwise the binning may result in unreliable estimates of the L_2^2 distance.

Along with the aforementioned results, we also provide a new proof of the hardness result of Shah and Peters [31]. Our proof is based on a coupling between an arbitrary absolutely

continuous distribution and a statistically independent distribution, which bears some resemblance to the coupling used in Lemma 14 of [31]. We use this coupling to show the fact that conditionally independent distributions are Wasserstein dense in the set of all absolutely continuous distributions of bounded support.

1.3. *Organization.* The paper is structured as follows. We present some basic background in Section 2. We revisit the hardness results of Shah and Peters [31] in Section 3. Minimax lower bounds on the critical radius are given in Section 4. Section 5 is devoted to developing tests of CI which match the lower bounds of Section 4. Section 6 gives examples for distributions satisfying the smoothness assumptions we impose in Sections 4 and 5. Section 7 provides a brief numerical study, which is meant to show that our nonparametric tests are in fact readily implementable and perform well in practice. Finally, a discussion is provided in Section 8.

2. Background. In this section, following some basic notation, we present some background on minimax testing and briefly introduce the various smoothness conditions we use in our minimax upper and lower bounds.

2.1. *Notation.* We make extensive usage of metrics on probability distributions in this paper. The total variation (TV) metric between two distributions p, q on a measurable space $(\Omega, \mathcal{F})$ is defined as

$$d_{\mathrm{TV}}(p,q) = \sup_{A\in\mathcal{F}} \big|p(A) - q(A)\big| = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\int \left|\frac{dp}{d\nu} - \frac{dq}{d\nu}\right| d\nu,$$

where the last identity assumes ν is a common dominating measure of p and q, that is, $p \ll \nu$ and $q \ll \nu$ and $\frac{dp}{d\nu}, \frac{dq}{d\nu}$ denote the densities of p and q with respect to ν (note here that ν can always be taken as $\nu = p + q$). Under the latter assumption, one can also define the L_2 distance between p and q as

$$\|p-q\|_2 = \left[\int \left|\frac{dp}{d\nu} - \frac{dq}{d\nu}\right|^2 d\nu\right]^{1/2}.$$

Assuming that $p \ll q$, we may define the χ^2-divergence between p and q as

$$d_{\chi^2}(p,q) = \int \left(\frac{dp}{dq} - 1\right)^2 dq.$$

If $p \ll q$ fails to hold, then we take $d_{\chi^2}(p,q) = \infty$.

Next, we formalize our notation for conditional distributions. If the triplet (X, Y, Z) has a distribution $p_{X,Y,Z}$, we will use $p_{X,Y|Z=z}$ to denote the conditional joint distribution of $X, Y|Z = z$. Additionally, $p_{X|Z=z}$ and $p_{Y|Z=z}$ will denote the marginal conditional distributions of $X|Z = z$ and $Y|Z = z$, respectively. The marginal distributions will be denoted with p_X, p_Y, p_Z and joint marginal distributions will be denoted with $p_{X,Y}, p_{Y,Z}, p_{X,Z}$. Furthermore, with a slight abuse of notation, $p_{X,Y|Z}(x, y|z)$ and $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$ will denote the densities of these distributions evaluated at the points x, y and z (or the corresponding probability mass functions when X and Y are discrete).

In addition, we will use $\lesssim$ and $\gtrsim$ to mean $\leq$ and $\geq$ up to positive universal constants (which may be different from place to place). If both $\lesssim$ and $\gtrsim$ hold, we denote this as $\asymp$. For an integer $n \in \mathbb{N}$, we use the convenient shorthand $[n] = \{1, 2, \ldots, n\}$.

2.2. *Minimax testing.* In order to characterize the complexity of CI testing, we use the minimax testing framework, introduced in the work of Ingster and coauthors [21, 22], and which has since then been considered by many authors (see, for instance, [2, 4–6, 11, 12, 14, 37]). Formally, consider the testing problem

$$H_0 : p \in \mathcal{H}_0 \quad \text{vs} \quad H_1 : p \in \mathcal{S}_1(\varepsilon), \tag{2.1}$$

where $\mathcal{S}_1(\varepsilon) := \{p \in \mathcal{H}_1 : \inf_{q \in \overline{\mathcal{H}}_0} \|p - q\|_1 \geq \varepsilon\}$, and $\mathcal{H}_0 \subseteq \overline{\mathcal{H}}_0$ and $\mathcal{H}_1$ are prespecified sets of distributions. We define the minimax risk of testing as

$$R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) = \inf_{\psi} \Big\{ \sup_{p \in \mathcal{H}_0} \mathbb{E}_p[\psi(\mathcal{D}_n)] + \sup_{p \in \mathcal{S}_1(\varepsilon)} \mathbb{E}_p[1 - \psi(\mathcal{D}_n)] \Big\}^{1}, \tag{2.2}$$

where the infimum is taken over all Borel measurable test functions $\psi : \text{supp}(\mathcal{D}_n) \mapsto [0, 1]$ (which gives the probability of rejecting the null hypothesis), and $\text{supp}(\mathcal{D}_n)$ is the support of the random variables $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots (X_n, Y_n, Z_n)\}$. We note that it is common to choose the sets $\mathcal{H}_0$ and $\overline{\mathcal{H}}_0$ to be identical. However, as will be clearer hereafter, in the setting of CI testing we will choose $\mathcal{H}_0$ to be a subset of distributions which are conditionally independent *and* appropriately smooth, while we will choose $\overline{\mathcal{H}}_0$ to be the set of *all* conditionally independent distributions.

In the minimax framework, our goal is to study the *critical radius* of testing defined as

$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) = \inf \Big\{ \varepsilon : R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \leq \frac{1}{3} \Big\}. \tag{2.3}$$

The constant $\frac{1}{3}$ above is arbitrary, and can be chosen as any small constant. The minimax testing radius or the critical radius, corresponds to the smallest radius ε at which there exists *some test* which distinguishes distributions in $\mathcal{H}_0$ from those in $\mathcal{H}_1$ which are appropriately far from $\mathcal{H}_0$. The critical radius provides a fundamental characterization of the statistical difficulty of the hypothesis testing problem in (2.1).

2.3. *Smoothness conditions.* In Sections 4 and 5, we derive upper and lower bounds on the minimax critical radius for conditional independence testing. However, in view of the results of Shah and Peters [31], and our own results in Section 3, we must impose some restrictions on the distributions under consideration in order to obtain nontrivial minimax rates. Broadly, we restrict our attention to settings where the conditional distributions are appropriately smooth.

We focus on two main settings in our work, the setting where X and Y are discrete but Z is continuous and when all three are continuous. For the case when X and Y are discrete and Z is continuous, we consider Z that is supported on $[0, 1]$. Define the set of distributions $\mathcal{E}'_{0,[0,1]}$ as distributions whose generating mechanism of the triple (X, Y, Z) supported on $\mathbb{R}^3$ is as follows: first, a Z from the distribution p_Z (which is absolutely continuous with respect to the Lebesgue measure) with support $[0, 1]$ is generated. Next, X and Y are generated from the distribution $p_{X,Y|Z}$, which is supported on[2] $[\ell_1] \times [\ell_2]$ for (almost) all Z. Denote by $\mathcal{P}'_{0,[0,1]} \subset \mathcal{E}'_{0,[0,1]}$ the set of null distributions (i.e., distributions such that $X \perp\!\!\!\perp Y|Z$) and let $\mathcal{Q}'_{0,[0,1]} = \mathcal{E}'_{0,[0,1]} \setminus \mathcal{P}'_{0,[0,1]}$. Similarly, in the case when X, Y and Z are continuous, we let $\mathcal{P}_{0,[0,1]^3} \subset \mathcal{E}_{0,[0,1]^3}$ be the set of distributions for which $X \perp\!\!\!\perp Y|Z$ and let $\mathcal{Q}_{0,[0,1]^3} = \mathcal{E}_{0,[0,1]^3} \setminus \mathcal{P}_{0,[0,1]^3}$.

With these preliminaries in place, we can define the various smoothness classes that we work with in this paper:

[1] Here and throughout, with a slight abuse of notation, we use $\mathbb{E}_p$ to denote expectation under i.i.d. data $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)\}$ where each observation is drawn from p.

[2] It is not crucial here that $X, Y|Z$ is supported on $[\ell_1] \times [\ell_2]$. It could be supported on any set $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = \ell_1$ and $|\mathcal{Y}| = \ell_2$. Here, for the sake of simplicity of presentation, we focus only on the case $[\ell_1] \times [\ell_2]$.

DEFINITION 2.1 (Null Lipschitzness).

1. Null TV Lipschitzness: Let $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L) \subset \mathcal{P}'_{0,[0,1]}$ (analogously $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L) \subset \mathcal{P}_{0,[0,1]^3}$) be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]$ we have

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le L|z - z'| \quad \text{and} \quad \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \le L|z - z'|,$$

where $p_{X|Z=z}$ and $p_{Y|Z=z}$ denote the conditional distributions of $X|Z=z$ and $Y|Z=z$ under $p_{X,Y,Z}$ respectively.

2. Null χ^2 Lipschitzness: Let $\mathcal{P}'_{0,[0,1],\chi^2}(L) \subset \mathcal{P}'_{0,[0,1]}$ (analogously $\mathcal{P}_{0,[0,1]^3,\chi^2}(L) \subset \mathcal{P}_{0,[0,1]^3}$) be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]$ we have

$$d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) \le L|z - z'| \quad \text{and} \quad d_{\chi^2}(p_{Y|Z=z}, p_{Y|Z=z'}) \le L|z - z'|,$$

where $p_{X|Z=z}$ and $p_{Y|Z=z}$ denote the conditional distributions of $X|Z=z$ and $Y|Z=z$ under $p_{X,Y,Z}$, respectively. The distance $d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'})$ is considered ∞ if $p_{X|Z=z} \ll p_{X|Z=z'}$ is violated.

3. Null Hölder Lipschitzness: Let $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L) \subset \mathcal{P}'_{0,[0,1]}$ be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]$ we have

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le \sqrt{L|z - z'|} \quad \text{and} \quad \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \le \sqrt{L|z - z'|},$$

where $p_{X|Z=z}$ and $p_{Y|Z=z}$ denote the conditional distributions of $X|Z=z$ and $Y|Z=z$ under $p_{X,Y,Z}$, respectively.

Under the alternative, we consider slightly different classes in the discrete and continuous cases. Formally, we define the following class for the discrete X and Y setting.

DEFINITION 2.2 (Alternative TV Lipschitzness). Let $\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L) \subset \mathcal{Q}'_{0,[0,1]}$ be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]$ we have

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \le L|z - z'|,$$

where $p_{X,Y|Z=z}$ denotes the conditional distribution of $X, Y|Z=z$ under $p_{X,Y,Z}$.

In the continuous case, we will further restrict our attention to distributions which in addition to being TV smooth (as above), also have smooth conditional density $p_{X,Y|Z}$. In order for us to impose proper smoothness on the density $p_{X,Y|Z}$, we will first define a Hölder smoothness class.

DEFINITION 2.3 (Hölder smoothness). Let $s > 0$ be a fixed real number, and let $\lfloor s \rfloor$ denote the maximum integer strictly smaller than s. Denote by $\mathcal{H}^{2,s}(L)$, the class of functions $f : [0,1]^2 \mapsto \mathbb{R}$, which posses all partial derivatives up to order $\lfloor s \rfloor$ and for all $x, y, x', y' \in [0,1]$ we have

$$\begin{aligned} &\sup_{k \le \lfloor s \rfloor} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x,y) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x',y') \right| \\ &\qquad \le L\big((x-x')^2 + (y-y')^2\big)^{\frac{s - \lfloor s \rfloor}{2}}, \end{aligned} \tag{2.4}$$

and in addition

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

In the above assumption, in the addition to the usual Hölder smoothness assumption, we assume that there is a uniform bound on all derivatives of lower than $\lfloor s \rfloor$ order. When $s = 1$, the above is simply the class of L-Lipschitz functions.

DEFINITION 2.4 (Alternative Lipschitzness). Let $\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L, s) \subset \mathcal{Q}_{0,[0,1]^3}$ be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0, 1]$ we have

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \leq L|z - z'|,$$

where $p_{X,Y|Z=z}$ denotes the conditional distribution of $X, Y|Z = z$ under $p_{X,Y,Z}$. In addition, we assume that for all $z, x, y \in [0, 1]$: $p_{X,Y|Z}(x, y|z) \in \mathcal{H}^{2,s}(L)$.

We devote Section 6 to investigating various relationships between these different Lipschitzness assumptions, as well as to constructing broad nonparametric classes of distributions which satisfy these Lipschitzness conditions.

3. The hardness of CI testing revisited. In this section, we revisit the recent work of Shah and Peters [31]. In order for us to review their results, and to build upon them, we will recall their notation. Let $\mathcal{E}_0$ denote the set of all distributions for (X, Y, Z) on $\mathbb{R}^{d_X+d_Y+d_Z}$, which are absolutely continuous with respect to the Lebesgue measure. Define the set of conditionally independent distributions, that is, distributions such that $X \perp\!\!\!\perp Y|Z$, as $\mathcal{P}_0 \subset \mathcal{E}_0$. Let $\mathcal{E}_{0,M} \subseteq \mathcal{E}_0$ be the set of distributions whose support is contained within an L_∞ ball of radius M. Define the set of alternative distributions as $\mathcal{Q}_0 = \mathcal{E}_0 \setminus \mathcal{P}_0$ and $\mathcal{P}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{P}_0$ and $\mathcal{Q}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{Q}_0$.

In their Proposition 5, Shah and Peters argue that the null and alternative sets of distributions $\mathcal{P}_{0,M}$ and $\mathcal{Q}_{0,M}$ are separated in TV distance. Here, separated is meant in the sense that there exists a distribution from $\mathcal{Q}_{0,M}$ which is at least $1/24$ apart in TV distance from any distribution in $\mathcal{P}_{0,M}$. Similarly, in Proposition 16, Shah and Peters argue that the sets of distributions $\mathcal{P}_0$ and $\mathcal{Q}_0$ are separated in KL divergence (in this proposition they consider only the case $(X, Y, Z) \in \mathbb{R}^3$). In contrast, the first result of this section will show that when the Wasserstein distance is considered, the set of distributions $\mathcal{P}_{0,M}$ is dense in the set $\mathcal{Q}_{0,M}$. Let us first define the Wasserstein distance.

DEFINITION 3.1 (Wasserstein distance). Let $p \geq 1$ be a real number. Let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of measures μ on $(\mathbb{R}^d, \|\cdot\|_2)$, such that there exists $x_0 \in \mathbb{R}^d$ for which

$$\int_{\mathbb{R}^d} \|x - x_0\|_2^p d\mu(x) < \infty.$$

For two probability measures, μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ the p^{th} Wasserstein distance between μ and ν is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^p d\gamma(x, y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is set of all couplings between the measures μ and ν, that is, all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$, with marginals μ and ν.

We are now ready to state the first result of this section.

LEMMA 3.2 (Wasserstein denseness). *Take any distribution $P \in \mathcal{E}_{0,M}$ for some $M > 0$. Then for any $p \geq 1$ and any $\varepsilon > 0$ there exists a distribution $Q \in \mathcal{P}_{0,M}$ such that*

$$W_p(P, Q) \leq \varepsilon.$$

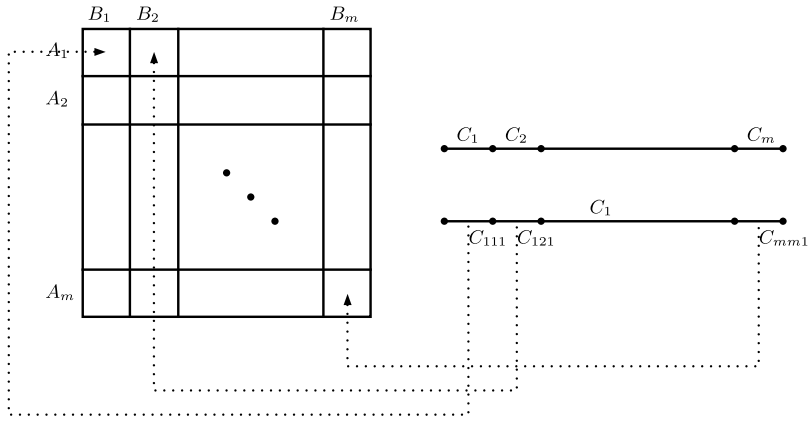

FIG. 1. *This schematic describes the construction of Q from P. $[-M, M]$ is divided in intervals $\{A_1, \ldots, A_m\}$, $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$. Next, each interval C_k is subdivided into m^2 smaller subintervals. The interval C_1 is displayed along with its subdivisions of C_{ij1} for $i, j \in [m]$. Each little interval C_{ij1} corresponds to a pair (A_i, B_j) or equivalently to a cell $A_i \times B_j$ in $[-M, M]^2$.*

PROOF. For simplicity, we will prove this result for the one-dimensional case $d_X = d_Y = d_Z = 1$. The proof extends trivially to the more general case. First, note that since both $P, Q \in \mathcal{E}_{0,M} \subseteq \mathcal{P}_p(\mathbb{R}^3)$, the Wasserstein distance between P and Q is well defined. We will now construct Q from P by describing a coupling between the two distributions.

Let $\{A_1, \ldots, A_m\}$ denote an equipartition of $[-M, M]$ in intervals. Similarly, let $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$ be equipartitions of $[-M, M]$. Divide each C_k further in m^2 subintervals of equal length denoted by C_{ijk}, so that each of these small intervals corresponds to a pair (A_i, B_j). Refer to Figure 1 for a visualization of this construction. The lengths of each interval A_i, B_i or C_i is $\frac{2M}{m}$, while the length of an interval C_{ijk} is $\frac{2M}{m^3}$. Given a draw $(X, Y, Z) \sim P$, we construct $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}) \sim Q$ as follows. Suppose that $X \in A_i$, $Y \in B_j$ and $Z \in C_k$. Then we generate uniformly $\widetilde{Z} \in C_{ijk}$ and $(\widetilde{X}, \widetilde{Y})$ uniformly in $A_i \times B_j$. By definition then $\widetilde{X} \perp\!\!\!\perp \widetilde{Y} | \widetilde{Z}$, and further $\widetilde{X}, \widetilde{Y}, \widetilde{Z} \in [-M, M]$. Hence $Q \in \mathcal{P}_{0,M}$.[3] Furthermore, we can bound the Wasserstein distance for this particular coupling as

$$W_p(P, Q)^p \leq \mathbb{E}\mathbb{E}_{(\widetilde{X}, \widetilde{Y}, \widetilde{Z})|(X,Y,Z)} \|(X, Y, Z) - (\widetilde{X}, \widetilde{Y}, \widetilde{Z})\|_2^p \leq \left(\sqrt{3}\frac{2M}{m}\right)^p.$$

Since m can be selected arbitrarily large, the above can be made smaller than ε^p. This completes the proof. □

The construction used to obtain Q from P in the above result captures intuitively the essence of the "hardness" of CI testing with continuous Z. The set $\mathcal{P}_{0,M}$ contains distributions, which allow the conditional distributions of $X, Y | Z = z$ to be "wildly discontinuous" as functions of z. This in turn allows for the existence of distributions in $\mathcal{P}_{0,M}$ capable of approximating any distribution in $\mathcal{E}_{0,M}$ in the Wasserstein metric. Later in this paper we will see that, if we disallow distributions in $\mathcal{E}_{0,M}$ whose conditional distributions can be wildly variable in z, CI testing becomes possible. We would also like to point out the intuition why the Wasserstein distance yields a result like Lemma 3.2 in contrast to using TV distance or KL divergence. The Wasserstein distance is based on a metric (in our case the L_2 metric) on the underlying sample space, and as a consequence has the critical feature (unlike the KL divergence or TV distance) that it is robust to small perturbations in the sample space (on

[3]For a precise expression of the density of Q, refer to Appendix A in the supplementary material [25].

the other hand, metrics like the TV metric are typically stable to small perturbations in the probability space). Indeed, the heart of the construction of Shah and Peters [31] and of our own result, is the idea that given a sample from a conditionally dependent distribution, one can perturb it slightly (effectively "encoding" the value of X in Z) to create a conditionally independent distribution. This operation, effectively a small perturbation in the sample space, does not change the Wasserstein distance much, but can have a large effect on the TV distance or KL divergence.

Lemma 3.2 suggests, but does not imply that CI testing is "hard." Building on the construction of Lemma 3.2, we give a new simpler proof of the "no-free-lunch" theorem of Shah and Peters [31], see Theorem 2. For convenience of the reader, we restate the no-free-lunch theorem below, and give a complete proof in Appendix A of the supplementary material. Let $d = d_X + d_Y + d_Z$, and suppose that we observe n observations $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$.

THEOREM 3.3 (No-free-lunch). *Given any $n \in \mathbb{N}$, $\alpha \in (0, 1)$, $M \in (0, \infty]$ and a potentially randomized test $\psi_n : \mathbb{R}^{nd} \times [0, 1] \mapsto \{0, 1\}$, that has valid level α for the null hypothesis $\mathcal{P}_{0,M}$, we have that $\mathbb{P}_Q(\psi_n = 1) \leq \alpha$ for all $Q \in \mathcal{Q}_{0,M}$.*

As stated, Theorem 3.3 assumes that (X, Y, Z) have a distribution which is continuous with respect to the Lebesgue measure. Suppose that ℓ_1 and ℓ_2 are two fixed and finite integers. We assume that X and Y are supported on $[\ell_1]$ and $[\ell_2]$, respectively, and that Z is supported on $[-M, M]^{d_Z}$ and has a continuous density with respect to the Lebesgue measure. The generating mechanism of the triple (X, Y, Z) is as follows: first a Z from the distribution P_Z is generated. Next, X and Y are generated from the distribution $P_{X,Y|Z}$ which is supported on $[\ell_1] \times [\ell_2]$ for (almost) all Z. Denote the set of all such distributions with $\mathcal{E}'_{0,M}$ (where we omit the dependence of $\mathcal{E}'_{0,M}$ on d_Z, ℓ_1, ℓ_2 for simplicity). Let $\mathcal{P}'_{0,M} \subset \mathcal{E}'_{0,M}$ be the subset of $\mathcal{E}'_{0,M}$ consisting of distributions such that $X \perp\!\!\!\perp Y|Z$ and $\mathcal{Q}'_{0,M} = \mathcal{E}'_{0,M} \setminus \mathcal{P}'_{0,M}$. Again as before we assume that we observe n observations $\mathcal{D}_n$. We have the following simple corollary to Theorem 3.3, which was alluded to by Shah and Peters [31].

COROLLARY 3.4 (Discrete no-free-lunch). *Given $n \in \mathbb{N}$, $\alpha \in (0, 1)$, $M \in (0, \infty)$ and a potentially randomized test ψ_n, that has a valid level α for the null hypothesis $\mathcal{P}'_{0,M}$, we have that $\mathbb{P}_Q(\psi_n = 1) \leq \alpha$ for all $Q \in \mathcal{Q}'_{0,M}$.*

Intuitively, Corollary 3.4 reveals that it is the continuity of Z that makes CI testing "hard," and not the continuity of X and Y.

4. Minimax lower bounds. In this section, we present our minimax lower bounds on the critical radius for conditional independence testing in various settings. Our first main result (Theorem 4.1) develops a lower bound on the critical radius in the case when X and Y are discrete and Z is continuous. Our next main result (Theorem 4.2) develops an analogous bound for the setting when X, Y and Z all have continuous distributions.

4.1. *X and Y discrete, Z continuous case.* We begin by recalling the Lipschitzness classes $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ and $\mathcal{P}'_{0,[0,1],\chi^2}(L)$ introduced in Definition 2.1, and $\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L)$ introduced in Definition 2.2. In this section, we develop a lower bound on the critical radius for distinguishing the conditionally independent distributions in any one of the null classes $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ and $\mathcal{P}'_{0,[0,1],\chi^2}(L)$ from the alternative class of conditionally dependent distributions $\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L)$. Formally, we have the following result.

THEOREM 4.1 (Critical radius lower bound). *Let $\overline{\mathcal{H}}_0 = \mathcal{P}'_{0,[0,1]}$. Suppose that $\mathcal{H}_0$ is either of $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ or $\mathcal{P}'_{0,[0,1],\chi^2}(L)$, while $\mathcal{H}_1 = \mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L)$ for some fixed $L \in \mathbb{R}^+$. Then for some absolute constant $c_0 > 0$ the critical radius defined in (2.3) is bounded as*

$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \geq c_0\bigg(\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}} \wedge 1\bigg).$$

REMARKS.

- In the case when ℓ_1 and ℓ_2 are constant, our lower bound on the critical radius
$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \geq c_0 n^{-2/5},$$
scales as the familiar rate for goodness-of-fit testing in the nonparametric setting of $\varepsilon_n \asymp n^{-2s/(4s+d)}$ [2, 5, 22] (where in our setting we take $d = 1$ and $s = 1$, corresponding to the one-dimensional Lipschitz smooth component Z).

  We note that as is typical in hypothesis testing problems this rate is faster than the $n^{-1/3}$ rate that we would expect for estimating a univariate Lipschitz smooth density, highlighting the fact that in many cases, from a statistical perspective, hypothesis testing is easier than estimation.
- On the other hand, the scaling of ε_n with ℓ_1 and ℓ_2 has a typical square-root dependence seen in *parametric* hypothesis testing problems [22, 37], where roughly we see that the critical radius shrinks provided that $\sqrt{\ell_1\ell_2}/n \to 0$. Once again this is in contrast to the linear dependence we would expect in estimating a multinomial distribution on $\ell_1 \times \ell_2$ categories, which would require $\ell_1\ell_2/n \to 0$ for consistent estimation.

  Thus we see that the lower bound we obtain for CI testing in the setting where X and Y are discrete, and Z is continuous blends parametric and nonparametric hypothesis testing rates. In Section 5, we develop matching upper bounds in various settings.
- We note in passing that our lower bound applies when the null distribution is restricted to belong to any of the three Lipschitzness classes introduced in Definition 2.1.
- We give the proof of Theorem 4.1 in Appendix B of the supplementary material. We note that at a high level we follow the strategy of Ingster [21] of creating a carefully chosen collection of possible densities under the alternative, and lower bounding the performance of the (optimal) likelihood ratio test in distinguishing a fixed null distribution against a uniform mixture of the selected distributions under the alternative. However, in our setting additional care is needed when perturbing the X and Y components in order to ensure that they remain valid discrete distributions (see Figure 1, and the associated construction), and to characterize the distance of our perturbed distributions from the manifold of conditionally independent distributions.

4.2. *X, Y and Z continuous case.* We first recall the Lipschitzness classes $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ introduced in Definition 2.1, and $\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,s)$ introduced in Definition 2.4. We derive a lower bound on the critical radius for distinguishing the conditionally independent distributions in either of the null classes $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ from the alternative class of conditionally dependent distributions $\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,s)$. Formally, we have the following result.

THEOREM 4.2 (Critical radius lower bound). *Let $\overline{\mathcal{H}}_0 = \mathcal{P}_{0,[0,1]^3}$. Suppose that $\mathcal{H}_0$ is either $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ or $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$, and $\mathcal{H}_1 = \mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,s)$ for some fixed $L \in \mathbb{R}^+$. Then we have that for some absolute constant $c_0 > 0$,*

$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \geq \frac{c_0}{n^{2s/(5s+2)}}.$$

REMARK.

- We note that our lower bound applies when the null distribution is restricted to belong to either of the classes $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$. Our proof in this setting builds on that of Theorem 4.1. In this case, to create a collection of distributions under the alternative we perturb the null distribution by smooth, infinitely differentiable bumps along all three coordinates in a carefully constructed fashion. By an appropriate choice of various parameters, we ensure that the distributions we construct satisfy the Lipschitzness and Hölder smoothness conditions required by the class $\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,s)$, while still remaining sufficiently far from the conditional independence manifold. We provide the details of our construction, as well as the subsequent analysis of the likelihood ratio test in Appendix B of the supplementary material.

5. Minimax upper bounds. In this section, we provide matching (in certain regimes) upper bounds to the lower bounds given in Section 4.

5.1. *Upper bound with finite discrete X and Y.* In this section, we will suggest a conditional independence test to match the lower bound of Section 4.1 when $\ell_1, \ell_2 = O(1)$ are not allowed to scale with n. In this case, the bound of Theorem 4.1 simply states that the critical radius is bounded from below by $cn^{-2/5}$, for some sufficiently small constant $c > 0$. To start the preparation for our test statistic, we will first reintroduce certain unbiased estimators from the work of Canonne et al. [12]. Our exposition and treatment of their estimators is novel, and builds on classical work on U-statistics [19, 30].

Suppose we observe $\sigma \geq 4$ observations of two discrete covariates X' and Y' taking values in $[\ell_1]$ and $[\ell_2]$.[4] Denote the joint distribution of (X', Y') by $p_{X',Y'}$. As usual we denote the marginals as $p_{X'}$ and $p_{Y'}$ (i.e., $p_{X'}(x) = \sum_{y\in[\ell_2]} p_{X',Y'}(x,y)$ and similarly for $p_{Y'}$). We are interested in finding an unbiased estimate of the following expression:

$$\|p_{X',Y'} - p_{X'}p_{Y'}\|_2^2 = \sum_{x\in[\ell_1],y\in[\ell_2]} \big(p_{X',Y'}(x,y) - p_{X'}(x)p_{Y'}(y)\big)^2. \tag{5.1}$$

The above expression is nothing but the L_2^2 distance between $p_{X',Y'}$ and the product of the marginals $p_{X'}p_{Y'}$. In order for us to unbiasedly estimate this quantity, we will use a U-statistic, and at least 4 observations. Before we define the U-statistic, let us define its kernel. Let $i, j \in [\sigma]$ be two observations. Define

$$\phi_{ij}(xy) = \mathbb{1}\big(X_i' = x, Y_i' = y\big) - \mathbb{1}\big(X_i' = x\big)\mathbb{1}\big(Y_j' = y\big). \tag{5.2}$$

Next, take 4 distinct observations $i, j, k, l \in [\sigma]$, and define the kernel function

$$h_{ijkl} = \frac{1}{4!}\sum_{\pi\in[4!]}\sum_{x\in[\ell_1],y\in[\ell_2]} \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy),$$

where π is a permutation of i, j, k, l. Clearly, since $i, j, k, l \in [\sigma]$ are distinct, the above is an unbiased estimate of (5.1). Next, we construct the U-statistic

$$U(\mathcal{D}) := \frac{1}{\binom{\sigma}{4}} \sum_{i<j<k<l:(i,j,k,l)\in[\sigma]} h_{ijkl}, \tag{5.3}$$

where we denoted $\mathcal{D} = \{(X_1', Y_1'), \ldots, (X_\sigma', Y_\sigma')\}$. The U-statistic (5.3) is an unbiased estimate of the L_2^2 distance in (5.1). It is not obvious that this estimator is the same as the one defined

[4] As in the lower bound, it is not crucial that the supports of X' and Y' are $[\ell_1]$ and $[\ell_2]$. We focus on this case simply for the sake of clarity.

in equation (18) of Canonne et al. [12]. However, using Proposition 4.2 of [12] and the fact that the U-statistic in (5.3) is a symmetric estimator, one can deduce that the two estimators must coincide.

In order to analyze our hypothesis test, we will appropriately bound the mean and variance of our test statistic under the null and under the alternative. Since our test is based on the U-statistic in (5.3), we will need to bound its variance. In principle, one can directly reuse the bound on the variance of the U-statistic in (5.3) given in [12]. Since the original derivation of this bound is complicated, we give a novel derivation starting from first principles, building on the extensive theory for U-statistics. We have the following result.

LEMMA 5.1 (Variance upper bound). *There exists some absolute constant C such that*

$$\operatorname{Var}\big[U(\mathcal{D})\big] \le C\bigg(\frac{\mathbb{E}[U(\mathcal{D})]\max(\|p_{X',Y'}\|_2, \|p_{X'}p_{Y'}\|_2)}{\sigma} + \frac{\max(\|p_{X',Y'}\|_2^2, \|p_{X'}p_{Y'}\|_2^2)}{\sigma^2}\bigg).$$

Now that we have defined the statistic U and have bounded its variance, we are ready to introduce our test statistic. Before that we include a randomization device in the test:

Draw $N \sim \operatorname{Poi}(\frac{n}{2})$. If $N > n$ accept the null hypothesis. If $N \le n$, take arbitrary N out of the n samples and work with them. The next step is to discretize the variable Z into d bins of equal size. Denote those bins with $\{C_1, \dots, C_d\}$, so that $\bigcup_{i\in[d]} C_i = [0,1]$, and each C_i is an interval of length $\frac{1}{d}$. Next, construct the datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i \in C_m, i \in [N]\}$. Let $\sigma_m = |\mathcal{D}_m|$ be the sample size in each set $\mathcal{D}_m$, so that $\sum_{m\in[d]} \sigma_m = N$. For bins $\mathcal{D}_m$ with at least $\sigma_m \ge 4$ observations, let for brevity $U_m = U(\mathcal{D}_m)$. Each U_m can be thought of as a local test of independence within the bin C_m—if the value of U_m is close to 0 then intuitively independence holds within that bin, while if the value of U_m is large, independence is potentially violated within that bin. In order to combine these different statistics, we follow Canonne et al. [12] and consider the following test statistic:

$$T = \sum_{m\in[d]} \mathbb{1}(\sigma_m \ge 4)\sigma_m U_m. \tag{5.4}$$

We will prove that under the null hypothesis the value of T is likely to be below a threshold τ (to be specified), while under the alternative hypothesis T will likely exceed the value τ. Define the test

$$\psi_\tau(\mathcal{D}_N) = \mathbb{1}(T \ge \tau),$$

where $\mathcal{D}_N = \{(X_1, Y_1, Z_1), \dots, (X_N, Y_N, Z_N)\}$. Recall the definitions of the null Lipschitzness classes $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1],\chi^2}(L)$, $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ and the alternative Lipschitzness classes $\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L)$ (see Definitions 2.1 and 2.2 in Section 2.3). We are now ready to state the main result of this section.

THEOREM 5.2 (Finite discrete X, Y upper bound). *Set $d = \lceil n^{2/5} \rceil$ and let $\tau = \zeta n^{1/5}$ for a sufficiently large absolute constant ζ (depending on L). Finally, suppose that $\varepsilon \ge cn^{-2/5}$, for a sufficiently large constant c (depending on ζ, L, ℓ_1, ℓ_2). Then we have that*

$$\sup_{p\in\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)\cup\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)\cup\mathcal{P}'_{0,[0,1],\chi^2}(L)} \mathbb{E}_p\big[\psi_\tau(\mathcal{D}_N)\big] \le \frac{1}{10},$$

$$\sup_{p\in\{p\in\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L):\inf_{q\in\mathcal{P}'_{0,[0,1]}}\|p-q\|_1\ge\varepsilon\}} \mathbb{E}_p\big[1-\psi_\tau(\mathcal{D}_N)\big] \le \frac{1}{10} + \exp(-n/8).$$

REMARKS.

- In the above theorem, the constants $\frac{1}{10}$ are arbitrary and can be made smaller (or larger) by appropriately adjusting the constants ζ and c. In the case when ℓ_1 and ℓ_2 are of constant order, the above test is optimal, in the sense that the critical radius rate $n^{-2/5}$ matches the lower bound given in Theorem 4.1.
- When ℓ_1 and ℓ_2 are allowed to scale with n, the test no longer results in the correct order for the critical radius (in particular, we can no longer treat the quantity c as a constant and its dependence on ℓ_1 and ℓ_2 is not optimal). In the next section, we provide more sophisticated test which is capable of matching the bound proved in Theorem 4.1 for some regimes of ℓ_1 and ℓ_2.
- In order to show that our test has high power for sufficiently large ε_n, we follow a classical strategy of upper bounding the variance of our test statistic under the null and alternative, upper bounding its expectation under null, and lower bounding its expectation under the alternative. These bounds together with a careful choice of the threshold τ, and an application of Chebyshev's inequality, are used to characterize the power of our proposed test. We detail these calculations in Appendix C of the supplementary material.
- A recurring complication, one that we need to address in the analysis of our tests in both the discrete and continuous X, Y setting is that our test statistic does not have expectation zero under the null. This is in sharp contrast to typical tests for goodness-of-fit and two-sample tests (for instance, those analyzed in [2, 4, 5, 14, 37]). In more detail, under the null, the binning operation used to discretize the Z variable, moves us off the manifold of conditionally independent distributions (i.e., the discretized distribution need not satisfy conditional independence even if the original distribution does).

  Exploiting the Lipschitzness assumptions in Definition 2.1, we can argue that under the null, for sufficiently small bins, we do not move too far from the collection of conditionally independent distributions (say in the total variation sense). A naive reduction would yield an imprecise null hypothesis testing problem of attempting to distinguish distributions, which are near-conditionally independent from those which are relatively far from conditionally independent. This imprecise null testing problem is however statistically challenging [38], and this naive reduction fails to yield the optimal rates described in our upper bounds.

  Instead, avoiding this indirect reduction, we take a more direct approach of uniformly upper bounding the expectation of our test statistic under the null. By directly using the Lipschitzness assumptions, and the factorization structure of distributions under the null, we are able to obtain tighter bounds on the expected value of our test statistic under the null. This in turn yields near-optimal upper bounds on the critical radius.

5.2. *Upper bound with scaling discrete X and Y.* In this section, we present a more sophisticated test procedure which is capable of matching the bound of Theorem 4.1 for some regimes of the sizes of the supports of X and Y — ℓ_1 and ℓ_2. In contrast to the previous section, we now no longer assume that $\ell_1, \ell_2 = O(1)$. We note that throughout this section, without any loss of generality, we focus on the case when $\sqrt{\ell_1 \ell_2}/n \lesssim 1$. When this condition is not satisfied, the lower bound in Theorem 4.1 shows that the critical radius must be at least a constant, and in this regime upper bounds are trivial. Since we only characterize the critical radius up to constants, when we choose the separation between the null and alternate ε to be a sufficiently large constant (say 2), there are no longer any distributions in the alternate, and the CI testing problem is trivial.

The key idea of this section is to use a weighted U-statistic in place of the (unweighted) U-statistic from Section 5.1. This weighting is sometimes referred to as "flattening"; see, for

example, [12, 14]. A careful choice of the weighting yields a U-statistic with smaller variance (see Lemma 5.4), and the resulting test has higher power.

To describe the weighting, consider again the same scenario as in Section 5.1. Suppose we observe $\sigma \geq 4$ samples of two discrete covariates (X', Y') supported on $[\ell_1] \times [\ell_2]$. Let $\mathcal{D} = \{(X'_1, Y'_1), \ldots, (X'_\sigma, Y'_\sigma)\}$ and $p_{X',Y'}$ be the distribution of (X', Y'). By losing at most three samples, we may assume that $\sigma = 4 + 4t$ for some $t \in \mathbb{N}$. Define $t_1 := \min(t, \ell_1)$ and $t_2 := \min(t, \ell_2)$. Next we split $\mathcal{D}$ into three datasets of sizes t_1, t_2 and $2t + 4$, respectively: $\mathcal{D}_{X'} = \{X'_i : i \in [t_1]\}$, $\mathcal{D}_{Y'} = \{Y'_i : t_1 + 1 \leq i \leq t_1 + t_2\}$ and $\mathcal{D}_{X',Y'} = \{(X'_i, Y'_i) : 2t + 1 \leq i \leq \sigma\}$. The idea behind defining those three datasets is that the first two datasets—$\mathcal{D}_{X'}$ and $\mathcal{D}_{Y'}$, will be used to calculate weights, while the last dataset $\mathcal{D}_{X',Y'}$, which has at least 4 observations, will be used to calculate the U-statistic. Construct the integers

$$1 + a_{xy} = (1 + a_x)(1 + a'_y),$$

where a_x are the number of occurrences of x in $\mathcal{D}_{X'}$ and a'_y is the number of occurrences of y in $\mathcal{D}_{Y'}$.

Next, take 4 distinct observations indexed by i, j, k, l from the dataset $\mathcal{D}_{X',Y'}$, and define the (weighted) kernel function

$$h^{\boldsymbol{a}}_{ijkl} = \frac{1}{4!} \sum_{\pi \in [4!]} \sum_{x \in [\ell_1], y \in [\ell_2]} \frac{\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)}{1 + a_{xy}},$$

where π is a permutation of i, j, k, l and recall the definition of $\phi_{ij}(xy)$ (5.2). Here the super-indexing with $\boldsymbol{a}$ of $h^{\boldsymbol{a}}_{ijkl}$, indicates that the statistic is weighted by the numbers $1 + a_{xy}$ for $x \in [\ell_1]$, $y \in [\ell_2]$. Notice that the idea of this weighting is similar to the weighting in a Pearson's χ^2 test of independence. Indeed the quantity a_{xy} is in expectation proportional to the product $p_{X'}(x)p_{Y'}(y)$. On the other hand, the expression $\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)$ is unbiased for $(p_{X',Y'}(x, y) - p_{X'}(x)p_{Y'}(y))^2$. Next, to reduce the variance of $h^{\boldsymbol{a}}_{ijkl}$, we construct the (weighted) U-statistic

$$U_W(\mathcal{D}) := \frac{1}{\binom{2t+4}{4}} \sum_{i<j<k<l : (i,j,k,l) \in \mathcal{D}_{X',Y'}} h^{\boldsymbol{a}}_{ijkl}, \tag{5.5}$$

where we abused notation slightly for $(i, j, k, l) \in \mathcal{D}_{X',Y'}$ to mean taking four observations from the dataset $\mathcal{D}_{X',Y'}$. For convenience of notation, we now give a definition from [14].

DEFINITION 5.3 (Split distribution). Given a discrete distribution p over $[d_1] \times [d_2]$ and a multiset S of elements of $[d_1] \times [d_2]$ we now define the split distribution p_S. Let $b_{xy} = \sum_{(x',y') \in S} \mathbb{1}((x, y) = (x', y'))$. Thus $\sum_{(x,y) \in [d_1] \times [d_2]} 1 + b_{xy} = d_1 d_2 + |S|$. Define the set $B_S = \{(x, y, i) | (x, y) \in [d_1] \times [d_2], 1 \leq i \leq 1 + b_{xy}\}$. The split distribution p_S is supported on B_S and is obtained by sampling (x, y) from p and i uniformly from the set $[1 + b_{xy}]$.

Given S and b_{xy} as in Definition 5.3, for any two discrete distributions p and q over $[d_1] \times [d_2]$ it follows that

$$\|p_S - q_S\|_2^2 = \sum_{(x,y) \in [d_1] \times [d_2]} \frac{(p(x, y) - q(x, y))^2}{1 + b_{xy}}.$$

Similarly, for the split distribution p_S we have that

$$\|p_S\|_2^2 = \sum_{(x,y) \in [d_1] \times [d_2]} \frac{p^2(x, y)}{1 + b_{xy}}.$$

Construct a multiset A by adding a_{xy} occurrences of the pair (x, y) to A. Using this notation, it now follows that

$$\begin{aligned}\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}] &= \|p_{X',Y',A} - p^{\Pi}_{X',Y',A}\|_2^2 \\ &= \sum_{(x,y)\in[\ell_1]\times[\ell_2]} \frac{(p_{X',Y'}(x,y) - p_{X'}(x)p_{Y'}(y))^2}{1 + a_{xy}},\end{aligned}$$

where $p_{X',Y',A}$ is the A-split distribution $p_{X',Y'}$, and $p^{\Pi}_{X',Y',A}$ is the A-split distribution $p^{\Pi}_{X',Y'}$ where $p^{\Pi}_{X',Y'} = p_{X'}p_{Y'}$. We will now show an analogous variance bound to the one in Lemma 5.1. We have the following.

LEMMA 5.4 (Variance upper bound). *For some absolute constant C, the following holds*:

$$\begin{aligned}&\mathrm{Var}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}] \\ &\le C\Bigg(\frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]\|p^{\Pi}_{X',Y',A}\|_2}{\sigma} \\ &\quad + \frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]^{3/2}}{\sigma} + \frac{\|p^{\Pi}_{X',Y',A}\|_2^2}{\sigma^2} + \frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]}{\sigma^2}\Bigg).\end{aligned}$$

In comparing to the result of Lemma 5.1, we see roughly that the variance bound now depends on the (typically much smaller) L_2-norm of the flattened or split distribution $p^{\Pi}_{X',Y',A}$, instead of the L_2 norm of the original distribution $p_{X',Y'}$. As emphasized in [12, 14], this variance reduction achieved through flattening is critical for designing minimax optimal tests (particularly when ℓ_1 and ℓ_2 are allowed to grow with the sample-size n).

Now we are ready to define our test statistic. As before, the first step is to draw a random sample size $N \sim \mathrm{Poi}(\frac{n}{2})$ and take N subsamples of the n observations, with the convention that if $N > n$ we accept the null hypothesis. Next, bin the support of the variable Z into d bins of equal size. Denote those bins with $\{C_1, \ldots, C_d\}$, so that $\bigcup_{i\in[d]} C_i = [0, 1]$, and each C_i is an interval of length $\frac{1}{d}$. Construct the datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i \in C_m, i \in [N]\}$. Let $\sigma_m = |\mathcal{D}_m|$ be the sample size in each set $\mathcal{D}_m$, so that $\sum_{m\in[d]} \sigma_m = N$. Recall that each set $\mathcal{D}_m$ will be further separated into three sets $\mathcal{D}_{m,X}$, $\mathcal{D}_{m,Y}$ and $\mathcal{D}_{m,X,Y}$, the first two of which are used for calculating weights, while the last one is used for the calculation of the weighted U-statistic. For bins $\mathcal{D}_m$ with at least $\sigma_m \ge 4$ observations, let for brevity $U_m = U_W(\mathcal{D}_m)$. We now combine these different independence testing statistics into one CI testing statistic as follows. Let

$$T = \sum_{m\in[d]} \mathbb{1}(\sigma_m \ge 4)\sigma_m\omega_m U_m, \tag{5.6}$$

where $\omega_m = \sqrt{\min(\sigma_m, \ell_1)\min(\sigma_m, \ell_2)}$ is a weighting factor, which further weights the statistics U_m. The presence of ω_m is necessitated by the weighting of the U-statistic (5.5). In order to show that the test based on the statistic T has high power (and low type 1 error) we will prove that under the null hypothesis the value of T is likely to be below a threshold τ (to be specified), while under the alternative hypothesis T will likely exceed the value τ. Define the test

$$\psi_\tau(\mathcal{D}_N) = \mathbb{1}(T \ge \tau), \tag{5.7}$$

where $\mathcal{D}_N = \{(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)\}$. We have the following result.

THEOREM 5.5 (Scaling discrete X, Y upper bound). *Set* $d = \lceil \frac{n^{2/5}}{(\ell_1\ell_2)^{1/5}} \rceil$ *and set the threshold* $\tau = \sqrt{\zeta d}$ *for a sufficiently large absolute constant* ζ *(depending on* L*). Suppose that* $\ell_1 \geq \ell_2$ *satisfy the condition that* $d\ell_1 \lesssim n$*. Then when* $\varepsilon \geq c\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$*, for a sufficiently large absolute constant* c *(depending on* ζ, L*), we have that*

$$\sup_{p \in \mathcal{P}'_{0,[0,1],\chi^2}(L)} \mathbb{E}_p\big[\psi_\tau(\mathcal{D}_k)\big] \leq \frac{1}{10},$$

$$\sup_{p \in \{p \in \mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L) : \inf_{q \in \mathcal{P}'_{0,[0,1]}} \|p-q\|_1 \geq \varepsilon\}} \mathbb{E}_p\big[1 - \psi_\tau(\mathcal{D}_k)\big] \leq \frac{1}{10} + \exp(-n/8).$$

REMARKS.

- Some remarks regarding this result are in order. First, when $d\ell_1 \lesssim n$, the bound on the critical radius we obtain matches the information-theoretic limit derived in Theorem 4.1. An important special case (that we will use in our tests in the continuous X and Y setting) when this condition is automatically implied is when $\ell_1 \asymp \ell_2$.

  To see this, observe that when $\ell_1 \asymp \ell_2$ we have that $d\ell_1 \lesssim n$ is equivalent to $(\frac{n}{\ell_1})^{2/5} \lesssim \frac{n}{\ell_1}$ (for our choice of d) which is implied by the condition that $\frac{\ell_1}{n} \lesssim 1$. When this latter condition is not satisfied, the lower bound on the critical radius in Theorem 4.1 is a universal constant (and the upper bound is trivial).

  We also note in passing that for our choice of d, the condition that $d\ell_1 \lesssim n$ is equivalent to the condition that $\frac{\ell_1^4}{\ell_2} \lesssim n^3$, which yields the claim in Section 1.2 that our test is minimax optimal when $\frac{\ell_1^4}{\ell_2} \lesssim n^3$.
- In contrast to Theorem 5.2, here we choose the null set of distributions as $\mathcal{P}'_{0,[0,1],\chi^2}(L)$. As we discussed following Theorem 5.2, one of the key difficulties is to characterize the effect of discretization of the Z variable, in order to upper bound the expectation of our test statistic under the null, over the appropriate Lipschitzness class. When ℓ_1 and ℓ_2 are allowed to scale, we show an upper bound on this expectation in terms of the χ^2-divergence between the discretized null distribution and the product of its marginals (see equation (C.23) in Appendix C) of the supplement. We in turn show that this discretization error due to binning is appropriately small when the null distribution satisfies the χ^2 Lipschitzness condition, that is, belongs to $\mathcal{P}'_{0,[0,1],\chi^2}(L)$.
- As we detail further in Appendix C of the supplement, when the condition that $d\ell_1 \lesssim n$ is not satisfied we still provide upper bounds on the critical radius but these upper bounds do not match the lower bound in Theorem 4.1. As we discuss further in Section 8, we believe that sharpening either the lower or upper bound is challenging, requiring substantially different ideas, and we defer this to future work.
- From a technical standpoint, analyzing the power of the test statistic in (5.6) is substantially more involved than the analysis of its fixed ℓ_1, ℓ_2 counterpart in (5.4). Several complications are introduced in ensuring that the flattening weights (the terms a_{xy} in the definition of our U-statistic in (5.5)) are well behaved. In a classical fixed dimensional setup (where ℓ_1, ℓ_2 and the number of bins d are all held fixed), it would be relatively straightforward to argue that the flattening weights concentrate tightly around their expected values. In the high-dimensional setting that we consider, these weights can have high variance and substantial work is needed to tightly bound the mean and variance of our test statistic.

  This also highlights an important difference from the goodness-of-fit problem considered in [5, 10, 37]. In the goodness-of-fit problem, where we test fit of the data to a *known*

distribution p_0 the corresponding weights in the Pearson χ^2 statistic are fixed and known to the statistician. In conditional independence testing, these weights are estimated from data.

5.3. *Upper bound in the continuous case.* In this section, consider testing for CI when (X, Y, Z) are supported on $[0, 1]^3$ and have a distribution which is absolutely continuously with respect to the Lebesgue measure. In view of the notation in Section 4.2, this is equivalent to assuming that $p_{X,Y,Z} \in \mathcal{E}_{0,[0,1]^3}$. We begin our discussion with formally describing the test.

The testing strategy is related to the test described in Section 5.2. First, draw $N \sim \mathrm{Poi}(\frac{n}{2})$, and take arbitrary N out of the n observations in the case when $N \leq n$, and accept the null hypothesis if $N > n$. Next, we bin the support $[0, 1]$ with bins $\{C_1, C_2, \ldots, C_d\}$, where the sizes of those bins are equal and $\bigcup_{i\in[d]} C_i = [0, 1]$. These bins will be used to discretize Z. In addition, we create a second rougher (in the case $s \geq 1$) partition of $[0, 1]$ into $d' := \lceil d^{1/s} \rceil$ intervals $\bigcup_{i\in[d']} C'_i = [0, 1]$. These second bins will be used to discretize X and Y. Specifically, we use these two sets of bins to discretize the observations $\mathcal{D}_N = \{(X_i, Y_i, Z_i)\}_{i\in[N]}$ as follows. First, define the discretization function $g : [0, 1] \mapsto [d']$ by $g(x) = j$ iff $x \in C'_j$. Next, consider the set of observations $\mathcal{D}'_N = \{(g(X_i), g(Y_j), Z_i)\}_{i\in[N]}$. We can now use the test defined in (5.7): $\psi_\tau(\mathcal{D}'_N)$ with an appropriately selected threshold τ and the bins $\{C_1, C_2, \ldots, C_d\}$ to discretize Z with in order to test for CI. We have the following result.

THEOREM 5.6 (Continuous X, Y, Z upper bound). *Set $d = \lceil n^{2s/(5s+2)} \rceil$ and set the threshold $\tau = \sqrt{\zeta d}$ for a sufficiently large ζ (depending on L). Let $\mathcal{H}_0(s) = \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L) \cup \mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ when $s \geq 1$ and $\mathcal{H}_0(s) = \mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ when $s < 1$. Then, for a sufficiently large absolute constant c (depending on ζ, L), when $\varepsilon \geq cn^{-2s/(5s+2)}$, we have that*

$$\sup_{p\in\mathcal{H}_0(s)} \mathbb{E}_p\big[\psi_\tau(\mathcal{D}'_k)\big] \leq \frac{1}{10},$$

$$\sup_{p\in\{p\in\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,s):\inf_{q\in\mathcal{P}_{0,[0,1]^3}} \|p-q\|_1\geq\varepsilon\}} \mathbb{E}_p\big[1 - \psi_\tau(\mathcal{D}'_k)\big] \leq \frac{1}{10} + \exp(-n/8).$$

REMARKS.

- Theorem 5.6 shows that the test $\psi_\tau(\mathcal{D}'_N)$ matches the lower bound derived in Theorem 4.2, showing that under appropriate Lipschitzness conditions our test is a minimax optimal nonparametric test for conditional independence.
- We note that in this setting, a careful analysis of the expectation of our statistic under the null shows that the null set of distributions can be taken as $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L) \cup \mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ which is a larger set of distributions in comparison to that of Theorem 5.5.
- Finally, the analysis in the continuous setting builds extensively on our analysis for the test in (5.7). However, as we detail in Appendix C of the supplement (see Lemmas C.16, C.17 and C.18), careful analysis is needed to show that the additional discretization error of the X and Y variables does not change the mean and variance of our test statistic too much (under both the null and alternative).

6. Investigating Lipschitzness conditions. In our upper and lower bounds, in order to tractably test conditional independence in the nonparametric setting, we impose various Lipschitzness conditions on the distributions under consideration. In order to build further intuition for these conditions, in this section we derive several inclusions which relate the Lipschitzness classes defined in Sections 4.1 and 4.2. We then give examples of natural classes of distributions which satisfy our various Lipschitzness conditions.

6.1. *Relationships between the Lipschitzness classes.* Recall the definitions of the null Lipschitzness classes in Definition 2.1. Our first result shows that the class of Hölder smooth distributions contains the class of TV smooth distributions and χ^2 smooth distributions.

LEMMA 6.1. *We have the following inclusions*:

$$\mathcal{P}'_{0,[0,1],\chi^2}(L) \subseteq \mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L), \tag{6.1}$$

$$\mathcal{P}'_{0,[0,1],\mathrm{TV}}(\sqrt{L}) \subseteq \mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L). \tag{6.2}$$

PROOF. To prove this result, we state a simple but useful direct corollary of the Cauchy–Schwarz inequality, which is also known in the literature as the T2 Lemma.

LEMMA 6.2 (T2 Lemma). *For positive reals* $\{u_i\}_{i\in[k]}$ *and* $\{v_i\}_{i\in[k]}$, *we have*

$$\frac{(\sum_{i\in[k]} u_i)^2}{\sum_{i\in[k]} v_i} \le \sum_{i\in[k]} \frac{u_i^2}{v_i}.$$

By the T2 Lemma, it is simple to see that

$$\begin{aligned} d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) &= \sum_x \frac{(p_{X|Z}(x|z) - p_{X|Z}(x|z'))^2}{p_{X|Z}(x|z')} \\ &\ge \frac{(\sum_x |p_{X|Z}(x|z) - p_{X|Z}(x|z')|)^2}{\sum_x p_{X|Z}(x|z')} = \|p_{X|Z=z} - p_{X|Z=z'}\|_1^2. \end{aligned}$$

Hence we have that

$$d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) \le L|z - z'| \quad \Longrightarrow \quad \|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le \sqrt{L|z - z'|},$$

and, therefore, we obtain the inclusion in (6.1).

To derive the second inclusion, note that when $z, z' \in [0, 1]$ we have $|z - z'| \le \sqrt{|z - z'|}$ and, therefore,

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le \sqrt{L}|z - z'| \quad \Longrightarrow \quad \|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le \sqrt{L|z - z'|}. \quad \square$$

In Definition 2.1, we assume that the marginal distributions of X and Y conditional on Z are each smooth. Our next result shows that up to a factor of 2 this is equivalent to assuming TV Lipschitzness on the joint distribution of (X, Y) conditional on Z.

LEMMA 6.3. *Define the class of distributions* $\mathcal{P}''_{0,[0,1],\mathrm{TV}}(L) \subset \mathcal{P}'_{0,[0,1]}$ *such that for each* $p_{X,Y,Z} \in \mathcal{P}''_{0,[0,1],\mathrm{TV}}(L)$ *and all* $z, z' \in [0, 1]$:

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \le L|z - z'|.$$

Then

$$\mathcal{P}''_{0,[0,1],\mathrm{TV}}(L) \subseteq \mathcal{P}'_{0,[0,1],\mathrm{TV}}(L) \quad \textit{and} \quad \mathcal{P}'_{0,[0,1],\mathrm{TV}}(L) \subseteq \mathcal{P}''_{0,[0,1],\mathrm{TV}}(2L).$$

PROOF. The first inclusion is a consequence of the triangle inequality:

$$\max\big(\|p_{X|Z=z} - p_{X|Z=z'}\|_1, \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1\big) \le \|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1.$$

To obtain the second inclusion, we note that $p_{X,Y|Z=z} = p_{X|Z=z}p_{Y|Z=z}$ and $p_{X,Y|Z=z'} = p_{X|Z=z'}p_{Y|Z=z'}$, and that d_{TV} is subadditive on product distributions [37] so that

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \le \|p_{X|Z=z} - p_{X|Z=z'}\|_1 + \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1.$$

□

Similar statements to Lemma 6.3 hold for the classes $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ and $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}$. For brevity, we do not state them here. We now state another similar result for the Lipschitzness class $\mathcal{P}'_{0,[0,1],\chi^2}$.

LEMMA 6.4. *Define the class of distributions* $\mathcal{P}''_{0,[0,1],\chi^2}(L) \subset \mathcal{P}'_{0,[0,1]}$, *such that for each* $p_{X,Y,Z} \in \mathcal{P}''_{0,[0,1],\chi^2}(L)$ *and all* $z, z' \in [0,1]$ *we have*

$$d_{\chi^2}(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \le L|z - z'|.$$

Then

$$\mathcal{P}''_{0,[0,1],\chi^2}(L) \subseteq \mathcal{P}'_{0,[0,1],\chi^2}(L) \quad \text{and} \quad \mathcal{P}'_{0,[0,1],\chi^2}(L) \subseteq \mathcal{P}''_{0,[0,1],\chi^2}(2L + L^2).$$

PROOF. We start by showing the first inclusion. Note that by the T2 lemma

$$\begin{aligned} d_{\chi^2}(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) &= \sum_{x,y} \frac{p^2_{X,Y|Z}(x, y|z)}{p_{X,Y|Z}(x, y|z')} - 1 \\ &\ge \sum_{x} \frac{(\sum_y p_{X,Y|Z}(x, y|z))^2}{\sum_y p_{X,Y|Z}(x, y|z')} - 1 = d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}). \end{aligned}$$

By symmetry, it also follows that $d_{\chi^2}(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \ge d_{\chi^2}(p_{Y|Z=z}, p_{Y|Z=z'})$ which shows the first inclusion. For the second inclusion using the fact that $p_{X,Y|Z=z} = p_{X|Z=z} p_{Y|Z=z}$ and $p_{X,Y|Z=z'} = p_{X|Z=z'} p_{Y|Z=z'}$, it is simple to verify that

$$\begin{aligned} d_{\chi^2}(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) &= d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) + d_{\chi^2}(p_{Y|Z=z}, p_{Y|Z=z'}) \\ &\quad + d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) d_{\chi^2}(p_{Y|Z=z}, p_{Y|Z=z'}), \end{aligned}$$

which yields the desired conclusion by noting that this expression in turn is smaller than $2L|z - z'| + L^2|z - z'|^2 \le 2L|z - z'| + L^2|z - z'|$, when $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(L)$. □

A similar result also holds for the set $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ but once again we do not state the result here for brevity.

6.2. *Distribution families in our Lipschitzness classes*. Next, we give some concrete examples of distributions which belong to the different Lipschitzness classes. We begin by showing that smoothness of the log-conditional density is sufficient to ensure that the distribution belongs to both the TV and χ^2 Lipschitzness classes. We then show that a broad subset of exponential family distributions have a smooth log-conditional distributions.

LEMMA 6.5. *Take a distribution* $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1]}$. *Suppose that the functions* $\log p_{X|Z}(x|z)$, $\log p_{Y|Z}(y|z)$ *are L-Lipschitz in z for all values of x and y. Then the distribution* $p_{X,Y,Z}$ *belongs to* $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(e^L - 1) \cap \mathcal{P}'_{0,[0,1],\chi^2}(e^L - 1)$.

PROOF. We begin by showing that $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(e^L - 1)$. Note that

$$\sum_x \frac{p^2_{X|Z}(x|z)}{p_{X|Z}(x|z')} - 1 = \sum_x \left(\frac{p_{X|Z}(x|z)}{p_{X|Z}(x|z')} - 1 \right) p_{X|Z}(x|z).$$

As a consequence it suffices to show that

$$\frac{p_{X|Z}(x|z)}{p_{X|Z}(x|z')} - 1 \le (e^L - 1)|z - z'|,$$

for all $z, z' \in [0,1]$ and all x (and the analogous claim for $p_{Y|Z}$) in order to conclude that $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(e^L - 1)$. Since $\log p_{X|Z}(x|z)$ is L-Lipschitz in z, it follows that for values of $|z - z'| \le 1$:

$$\begin{aligned}\frac{p_{X|Z}(x|z)}{p_{X|Z}(x|z')} - 1 &\le \exp\bigl(L|z - z'|\bigr) - 1 = L|z - z'| + \sum_{k\ge 2}\bigl(L|z - z'|\bigr)^k/k! \\ &\le L|z - z'| + L|z - z'| \sum_{k\ge 2} L^{k-1}/k! = L|z - z'| + L|z - z'|(e^L - 1 - L)/L \\ &= (e^L - 1)|z - z'|.\end{aligned}$$

This, together with an identical claim for $p_{Y|Z}$, proves the first claim, that is, $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(e^L - 1)$. To establish the second claim, note that

$$\begin{aligned}&\sum_x \bigl|p_{X|Z}(x|z) - p_{X|Z}\bigl(x|z'\bigr)\bigr| \\ &\quad= \sum_x \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1\right) \min\bigl(p_{X|Z}(x|z), p_{X|Z}\bigl(x|z'\bigr)\bigr) \\ &\quad\le \sum_x \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1\right) p_{X|Z}(x|z).\end{aligned}$$

Hence the same proof as above applies. This completes the proof. □

We now state several similar and related results without proof, noting that their proofs are nearly identical to the proof of Lemma 6.5.

LEMMA 6.6. *Take a distribution $p_{X,Y,Z} \in \mathcal{Q}'_{0,[0,1]}$. Suppose that the function $\log p_{X,Y|Z}(x, y|z)$ is L-Lipschitz in z for all values of x and y. Then $p_{X,Y,Z} \in \mathcal{Q}'_{0,[0,1],\mathrm{TV}}(e^L - 1)$.*

LEMMA 6.7. *Let $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3}$. Suppose that the functions $\log p_{X|Z}(x|z)$, $\log p_{Y|Z}(y|z)$ are L-Lipschitz in z for all values of x and y. Then the distribution $p_{X,Y,Z}$ also belongs to $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(e^L - 1) \cap \mathcal{P}_{0,[0,1]^3,\chi^2}(e^L - 1)$.*

LEMMA 6.8. *Let $p_{X,Y,Z} \in \mathcal{Q}_{0,[0,1]^3}$. Suppose that the function $\log p_{X,Y|Z}(x, y|z)$ is L-Lipschitz in z for all x and y, and further that the function $p_{X,Y|Z}(x, y|z)$ is jointly C-Lipschitz in x and y, for all z, that is,*

$$\bigl|p_{X,Y|Z}(x, y|z) - p_{X,Y|Z}\bigl(x', y'|z\bigr)\bigr| \le C\bigl(|x - x'| + |y - y'|\bigr). \tag{6.3}$$

Then $p_{X,Y,Z} \in \mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}((e^L - 1) \vee \sqrt{2}C, 1)$.

Lemmas 6.6 and 6.8 are regarding the continuous case, and are therefore slightly different from Lemmas 6.5 and 6.7. Hence for completeness, we give the proof of Lemma 6.8 in the supplement. Roughly, these results taken together show that Lipschitzness of the log conditional density imply the various Lipschitzness conditions we impose. Our next set of results shows that a broad class of natural exponential family type distributions, in fact, have smooth log conditional densities.

LEMMA 6.9. *Consider the density* $p_{W|Z}(w|z) \propto \exp(g(w,z))$, *where* $g(w,z)$ *is an* L*-Lipschitz function in* $z \in [0,1]$ *for all values of* w. *Then the function* $\log p_{W|Z}(w|z)$ *is* $2L$*-Lipschitz.*

We note that in the lemma above, W can be taken as a vector of any dimension so the lemma applies to $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$ as well as to $p_{X,Y|Z}(x,y|z)$. The lemma also applies in both discrete W as well as continuous W cases.

PROOF. We consider the differences

$$\log \frac{\exp(g(w,z))}{\sum_w \exp(g(w,z))} - \log \frac{\exp(g(w,z'))}{\sum_w \exp(g(w,z'))}$$
$$\leq \big(g(w,z) - g\big(w,z'\big)\big) - \log \frac{\sum_w \exp(g(w,z))}{\sum_w \exp(g(w,z'))}.$$

Next, we use Jensen's inequality and the fact that—log is a convex function to show that

$$\begin{aligned} -\log \frac{\sum_w \exp(g(w,z))}{\sum_w \exp(g(w,z'))} &= -\log \frac{\sum_w \exp(g(w,z')) \exp(g(w,z) - g(w,z'))}{\sum_w \exp(g(w,z'))} \\ &\leq \sum_w \frac{\exp(g(w,z'))}{\sum_w \exp(g(w,z'))} \big(g\big(w,z'\big) - g(w,z)\big) \\ &\leq \sum_w \frac{\exp(g(w,z'))}{\sum_w \exp(g(w,z'))} \big|g\big(w,z'\big) - g(w,z)\big| \\ &\leq L\big|z - z'\big|. \end{aligned}$$

Putting things together, we get

$$\log \frac{\exp(g(w,z))}{\sum_w \exp(g(w,z))} - \log \frac{\exp(g(w,z'))}{\sum_w \exp(g(w,z'))}$$
$$\leq \big|g(w,z) - g\big(w,z'\big)\big| + L\big|z - z'\big| \leq 2L\big|z - z'\big|.$$

Reversing the roles of z and z' we conclude. The same proof goes through in the continuous case, where summations have to be substituted with integrals. □

Finally, in the continuous case we provide a family of distributions for which $\log p_{X,Y|Z}(x,y|z)$ is L-Lipschitz in z and $p_{X,Y|Z}(x,y|z)$ is C-Lipschitz in x and y as required in Lemma 6.8.

LEMMA 6.10. *Suppose that* $g(x,y,z) : [0,1]^3 \mapsto [-M, M]$ *is a bounded* L*-Lipschitz function, that is,* $|g(x,y,z) - g(x',y',z')| \leq L(|x - x'| + |y - y'| + |z - z'|)$. *Take* $p_{X,Y,Z}(x,y,z) \propto \exp(g(x,y,z))$. *Then*

$$p_{X,Y|Z}(x,y|z) = \frac{\exp(g(x,y,z))}{\int_{[0,1]^2} \exp(g(x,y,z))\,dx\,dy},$$

satisfies (6.3) *with a constant* $C = Le^{2M}$ *and, furthermore,* $\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \leq (e^{2L} - 1)|z - z'|$.

PROOF. By Lemmas 6.8 and 6.9, since g is L-Lipschitz in z for all x, y we have that $\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \leq (e^{2L} - 1)|z - z'|$. It remains to show that (6.3) holds with the

appropriate constant C. By definition, we have

$$\frac{|\exp(g(x,y,z)) - \exp(g(x',y',z))|}{\int_{[0,1]^2} \exp(g(x,y,z))\,dx\,dy} \le \exp(M)\big|\exp\big(g(x,y,z)\big) - \exp\big(g(x',y',z)\big)\big|.$$

Denote for brevity $g = g(x,y,z)$ and $g' = g(x',y',z)$ and note that $|g|, |g'| \le M$. By a Taylor expansion,

$$\begin{aligned}|e^g - e^{g'}| &\le |g - g'| \sum_{k=1}^{\infty} \frac{\sum_{i=0}^{k-1} |g|^i |g'|^{(k-1-i)}}{k!} \le |g - g'| \exp(M) \\ &\le L \exp(M)\big(|x - x'| + |y - y'|\big).\end{aligned}$$

We conclude that

$$\frac{|\exp(g(x,y,z)) - \exp(g(x',y',z))|}{\int_{[0,1]^2} \exp(g(x,y,z))\,dx\,dy} \le L \exp(2M)\big(|x - x'| + |y - y'|\big),$$

which is our desired result. □

7. Simulations. In this section, we report some numerical results on synthetic data to validate some of our theoretical predictions.

We note that all of our procedures require specifying a rejection threshold τ for the different tests. While we know the precise order of τ we do not know the appropriate constant. In order to handle this in practice, we use a permutation approach which is often used in practice (see, for instance, [41]). In more detail, we calculate the statistic T, and perform a permutation to obtain a reference distribution for the test statistic T under the null hypothesis. Recall that we construct the datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i \in C_m, i \in [N]\}$ for each of the d bins C_m. For each $\mathcal{D}_m$ we permute the X_i and Y_i values to simulate independently drawn values. Suppose that σ_m samples fall in the bin C_m, then for a permutation $\pi : [\sigma_m] \mapsto [\sigma_m]$ we consider $\mathcal{D}_m^{\pi} = \{(X_{\pi(i)}, Y_i) : Z_i \in C_m, i \in [N]\}$. We recalculate the statistic T over different sets $\mathcal{D}_m^{\pi}$ (using different permutations π for each set), and we repeat this M times, each time denoting the value permuted statistic with T_i for $i \in [M]$. Finally, we compare our statistic T with the values of the statistics in the set $\{T_1, \dots, T_M\}$ and return the value $M^{-1} \sum_{i \in [M]} \mathbb{1}(T_i > T)$. We would then reject the null hypothesis if this value is smaller than some pre-specified cutoff (say 0.05).

This procedure is motivated by the intuition that permuting indexes within bins $Z_i \in C_m$ generates approximately conditionally independent samples. While this intuition is apparent, in contrast to the settings of two-sample testing and independence testing, it is not straightforward to show that this procedure correctly controls the Type I error. We note that this permutation procedure works remarkably well in practice. However, rigorously proving the validity of this permutation procedure, and studying its power, warrants further research and is delegated to future work.

7.1. *Finite discrete X and Y.* In this subsection, we consider finite discrete X and Y with fixed number of categories $\ell_1 = 2$ and $\ell_2 = 3$. In order for us to construct examples that satisfy the conditions of Theorems 5.2 or 5.5, we rely on the examples studied in Section 6. Under the null hypothesis, we consider the following probabilities:

$$\begin{aligned}&p_{X,Y|Z}(1,1|z) \propto \exp\big(z + \tanh(z)\big), \qquad p_{X,Y|Z}(1,2|z) \propto \exp\big(z + \cos(z)\big), \\ &p_{X,Y|Z}(1,3|z) \propto \exp\big(z + \sin(z)\big), \qquad p_{X,Y|Z}(2,1|z) \propto \exp\big(\cos(z) - 1 + \tanh(z)\big), \\ &p_{X,Y|Z}(2,2|z) \propto \exp\big(\cos(z) - 1 + \cos(z)\big), \qquad p_{X,Y|Z}(2,3|z) \propto \exp\big(\cos(z) - 1 + \sin(z)\big).\end{aligned}$$

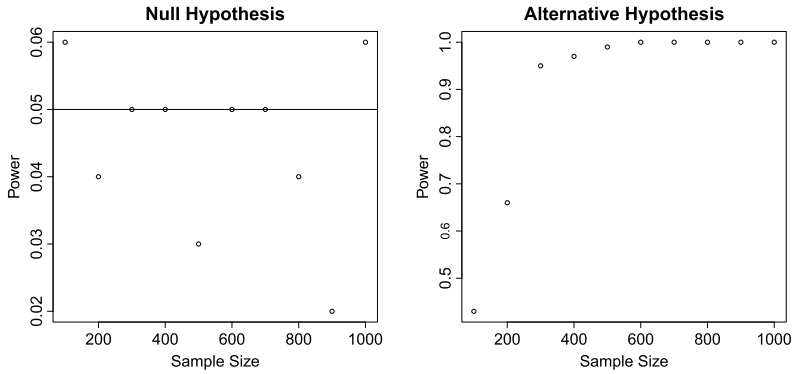

FIG. 2. *This figure displays the size and power of the test in the discrete X, Y and continuous Z example. We see that under the null hypothesis the size is gravitating around* 0.05 *which is also the most common size across all simulations. The power of the test increases steadily with the increase of the sample size, and reaches* 1 *when the sample size is* 1000.

In this setting, all of the exponents are Lipschitz, and can be decomposed so that the random variables are conditionally independent. Under the alternative, we consider the following distribution:

$$p_{X,Y|Z}(1,1|z) \propto \exp(z), \qquad p_{X,Y|Z}(1,2|z) \propto \exp\bigl(\tanh(z)\bigr),$$
$$p_{X,Y|Z}(1,3|z) \propto \exp\bigl(\sin(z)\bigr), \qquad p_{X,Y|Z}(2,1|z) \propto \exp\bigl(\cos(z)\bigr),$$
$$p_{X,Y|Z}(2,2|z) \propto \exp(z+1), \qquad p_{X,Y|Z}(2,3|z) \propto \exp\bigl(\tanh(z)-1\bigr).$$

In the example above, the probabilities do not factor as products so the variables are not conditionally independent; however, all functions in the exponents are still Lipschitz so that the distribution is TV smooth by Lemma 6.9. Figure 2 shows the results of running the weighted test of Section 5.2 on the above examples. For each sample size of $N = 100, 200, \ldots, 1000$, we perform 100 simulations. Within each simulation, we permute $M = 100$ times and compute the value $M^{-1}\sum_{i\in[M]} \mathbb{1}(T_i > T)$. The final size and power are calculated based on how many (out of the 100) values were smaller than or equal to 0.05.

7.2. *Continuous X, Y and Z.* In this subsection, we consider the following examples. Under H_0, we generate

$$X = \frac{U_1 + Z}{2} \quad \text{and} \quad Y = \frac{U_2 + Z}{2},$$

where $U_1, U_2, Z \sim U([0,1])$ are independent. Under the alternative, H_1, we generate

$$X = \frac{U_1 + U + Z}{3} \quad \text{and} \quad Y = \frac{U_2 + U + Z}{3},$$

where $U, U_1, U_2, Z \sim U([0,1])$ are independent. A straightforward calculation (see Appendix F of the supplement) shows that these distributions belong to the classes $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and $\mathcal{Q}_{0,[0,1]^3,\mathrm{TV}}(L,1)$ (respectively) for appropriately chosen constants L, so that the conditions of Theorem 5.6 hold.

Figure 3 shows the results of running the weighted continuous test described in Section 5.3 for these examples. For each sample size of $N = 100, 200, \ldots, 1000$, we perform 100 simulations. Within each simulation, we permute $M = 100$ times and compute the value $M^{-1}\sum_{i\in[M]} \mathbb{1}(T_i > T)$. The final size and power are calculated based on how many (out of the 100) values were smaller than or equal to 0.05.

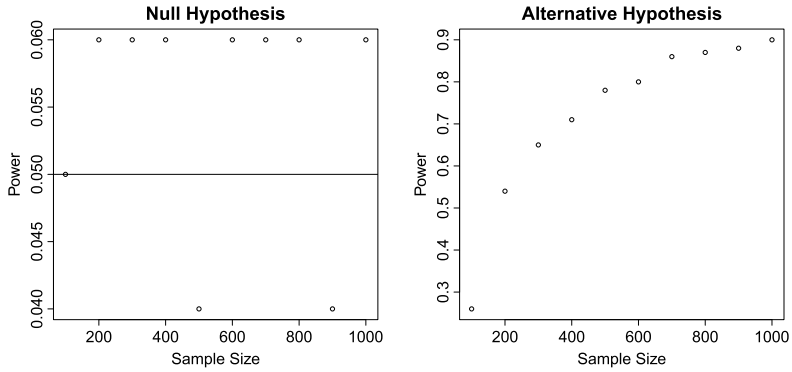

FIG. 3. *This figure displays the size and power of the test in the continuous X, Y, Z example. We see that under the null hypothesis the size is very slightly inflated at* 0.06 *for most of the simulations, which may be due to the limited number of replications of each simulation and also due to the limited number of permutations within each simulation. The power of the test increases steadily with the increase of the sample size, and reaches* 0.9 *when the sample size is* 1000.

8. Discussion. In this paper, we have studied nonparametric CI testing from a minimax perspective. We derived upper and lower bounds on the minimax critical radius in three main settings—(1) X, Y discrete and supported on a fixed number of categories, Z continuous on $[0, 1]$, (2) X, Y discrete on a growing number of categories Z continuous on $[0, 1]$ and (3) X, Y, Z absolutely continuous and supported on $[0, 1]$.[3] In order to develop interesting minimax bounds, we introduced and studied several natural Lipschitzness conditions for conditional distributions. In addition, we provided a novel construction of a coupling between a conditionally independent distribution and an arbitrary distribution of bounded support, leading to a new proof of the hardness result of Shah and Peters [31]. Finally, the CI tests that we developed are implementable and perform well in practice as evidenced by our simulation study in Section 7.

There are several open questions which we intend to investigate in our future work. Moving beyond the total variation metric, a natural challenge is to derive minimax rates for the critical radius in other metrics. Another technical challenge is to move beyond the requirement that $\frac{\ell_1^4}{\ell_2} \lesssim n^3$ (where $\ell_1 \geq \ell_2$), which we impose in the scaling ℓ_1, ℓ_2 case. We believe that the analysis in this case is challenging and would require designing new tests, or deriving new lower bound techniques, and is left for future research. Identifying conditions under which the natural permutation procedure of Section 7 correctly controls the Type I error and has high power is also a challenging direction that we hope to pursue.

Acknowledgments. The authors would like to thank the editor, associate editor and two anonymous referees whose comments and constructive suggestions led to significant improvements of the manuscript. The authors are also grateful to Ilmun Kim for various discussions on the topic.

The second author was partially supported by NSF DMS17130003, NSF CCF1763734.

The third author was supported by NSF DMS1713003.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Minimax optimal conditional independence testing" (DOI: 10.1214/20-AOS2030SUPP; .pdf). The supplementary material contains all omitted proofs as well as some extensions of the results presented in the main text.

REFERENCES

[1] AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* **7** 131–177. MR1173420

[2] ARIAS-CASTRO, E., PELLETIER, B. and SALIGRAMA, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.* **30** 448–471. MR3794401 https://doi.org/10.1080/10485252.2018.1435875

[3] AZADKIA, M. and CHATTERJEE, S. (2019). A simple measure of conditional dependence. Preprint. Available at arXiv:1910.12327.

[4] BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.* **12** 727–749. MR3834283 https://doi.org/10.1214/18-AOAS1155SF

[5] BALAKRISHNAN, S. and WASSERMAN, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

[6] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[7] BERGSMA, W. (2010). Nonparametric testing of conditional independence by means of the partial copula. Available at SSRN 1702981.

[8] BERGSMA, W. P. (2004). Testing conditional independence for continuous random variables. *Eurandom*.

[9] BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2018). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981 https://doi.org/10.1111/rssb.12340

[10] BLAIS, E., CANONNE, C. L. and GUR, T. (2019). Distribution testing lower bounds via reductions from communication complexity. *ACM Trans. Comput. Theory* **11** 1–37. MR3940784 https://doi.org/10.1145/3305270

[11] CANONNE, C. L. (2015). A survey on distribution testing: Your data is big. But is it blue? In *Electronic Colloquium on Computational Complexity* (*ECCC*) **22** 1–1.

[12] CANONNE, C. L., DIAKONIKOLAS, I., KANE, D. M. and STEWART, A. (2018). Testing conditional independence of discrete distributions. In *STOC'*18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 735–748. ACM, New York. MR3826290 https://doi.org/10.1145/3188745.3188756

[13] DAWID, A. P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B* **41** 1–31. MR0535541

[14] DIAKONIKOLAS, I. and KANE, D. M. (2016). A new approach for testing properties of discrete distributions. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 685–694. IEEE Computer Soc., Los Alamitos, CA. MR3631031

[15] DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. In *UAI* 132–141.

[16] FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* 489–496.

[17] GRETTON, A. and GYÖRFI, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* **11** 1391–1423. MR2645456

[18] GYÖRFI, L. and WALK, H. (2012). Strongly consistent nonparametric tests of conditional independence. *Statist. Probab. Lett.* **82** 1145–1150. MR2915081 https://doi.org/10.1016/j.spl.2012.02.023

[19] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294 https://doi.org/10.1214/aoms/1177730196

[20] HUANG, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* **38** 2047–2091. MR2676883 https://doi.org/10.1214/09-AOS770

[21] INGSTER, Y. I. Minimax nonparametric detection of signals in white Gaussian noise. *Problemy Peredachi Informatsii* **18** 61–73. MR0689340

[22] INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446 https://doi.org/10.1007/978-0-387-21580-8

[23] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

[24] MARGARITIS, D. (2005). Distribution-free learning of Bayesian network structure in continuous domains. In *AAAI* **5** 825–830.

[25] NEYKOV, M., BALAKRISHNAN, S. and WASSERMAN, L. (2021). Supplement to "Minimax optimal conditional independence testing." https://doi.org/10.1214/20-AOS2030SUPP.

[26] PATRA, R. K., SEN, B. and SZÉKELY, G. J. (2016). On a nonparametric notion of residual and its applications. *Statist. Probab. Lett.* **109** 208–213. MR3434980 https://doi.org/10.1016/j.spl.2015.10.011

[27] PEARL, J. (2014). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Elsevier.

[28] ROSENBAUM, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika* **49** 425–435. MR0760206 https://doi.org/10.1007/BF02306030

[29] SEN, R., THEERTHA SURESH, A., SHANMUGAM, K., DIMAKIS, A. G. and SHAKKOTTAI, S. (2017). Model-powered conditional independence test. In *Advances in Neural Information Processing Systems* 2951–2961.

[30] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[31] SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

[32] SONG, K. (2009). Testing conditional independence via Rosenblatt transforms. *Ann. Statist.* **37** 4011–4045. MR2572451 https://doi.org/10.1214/09-AOS704

[33] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

[34] SU, L. and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *J. Econometrics* **141** 807–834. MR2413488 https://doi.org/10.1016/j.jeconom.2006.11.006

[35] SU, L. and WHITE, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* **24** 829–864. MR2428851 https://doi.org/10.1017/S0266466608080341

[36] SU, L. and WHITE, H. (2014). Testing conditional independence via empirical likelihood. *J. Econometrics* **182** 27–44. MR3212759 https://doi.org/10.1016/j.jeconom.2014.04.006

[37] VALIANT, G. and VALIANT, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.* **46** 429–455. MR3614697 https://doi.org/10.1137/151002526

[38] VALIANT, G. and VALIANT, P. (2017). Estimating the unseen: Improved estimators for entropy and other properties. *J. ACM* **64** Art. 37, 41. MR3713795 https://doi.org/10.1145/3125643

[39] WANG, X. and HONG, Y. (2018). Characteristic function based testing for conditional independence: A nonparametric regression approach. *Econometric Theory* **34** 815–849. MR3815468 https://doi.org/10.1017/S026646661700010X

[40] YAO, Q. and TRITCHLER, D. (1993). An exact analysis of conditional independence in several 2×2 contingency tables. *Biometrics* **49** 233–236. MR1221407 https://doi.org/10.2307/2532617

[41] ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2012). Kernel-based conditional independence test and application in causal discovery. Preprint. Available at arXiv:1202.3775.